La traducció automàtica

Juan Alberto Alonso Martín

PID_00159192

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

La traducció automàtica (TA) va ser una de les primeres aplicacions no numèriques que va sorgir arran de l'aparició dels primers ordinadors, a finals de la dècada dels 50 i principis dels 60. Curiosament, és al mateix temps una de les aplicacions computacionals més difícils de tractar i, en conseqüència, una de les que probablement més trigaran a arribar a tenir uns resultats òptims.

Podríem definir la *traducció automàtica* com la branca de la lingüística computacional que s'ocupa del disseny, la implementació, l'avaluació i l'ús de programes d'ordinador per traduir textos d'un idioma a un altre.

Hi ha quelcom de paradoxal a la TA. Traduir és una activitat que moltes persones, incloent-hi persones sense cap formació acadèmica, fan normalment cada dia, aparentment sense gaires dificultats (si més no, pensem el que passa a Catalunya, on el llenguatge parlat i escrit salta contínuament entre català i castellà, en totes dues direccions). Això ens pot fer creure que, tot tenint en compte la sofisticació a què s'ha arribat en el camp de la informàtica, traduir pot ser una tasca que un programa d'ordinador podria arribar a fer sense grans dificultats. És això, de fet, el que van pensar els primers informàtics dels anys 50 i dels 60, quan, d'una manera més aviat ingènua, van intentar fer programes que traduïssin de l'anglès al rus. Aleshores es pensava que el procés de traducció era poc menys que una qüestió de tenir grans diccionaris emmagatzemats a l'ordinador. Tanmateix, a la vista dels primers resultats, força decebedors, aviat es van adonar que la tasca no era tan senzilla. Traduir no consisteix només a substituir paraules. S'ha de triar la paraula adequada en la llengua d'arribada i sovint s'han de fer canvis sintàctics força importants a l'estructura de la frase de sortida. Fins i tot la traducció entre dues llengües lingüísticament tan properes com són el català i el castellà presenta dificultats enormes per als programes de traducció automàtica. A què es deguda aquesta dificultat dels programes informàtics per manegar la traducció entre llengües? Doncs precisament al fet que la matèria primera amb la qual treballen aquests programes és el llenguatge humà, molt probablement el sistema simbòlic més complex que coneixem. Una de les característiques dels llenguatges humans, en oposició al que passa en els llenguatges d'ordinador (com ara el BASIC, el C o el Java) és que són intrínsecament ambigus. Sovint una mateixa paraula té més d'un significat (penseu en la paraula catalana *cap*) i fins i tot el significat de frases senceres pot ser ambigu (penseu en la frase *No es poden col·locar blocs **fins** que no hagin estat aprovats per la direcció d'obres: No se pueden colocar bloques **finos/hasta** que no hayan sido aprobados por la dirección de obras*). És aquesta ambigüitat la que fa que el llenguatge humà sigui tan difícil de tractar per un programa d'ordinador.

TA

A partir d'ara abreujarem *traducció automàtica* amb la sigla *TA*.

Com veurem en aquest mòdul, els programes de traducció automàtica fan ús de tot l'arsenal d'eines lingüístiques que hi ha actualment a l'abast: lematitzadors i etiquetadors morfològics, analitzadors morfològics i sintàctics, lèxics computacionals monolingües i bilingües, gramàtiques complexes amb centenars de regles sintàctiques, mecanismes d'assignació de funció sintàctica i papers temàtics, mecanismes de resolució d'anàfora, etc. De manera alternativa, com també veurem, darrerament han aparegut sistemes de traducció automàtica que no fan servir coneixement lingüístic, sinó algorismes estadístics que intenten extreure informació probabilística a partir de grans corpus monolingües i bilingües, informació que després es fa servir per generar traduccions.

Objectius

En aquest mòdul farem un cop d'ull a alguns aspectes relacionats amb la traducció automàtica: les seves característiques i les diferències amb altres disciplines, la seva història, les seves limitacions, els tipus de sistemes de TA que existeixen actualment, els principals components d'aquests sistemes, el procés que segueix un sistema de TA per dur a terme la traducció d'un text, les aplicacions actuals de la TA i el seu ús a Internet.

En l'espai assignat a aquest mòdul és impossible fer una descripció detallada de tot el que té a veure amb la TA. Per això, hem hagut de deixar de banda alguns punts que podrien semblar rellevants (exemples comentats de sistemes de TA existents, tècniques de disseny i desenvolupament de sistemes de TA, tècniques d'avaluació de la qualitat d'un sistema de TA, etc.) i no sempre hem pogut dedicar tota l'extensió desitjable als punts que hem cobert al mòdul.

Tanmateix, esperem que, després de llegir el mòdul, el lector pugui tenir una idea bastant exacta de què és la *traducció automàtica*, de com funciona i de què es pot esperar dels sistemes de TA actuals.

1. Què és i què no és la *TA*?

La TA és una de les principals branques de la lingüística computacional. Com ja s'ha dit a la introducció, el seu àmbit d'aplicació és la traducció de textos entre dues llengües.

El fet que la traducció sigui l'eix central fa que, de vegades, altres disciplines i programes informàtics que també tenen a veure amb les tasques de traducció siguin confosos amb programes de traducció automàtica.

Més concretament, **no** és traducció automàtica:

- La traducció simultània: encara que la distinció sembli òbvia, hi ha molta gent que confon traducció automàtica i traducció simultània, segurament degut a la similitud entre els dos termes. No cal insistir que no tenen res a veure l'una amb l'altra: la traducció automàtica és una branca de la lingüística computacional que s'ocupa de la traducció automàtica o semiautomàtica de textos mitjançant programes d'ordinador, mentre que la traducció simultània és una disciplina de la traducció humana.

- Els diccionaris electrònics: són programes que donen la traducció de paraules aïllades.

- Els programes de reconeixement i generació de veu: són programes que converteixen frases parlades en el corresponent text escrit (reconeixement de veu). Els programes de generació de veu passen d'un text escrit a la seva versió parlada.

- Els correctors ortogràfics i gramaticals: són programes que detecten i, quan és possible, corregeixen els errors ortogràfics i gramaticals d'un text escrit.

2. Breu història de la TA

Es pot dir que la idea de la traducció automàtica va sorgir fa tres segles, amb les propostes fetes per Descartes i Leibniz al segle XVII per fer diccionaris basats en codis numèrics. Ja al nostre segle, durant la dècada dels anys 30, es van presentar alguns enginys mecànics que podien traduir textos paraula per paraula amb l'ajut d'una cinta perforada de paper.

Podríem dir que la primera presentació d'un sistema real de traducció automàtica es va fer als Estats Units el 1954. Es tractava d'un prototipus amb només 250 entrades lèxiques i 6 regles gramaticals que podia traduir algunes frases "preparades" del rus a l'anglès. Aquesta demostració, tot i la seva simplicitat, va obrir grans expectatives i durant la resta de la dècada dels 50 i la primera meitat dels anys 60 es van dedicar molts diners i moltes hores de recerca per intentar aconseguir sistemes de traducció totalment automàtica d'alta qualitat[1].

(1)En anglès, *FAHQT*: *Full Automatic High-Quality Translation*.

Tanmateix, segons s'avançava en el desenvolupament d'aquests sistemes es va fer palès que les tècniques utilitzades no eren suficients per resoldre els nombrosos problemes lingüístics que sorgien. L'aparició el 1966 del famós informe ALPAC[2], al qual s'afirmava que, vistos els resultats, no es podia esperar la consecució de sistemes de TA utilitzables a curt o mitjà termini, va suposar una aturada en la inversió i investigació en aquest camp.

(2)Sigles corresponents al nom en anglès del Comitè consultiu per al processament automàtic del llenguatge.

Durant la dècada següent es va continuar amb la recerca sobre traducció automàtica, principalment al Canadà, Europa (incloent-hi Rússia), Israel i el Japó. El 1976 va aparèixer al Canadà el sistema METEO, destinat a traduir informes meteorològics entre l'anglès i el francès. El mateix any, la Comissió Europea va decidir instal·lar el sistema SYSTRAN, entre l'anglès i el francès. Més tard s'hi van afegir més parells de llengües. Al final de la dècada dels 70 es va començar el projecte Eurotra, que durant els anys 80 va concentrar els esforços de molts grups de recerca per aconseguir un sistema de TA multilingüe entre totes les llengües comunitàries. Al mateix temps, la Universitat d'Austin, Texas, un dels pocs centres dels Estats Units que van seguir treballant en traducció automàtica va construir el precursor del que més tard seria el sistema METAL, desenvolupat i comercialitzat a la dècada dels 80 per l'empresa Siemens a Europa, sistema que encara perviu en el seu successor LT Translator. En general, durant la dècada dels 70, la majoria dels sistemes desenvolupats utilitzaven una estratègia basada en la transferència.

Vegeu també

L'estratègia basada en la transferència es tracta al subapartat 4.2.3 d'aquest mòdul.

La dècada dels 80 va representar la consolidació dels sistemes basats en transferència i l'aparició d'alguns sistemes basats en tècniques d'interlingua, com, per exemple, el sistema KBMT de la Universitat de Carnegie Mellon, als Estats Units.

A la dècada dels 90, el més significatiu ha estat el salt dels sistemes de TA des dels laboratoris de recerca i desenvolupament a les botigues de programes informàtics i, darrerament, a Internet. Programes per a entorns PC MS-DOS/Windows, com ara Globalink, PC-Translator, Transcend i d'altres, han aparegut en els últims anys. La major part d'aquests programes tenen molt poca potència lingüística (pertanyen al tipus de sistemes anomenats *de traducció directa*) i els seus resultats quant a la qualitat de traducció són bastant pobres. Més recentment, s'han començat a utilitzar sistemes més potents en entorns específics (normalment xarxes locals en configuracions client-servidor) i la traducció automàtica ha pres posicions dintre d'Internet.

3. Els límits de la TA

3.1. La importància del "coneixement del món"

Quan els usuaris comproven la qualitat de la traducció oferta pels programes actuals de TA sovint es pregunten com és possible que, amb la potència dels ordinadors i la sofisticació dels programes informàtics d'avui, no es puguin assolir resultats millors.

Per què resulta més fàcil fer programes d'ordinador per dirigir automàticament una sonda espacial de la Terra a Mart que per traduir completament bé una frase del català al castellà? Això por semblar absurd a primera vista, però és totalment cert. La resposta rau en la matèria primera amb la qual treballen els programes de traducció humana: el llenguatge humà. Les llengües que fem servir cada dia són segurament, encara que no en siguem conscients, el sistema simbòlic més complicat que coneixem. No només utilitzem les paraules pròpies de cada idioma (el lèxic), sinó que aquestes paraules apareixen en unes formes (morfologia) i en un ordre (sintaxi) molt específics i, a més –i aquesta és la gran dificultat per als programes d'ordinador–, la informació que es transmet mitjançant el llenguatge es complementa amb inferències i suposicions que anem fent contínuament sobre allò de què es parla. Aquestes inferències i suposicions les podem fer d'acord amb el coneixement (inconscient) que tenim de com funciona el món que ens envolta.

Així, enfront de dues frases aparentment molt semblants com són:

a) Els pingüins poden nedar però no volen.

b) Els nens poden nedar però no volen.

la interpretació més normal que fem els humans és, per a la primera frase, que els pingüins poden nedar però no poden **volar**, mentre que per a la segona solem entendre que els nens poden nedar però no **volen** nedar. Aquesta diferència d'interpretació per a dues frases amb la mateixa estructura morfosintàctica (només canvia un element lèxic: ***pingüí***/***nen***!) la fem perquè sabem que:

- Els pingüins són aus que, tot i tenir ales, no poden volar.
- Els pingüins són aus que, per contra, poden nedar molt bé.
- Els nens no poden volar (tret que no sigui en avió).

- Dir d'un nen que *pot nedar però no vola* és possible però molt improbable; la frase hauria d'estar inclosa en un discurs molt específic per tenir sentit.

- Dir d'un pingüí que *pot nedar però no ho vol fer* és igualment possible, però és igualment improbable; és una frase que també hauria d'estar inclosa en un discurs molt específic per tenir sentit.

Aquestes "peces de coneixement" és el que anomenem *coneixement de món* i és precisament on rau la diferència fonamental entre els humans i els ordinadors. Als programes de traducció automàtica hi pot haver lèxics molt extensos, amb molta informació lexicogràfica, morfològica, sintàctica i semàntica; hi pot haver moltes regles gramaticals que permetin el programa analitzar l'estructura de constituents de frases sintàcticament molt complexes; però, ara per ara, és extraordinàriament difícil incloure en aquests programes informació sobre com funciona el món que ens envolta. Hi ha diverses raons que podem citar com a responsables d'aquesta dificultat:

- La informació sobre el coneixement del món és massa extensa per poder-la introduir de manera eficaç en un programa d'ordinador. Pensem que, si no ens restringim a un domini molt específic (per exemple, la informació meteorològica o la informació turística d'una ciutat petita), estem parlant no només de totes les característiques de totes les possibles entitats del món (éssers vius, objectes i conceptes), sinó de totes les possibles relacions rellevants entre aquestes entitats.

- La informació sobre el coneixement del món és massa complexa per poder-la formalitzar de manera que es pugui utilitzar eficaçment pels programes de TA.

La conseqüència directa de tot això que hem explicat és que els programes de TA actuals poden fer traduccions prou bones sempre que aquestes depenguin d'informació lèxica, morfològica, sintàctica o, en part, semàntica. Tanmateix, el grau de qualitat de la traducció baixa radicalment quan entren en joc aspectes que tenen a veure amb la pragmàtica.

Òbviament, hi ha altres factors que determinen la qualitat de la traducció oferta pel sistema: el grau de proximitat de les llengües entre les quals es tradueix, el tipus de text traduït, etc.

3.2. Què significa *traduir bé*?

Quan un traductor humà tradueix un text entre dues llengües, normalment el que fa és parafrasejar en la llengua d'arribada allò que ha entès en la llengua de partida. Justament és aquesta capacitat de posar en altres paraules d'una altra llengua el significat original el que caracteritza un bon traductor humà. Només cal pensar en algunes traduccions humanes que trobem de tant en tant, en

què queda clar que el traductor no ha entès el que està traduint (ja sigui per desconeixement de la llengua o del tema del text original) i, a causa d'això, la traducció feta no té sentit.

Traduir bé significa, doncs, **entendre** el que s'està traduint i reflectir-ne el significat amb les paraules i estructures adequades de la llengua d'arribada. En aquest sentit, traduir bé és quelcom que, ara per ara, només pot fer un bon traductor humà. Aquest és el límit fonamental de la traducció automàtica, i el seu repte, apropar-se tant com sigui possible a aquest objectiu.

Tornem a fer la pregunta: per què ni l'ordinador més potent del món és capaç de fer una traducció comparable a la que pot fer qualsevol traductor humà mínimament competent?

La clau d'aquest *perquè* rau en una paraula de la definició que hem donat abans del que vol dir *traduir:* "**entendre** el que s'està traduint i reflectir-ne el significat amb les paraules i estructures adequades de la llengua d'arribada". *Entendre* és la clau del problema. Un ordinador pot buscar la traducció d'una paraula en grans diccionaris molt més ràpidament que qualsevol humà. Pot fer anàlisis sintàctiques de les frases d'entrada i generar-ne les frases de sortida en qüestió de mil·lisegons. Però un ordinador no pot *entendre* el que està traduint, i per això sovint s'equivoca, i dóna la traducció incorrecta d'una paraula o agafa una interpretació sintàctica d'una frase que, tot i ser formalment possible, no correspon al significat real d'aquesta frase (al que qualsevol humà li donaria). La raó última de per què això és així és que la manera de processar informació que té el nostre cervell és radicalment diferent de la que fan servir els ordinadors. Una possible explicació detallada d'aquest factor es pot trobar a Hawkins (2004). A tall de resum, els ordinadors processen la informació d'una manera seqüencial i per això són imbatibles quan es tracta de dur a terme una tasca molt complicada en un temps molt curt, però amb la condició que aquesta tasca es pugui descompondre en un conjunt de subtasques que es puguin fer una darrera l'altra.

D'altra banda, el nostre cervell processa la informació de manera paral·lela, per nivells de complexitat i amb realimentació, amb un còrtex prefrontal que té infinitat de grups de neurones, organitzades en capes, interconnectades entre elles i amb altres grups; cadascun d'aquests grups aprèn a encarregar-se d'una tasca determinada –que pot ser molt simple: reconèixer una línia recta o un so determinat de la llengua, o més complexa: reconèixer una cara coneguda o una paraula o frase de la llengua. Això fa que, al contrari del que passa amb els ordinadors, nosaltres siguem molt lents a processar seqüències de tasques llargues (per exemple, fer operacions matemàtiques complexes) però que, en canvi, siguem molt eficients i ràpids a reconèixer i predir patrons –una cara coneguda enmig de molta gent, una melodia coneguda enmig del soroll–, cosa que els ordinadors no poden fer bé. Com ja hem dit abans, el llenguatge humà és segurament el sistema simbòlic més complex que es coneix i ha estat evolutivament modelat durant centenars de milers d'anys pels nostres cervells

(Pinker, 1994, 2007; Deacon, 1997). És per això que nosaltres no trobem cap dificultat per entendre una frase en una llengua que coneguem, ja que la forma que té el nostre cervell de processar la informació s'hi adapta perfectament, mentre que fer el mateix amb un processament seqüencial com el que fan servir els ordinadors és pràcticament impossible.

4. Tipus de sistemes de TA

4.1. Panoràmica actual dels diferents tipus de sistemes de TA

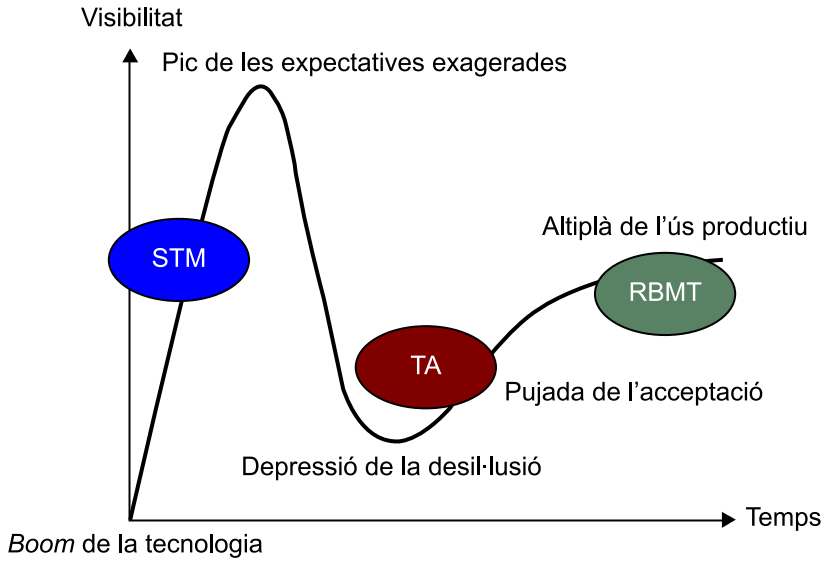
Mentre que els sistemes de TA basats en coneixement lingüístic –també anomenats *sistemes basats en regles*[3] – ja es fan servir de manera comercial o productiva des de fa anys i es pot considerar que és una tecnologia relativament estable, assentada i provada, durant els últims deu anys ha aparegut en escena una altra tecnologia de traducció automàtica radicalment diferent, la basada en tècniques i models estadístics[4].

[3] En anglès, *Rule Based Machine Translation* o RBMT.

[4] En anglès, *Statistical Machine Translation* o SMT.

La figura següent il·lustra l'estadi, dins de l'anomenat *cicle de sobreexpectació de Gartner*, en què es troben actualment els sistemes de TA basats en regles lingüístiques (RBMT), els sistemes basats en tècniques estadístiques (SMT) i la traducció automàtica en general (TA).

Figura 1. Cicle de sobreexpectació de Gartner



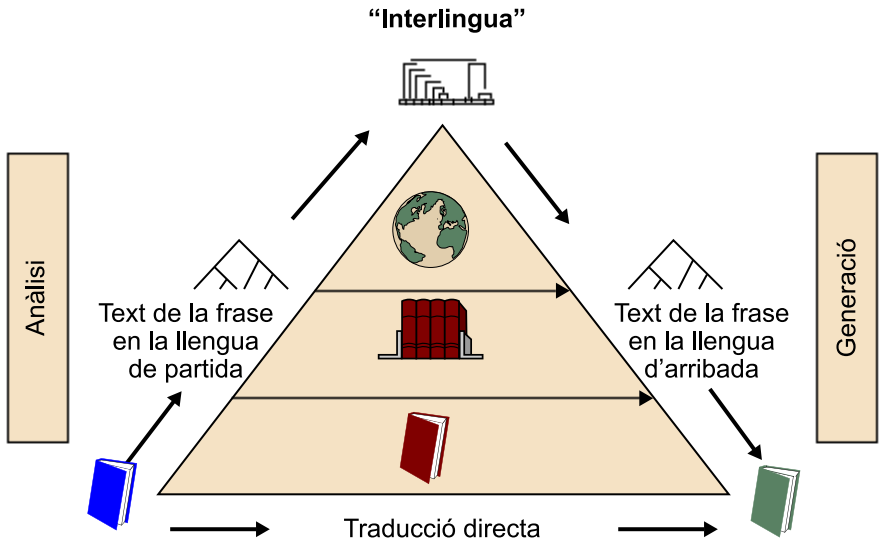
4.2. Sistemes basats en coneixement lingüístic (RBMT)

4.2.1. La "piràmide de la TA"

Tradicionalment, els sistemes de TA es classifiquen en tres grans tipus: sistemes de traducció directa, sistemes basats en transferència i sistemes basats en interlingua. Aquesta classificació es fa en funció de la "potència lingüística" del sistema (entenent per *potència lingüística* la quantitat d'informació lingüística

–lèxica, morfològica, sintàctica, semàntica etc.– que fa servir el sistema durant el procés de traducció). El triangle de la figura següent és una forma molt habitual de representar els tres tipus de sistemes.

Figura 2. Piràmide de la TA



Cal dir que aquesta classificació no és absoluta, en el sentit que un sistema específic no és purament de traducció directa, de transferència o d'interlingua. Els tres tipus representen més aviat franges d'un espectre continu, de manera que un sistema determinat pot ser classificat com "de transferència", però més cap a la banda de traducció directa, si la seva potència lingüística és més reduïda, o més cap a la banda d'interlingua, si la seva potència lingüística és més gran.

A continuació, veurem amb una mica més de detall cadascun d'aquests tipus i n'afegirem un altre no recollit al triangle: els sistemes de traducció basats en memòries de traducció. Finalment, farem esment d'altres aproximacions, de moment perifèriques, al problema de la traducció automàtica.

4.2.2. Sistemes de traducció directa

A la base de la piràmide de l'apartat anterior hi ha els sistemes de traducció directa. Aquests van ser els primers programes de traducció automàtica que van aparèixer. De fet, la filosofia que hi ha darrere d'aquests tipus de programes reflecteix la ingenuïtat de les primeres aplicacions de traducció automàtica, quan es pensava que tot era qüestió de tenir grans lèxics amb els quals poder traduir ràpidament les paraules de les frases del text.

Típicament, un sistema de traducció directa fa servir uns lèxics monolingües i bilingües molt grans. En alguns casos poden no tenir un mòdul d'anàlisi morfològica i, per tant, els lèxics monolingües han de tenir totes les formes per a cada lema (és a dir, no només *cantar*, sinó totes les formes del verb: *canto, cantes, canta*...). Per contra, la quantitat de coneixement lingüístic (morfo-

sintàctic) inclosa en aquests programes és molt limitada. No es fa una anàlisi sintàctica de les frases o, com a molt, se'n fa una de molt superficial. Molts dels sistemes de traducció que es comercialitzen actualment per a PC pertanyen a aquest tipus. Tenen dues característiques fonamentals: una gran rapidesa i una qualitat de traducció molt limitada. Clarament aquestes dues característiques estan relacionades entre elles: accedir a bases de dades lèxiques (els diccionaris del sistema) és una tasca que un ordinador pot fer d'una manera molt ràpida. Tanmateix, com ja hem vist, traduir és molt més que substituir paraules d'una llengua per les d'una altra.

4.2.3. Sistemes basats en transferència

Els sistemes basats en transferència van ser els primers sistemes de TA als quals es van aplicar tècniques de lingüística formal. Després del fracàs dels primers sistemes de traducció directa es va veure clar que per traduir una frase calia fer-ne una anàlisi lingüística tan a fons com fos possible.

Als sistemes de transferència, la traducció es realitza seguint tres fases: la fase d'anàlisi, la fase de transferència i la fase de generació.

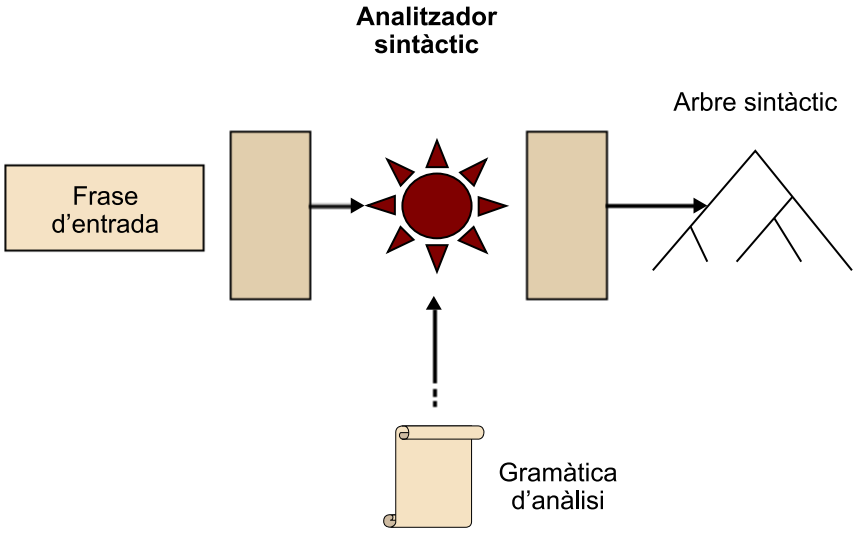
Fase d'anàlisi

Durant aquesta fase el programa fa una anàlisi morfològica de cada paraula de la frase. Tot seguit, se'n fa una anàlisi sintàctica, més o menys profunda, segons el sistema. L'anàlisi sintàctica dóna com a resultat un arbre amb l'estructura de constituents i informació morfològica (categoria, nombre, gènere, etc.), sintàctica (funció sintàctica del constituent respecte del seu nucli verbal, nominal o adjectival) i semàntica (tipus de nom, d'adjectiu, d'adverbi o de verb) associada a cada node de l'arbre.

És precisament aquesta informació morfològica, sintàctica i semàntica, que normalment està present als nodes de l'arbre en forma de parells tret-valor, juntament amb la informació estructural, la que es farà servir després a la fase de transferència per tal de triar les traduccions adequades segons el context de cada paraula a la frase i els canvis estructurals que calgui fer en traduir la frase a la llengua d'arribada.

Mentre que l'anàlisi morfològica és un aspecte que no presenta gaires problemes (tot i que encara no està resolt del tot, especialment pel que pertoca a la morfologia derivativa), és l'anàlisi sintàctica la que comporta un major grau de dificultat. L'anàlisi sintàctica es realitza mitjançant un programa anomenat *analitzador sintàctic* o ***parser***, el qual és guiat per un conjunt de regles d'estructura de frase que descriuen les diferents estructures sintàctiques de la llengua d'entrada (vegeu la figura següent).

Figura 3. Analitzador sintàctic



Segons el que acabem de dir, és obvi que la qualitat d'un sistema de TA basat en transferència és tan bona (o tan dolenta) com ho sigui la seva gramàtica d'anàlisi. Podríem dir que aquest és l'element clau del sistema i el més difícil de dissenyar i realitzar.

Fase de transferència

A partir de la informació present a l'arbre d'anàlisi, la fase de transferència s'encarrega de triar la traducció correcta per a cada paraula (en cas que aquesta en tingui més d'una, la qual cosa, com ja hem dit, passa força sovint). La informació que determina quina traducció s'ha d'agafar en cada cas és molt diversa. Així, en la traducció de verbs pot ser determinant la presència, absència o característiques de determinats arguments (subjecte, objecte directe, etc.); en la de noms, les característiques dels seus modificadors o complements adjectivals o preposicionals.

De vegades, quan se selecciona per a una paraula una traducció determinada, aquesta porta associada una transformació estructural que s'ha d'aplicar sobre la frase en la llengua d'arribada. Pensem, per exemple, en el cas del verb anglès *to love*. Podríem donar dues traduccions al català, depenent de si l'objecte directe és humà (*estimar*) o no humà (*agradar*). Mirem ara aquestes dues frases:

a) *John loves Mary* ⇒ *En John estima la Mary.*

b) *John loves music* ⇒ *A en John li agrada la música.*

Mentre que a (**a**) es conserva l'estructura sintàctica de l'anglès en la traducció catalana, a (**b**) el subjecte de *to love* (*John*) passa a ser l'objecte indirecte del verb català, i l'objecte directe del verb anglès (*music*) passa a ser el subjecte del

verb en la frase catalana. Aquest canvi de funcions sintàctiques és normalment especificat a l'entrada del diccionari bilingüe que relaciona *to love* amb *agradar*, i es duu a terme a la fase de transferència.

En resum, el resultat de la fase de transferència és un arbre sintàctic (l'arbre de transferència) semblant al de la fase d'anàlisi, però amb les paraules de l'idioma origen canviades per la seva traducció en la llengua d'arribada, i amb possibles canvis estructurals activats per determinades entrades del lèxic bilingüe.

Fase de generació

La fase de generació rep com a entrada l'arbre de transferència i s'ocupa de dur a terme un seguit de tasques pròpies de la llengua d'arribada, com ara:

1) Col·locar les paraules de la frase segons les regles d'ordre de constituents de la llengua d'arribada.

2) La inserció o eliminació de material lèxic (per exemple, la inserció dels pronoms febles *en* o *hi* a determinades frases quan traduïm del castellà al català, o l'eliminació en certs casos d'aquests mateixos pronoms febles si estem traduint del català al castellà).

3) La generació de les formes flexives adequades de les paraules de la frase de sortida, segons la informació present a l'arbre (per exemple, ***traduir***{1a. persona, plural, imperfecte d'indicatiu} ⇒ ***traduíem***).

4) La combinació i contracció d'elements lèxics (p. ex. en català, *dóna* + *el* + *hi* ⇒ *dóna-l'hi*, o *per el* ⇒ *pel*).

4.2.4. Sistemes d'interlingua

Els sistemes basats en interlingua són en realitat un cas extrem dels sistemes de transferència en què desapareix precisament la fase de transferència. En aquests sistemes, la fase d'anàlisi no consisteix només en una anàlisi morfosintàctica, sinó que el resultat final és una representació del significat de la frase en forma de xarxa semàntica (bàsicament, les entitats que apareixen a la frase representades com un conjunt de trets semàntics, i les relacions que hi ha entre aquestes entitats).

El que anomenem *interlingua* és precisament el llenguatge formal utilitzat per representar el significat de la frase.

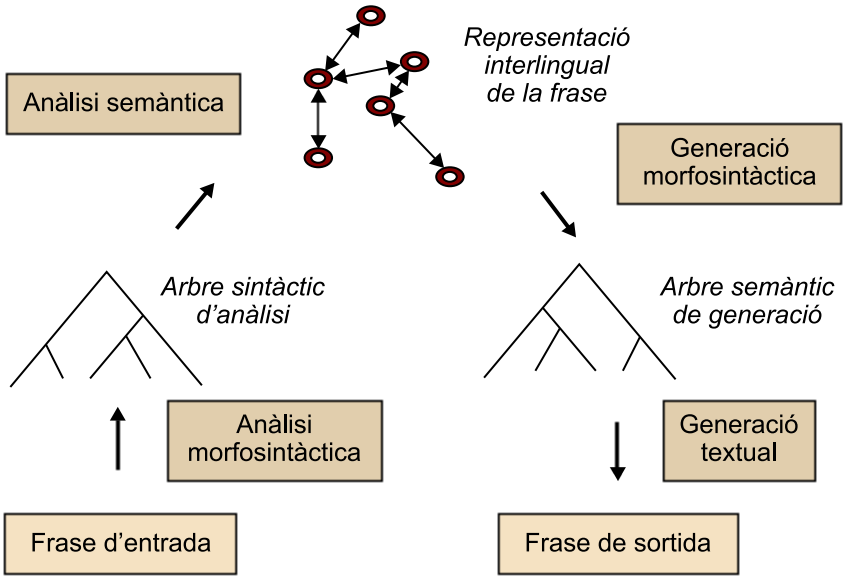
Interlingua

No hem de pensar que la interlingua és un llenguatge artificial però "humà", com ara una mena d'esperanto. Es tracta sempre d'un llenguatge formal, semblant als llenguatges d'ordinador.

La idea bàsica subjacent als sistemes d'interlingua és que, un cop "extret" el significat de la frase original i recollit en forma de representació formal, ja no cal una fase de transferència durant la qual es faci la traducció entre les paraules de la llengua de partida i les de la llengua d'arribada. De fet, la representació interlingual no conté paraules, sinó conceptes i relacions entre conceptes.

Així, la fase de generació consisteix a generar una frase amb les paraules i les estructures sintàctiques que expressin en la llengua d'arribada el significat de la representació interlingual. Això vol dir que, més que una traducció, el que fem és una paràfrasi de la frase original.

Figura 4. Esquema del procés interlingual



Encara que la idea dels sistemes d'interlingua és teòricament molt bona, a la pràctica ens trobem amb greus problemes de disseny i d'implementació.

En primer lloc, tenim el problema del disseny de la interlingua. Pensem el que significa tenir un llenguatge formal que sigui capaç de recollir acuradament el significat de qualsevol frase que ens pugui aparèixer a un text qualsevol. Aquest llenguatge hauria de contenir la descripció en forma de trets semàntics de totes les entitats (noms) que podem referenciar quan parlem (per exemple, ***taula, núvol, fred, gos, malenconia, estiu, política***, etc.), juntament amb la descripció de les propietats (adjectius) que cadascuna d'aquestes entitats pot tenir (per exemple, una **taula** pot ser **alta**, però no un **estiu**, ni una **malenconia;** una **política** pot ser **eficaç**, però un **núvol** o un **gos** no ho poden ser). A més s'han de poder representar totes les possibles relacions (verbs) entre les entitats (per exemple, un **gos** pot **bordar**, **mossegar**, **sentir fred** o **morir**, però un **estiu** no pot estar "connectat" amb cap d'aquestes relacions). És evident que, per a un domini no restringit (és a dir, per traduir qualsevol text sobre qualsevol tema) és pràcticament impossible dissenyar una interlingua amb aquestes característiques.

En segon lloc, fins i tot si imaginéssim que tenim una interlingua prou potent per recollir el significat de qualsevol frase, la capacitat computacional necessària per fer-la servir d'una manera eficaç al procés de traducció excedeix les possibilitats actuals dels ordinadors normals.

Així, doncs, podem concloure que els sistemes d'interlingua són una utopia, desitjable però impossible? De moment, és realment una utopia si estem parlant de fer servir aquesta tècnica per traduir qualsevol text, sense cap restricció de temes. Si parlem, però, de traduir textos en dominis restringits (i com més restringits, millor), aquests sistemes sí poden ser factibles.

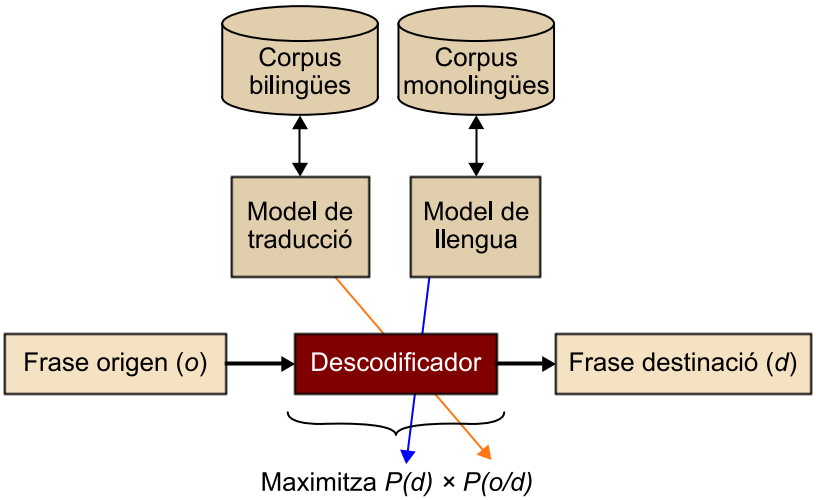
De fet, la tècnica d'interlingua s'ha fet servir en alguns sistemes de TA, especialment amb el japonès com a llengua de partida o de destinació, com, per exemple, al sistema ATLAS desenvolupat per l'empresa japonesa Fujitsu. En realitat, la tècnica utilitzada no és purament d'interlingua (sovint hi ha una petita fase de transferència), la qual cosa fa que aquests sistemes es puguin classificar com a sistemes híbrids transferència-interlingua (és a dir, classificables a la banda més alta del sistemes de transferència o a la banda més baixa dels sistemes d'interlingua).

4.3. Sistemes basats en models estadístics (SMT)

Els sistemes de TA basats en tècniques estadístiques es basen en algoritmes matemàtics que intenten relacionar grups de paraules de la llengua de partida amb grups de paraules en la llengua d'arribada, tot maximitzant les probabilitats que els grups de paraules de la llengua d'arribada siguin traduccions correctes de la llengua de partida. S'ha de tenir ben present que aquesta aproximació al problema de la traducció automàtica és radicalment diferent de la dels sistemes basats en regles lingüístiques. En aquest cas, es fan servir mètodes empírics, basats en la teoria de la informació i en l'assumpció que una traducció no és res més que la transmissió d'informació a través d'un canal sorollós (*noisy channel*).

Els components principals d'un sistema de TA estadístic són tres: el model de llengua, el model de traducció i el descodificador. El **model de llengua** assigna una probabilitat *P(d)* a cada cadena o seqüència ***d*** (de ***n*** paraules) en una llengua concreta –típicament, la llengua d'arribada. El **model de traducció** assigna una probabilitat *P(o|d)* a cada parell de cadenes o seqüències ***o*** en la llengua de partida i ***d*** en la llengua d'arribada. Finalment, **l'algorisme de descodificació** intenta maximitzar la probabilitat $P(d) \times P(o|d)$, és a dir, la probabilitat que la seqüència de paraules ***o*** de la llengua de partida tingui com a traducció la seqüència *d* de la llengua d'arribada i que la seqüència ***d*** de la llengua d'arribada sigui una seqüència correcta de paraules en aquesta llengua.

Figura 5. Components principals d'un sistema de TA estadístic



4.3.1. El model de llengua

Bàsicament, un model de llengua és un model probabilístic, sovint implementat en forma d'autòmat d'estats finits, que intenta descriure quines seqüències de paraules o quines estructures sintàctiques són més probables en una llengua concreta. En altres paraules, un model de llengua, posem per cas del català, indica a un sistema de TA estadístic fins a quin punt una seqüència de paraules en català és correcta o incorrecta. Així, un model de llengua del català assignaria a la seqüència ***d1*** de paraules *aquesta frase ha estat traduïda per* una probabilitat *P(d1)* molt alta de ser correcta, mentre que assignaria una probabilitat ***P(d2)*** molt baixa a la seqüència ***d2*** *aquesta estat frase per ha traduïda*.

Els models de llengua solen fer servir grups de ***n*** paraules, anomenats *n-grams*, en què ***n*** pot variar entre 2 i 7, depenent dels models. Les probabilitats es deriven normalment a partir de corpus monolingües i, òbviament, com més gran i més "net" (és a dir, com més correctes siguin les frases) sigui el corpus utilitzat, més acurat serà el model de llengua que se'n derivi.

4.3.2. El model de traducció

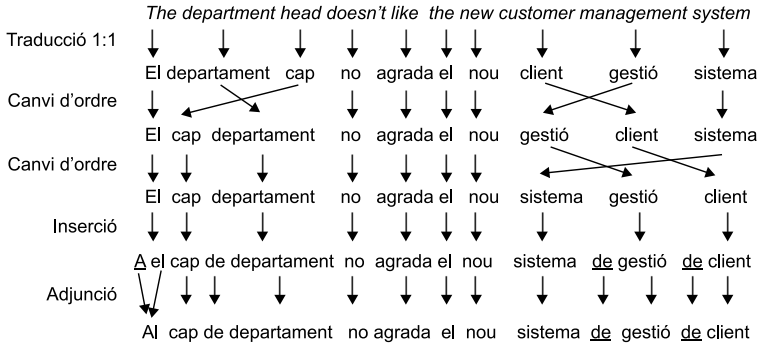
Com s'ha dit abans, el model de traducció és un model probabilístic que assigna probabilitats a parells de seqüències de paraules de la llengua de partida i de la llengua d'arribada. Aquestes probabilitats es deriven a partir de corpus bilingües paral·lelitzats i, igual que passava amb els models de llengua, com més extensos i ben alineats siguin aquests corpus, més "bo" serà el model de traducció que se'n derivi. Hi ha diferents tècniques per generar i refinar les probabilitats entre paraules o grups de paraules de la llengua de partida i la llengua d'arribada. La majoria d'aquestes tècniques impliquen processos d'aprenentatge (*training*) i iteració, bidireccionals, algorismes d'alineament de paraules a partir d'interseccions en corpus bilingües, tècniques heurístiques, etc. De vegades, les tècniques més apropiades depenen del parell de llengües en qüestió; no és el mateix intentar alinear paraules o seqüències de paraules entre el català i el castellà que entre el català i el turc, posem per cas. A ban-

da d'això, els models de traducció poden estar basats en probabilitats entre paraules origen → destinació o entre grups de paraules origen → destinació (grups que poden ser simplement seqüències de dues o més paraules i que no necessàriament han de coincidir amb sintagmes sencers amb una estructura sintàctica concreta).

4.3.3. El descodificador

El descodificador és el "cor" del sistema de TA estadístic. És un algorisme implementat en forma de programa informàtic que fa servir els models de llengua i de traducció (que s'han derivat prèviament a partir de corpus monolingües i bilingües durant el procés d'aprenentatge) i genera seqüències de paraules en la llengua d'arribada a partir de seqüències de paraules en la llengua de partida. Hi ha diferents tècniques de descodificació; com a exemple, una d'aquestes tècniques és la de substituir cada paraula de la frase d'origen per la corresponent paraula més probable en la llengua d'arribada (fent ús d'un model de traducció basat en parells de paraules) i després intentar maximitzar la probabilitat de la seqüència de destinació resultant segons el corresponent model de llengua aplicant iterativament accions com ara esborrar, inserir o canviar paraules, canviar-les d'ordre, ajuntar-les, etc. Vegeu l'exemple de la figura següent:

Figura 6. Exemple de descodificació



4.4. Sistemes híbrids

4.4.1. Què és un sistema híbrid?

S'anomenen *sistemes híbrids* els sistemes de TA que combinen tècniques lingüístiques (RBMT) amb tècniques estadístiques (SMT).

El perquè d'aquesta combinació de tècniques lingüístiques i estadístiques rau en el fet que gran part dels errors típics d'ambdós sistemes són complementaris entre ells. Per exemple, els sistemes estadístics donen molt bons resultats de traducció en aspectes relacionats amb la selecció lèxica o la naturalitat de les frases generades en la llengua d'arribada, precisament dos dels punts febles dels

sistemes basats en regles. Per contra, punts febles dels sistemes estadístics, com ara el tractament correcte de fenòmens morfològics i sintàctics –concordança, generació de les formes flexives correctes de noms, adjectius i verbs, etc.–, són tractats adequadament pels sistemes basats en regles.

4.4.2. Comparativa breu entre sistemes de TA lingüístics i sistemes de TA estadístics

La taula següent il·lustra els punts forts i els punts febles dels sistemes basats en regles lingüístiques i dels sistemes estadístics.

Taula 1. Punts forts i punts febles dels sistemes basats en regles lingüístiques i dels sistemes estadístics

	RBMT	SMT
Tractament de fenòmens sintàctics[5]	++	--
Tractament de fenòmens morfològics[6]	++	--
Selecció lèxica[7]	--	++
Llegibilitat de la traducció[8]	-	+
Sensibilitat a la qualitat de l'entrada[9]	-	+
Dependència de l'existència de corpus monolingües i bilingües extensos i de qualitat[10]	++	--
Dependència d'equips de desenvolupament especialitzats[11]	--	++
Rapidesa de desenvolupament[12]	--	++ (*)
Adequació per a la postedició[13]	+	-

(5)Com ara dependències i concordances de llarga distància (per exemple, *El senyal que vam veure ahir al vespre quan sortíem de casa era vermell → La señal que vimos ayer por la noche cuando salíamos de casa era roja*), generació d'auxiliars verbals, etc.

(6)Concordança de gènere/nombre/cas entre nom, adjectiu i/o determinants, de persona/nombre entre verb i subjecte, etc.

(7)*cap → cabeza/jefe/cabo/ninguno/hacia, taula → mesa/tabla*, etc.

(8)En seguir models de llengua, les traduccions generades pels sistemes estadístics *sonen* més "naturals" que les generades pels sistemes lingüístics, tot i que pot ser que les traduccions no siguin de vegades correctes (és a dir, no diguin allò que diu la frase original). Aquesta millor llegibilitat, juntament amb el millor tractament de la selecció lèxica, són dos factors clau que fan que la percepció dels usuaris "a primera vista" sigui que els sistemes estadístics tradueixen millor que els sistemes basats en regles. Tanmateix, una comparació més acurada de la qualitat d'ambdós tipus de sistemes dóna una major qualitat als sistemes basats en regles, sobretot per a dominis no restringits.

(9)Els sistemes basats en regles són molt més sensibles a errors ortogràfics i/o gramaticals o de puntuació a les frases del text d'origen.

(10)Una condició *sine qua non* per poder tenir un bon sistema de TA estadístic és disposar de corpus monolingües i bilingües extensos i de bona qualitat. Sovint aquests corpus no estan disponibles per a llengües o parells de llengües minoritaris.

(11)El desenvolupament i manteniment dels sistemes basats en regles només els poden fer personal molt especialitzat amb una bona formació en lingüística computacional.

(12)Els sistemes estadístics –un cop fet el programa nucli– permeten desenvolupar nous parells de llengües molt ràpidament, amb la condició que hi hagi corpus monolingües i bilingües extensos per a les llengües en qüestió.

(13)Els errors comesos pels sistemes basats en regles solen ser predictibles i sistemàtics, mentre que els que fan els sistemes estadístics no ho són. A més, la més bona llegibilitat de les frases generades pels sistemes estadístics sovint emmascara errors en la traducció. Això fa que la gent que es dedica a corregir traducció automàtica normalment prefereixi corregir traduccions provinents de sistemes basats en regles.

4.4.3. Mètodes d'hibridació

La combinació de tècniques lingüístiques i tècniques estadístiques es pot dur a terme de diferents maneres. A continuació, n'exposem les més habituals.

Ampliació de sistemes RBMT amb tècniques estadístiques

En aquest cas, partim d'un sistema de TA basat en regles lingüístiques i intentem millorar-ne el rendiment fent servir mecanismes estadístics en alguns dels seus mòduls o fases de traducció. En donem alguns exemples:

1) Postedició estadística

Es tracta d'entrenar un sistema estadístic amb un corpus bilingüe aconseguit alineant les frases traduïdes pel sistema basat en regles, d'una banda, i les mateixes frases corregides (o ja correctes), d'una altra banda. El sistema estadístic així entrenat faria de "posteditor automàtic" de les traduccions generades pel sistema de TA lingüístic.

2) Generació de lèxics bilingües amb tècniques estadístiques

La creació de lèxics bilingües amb terminologia sobre un nou domini específic o per a un nou parell de llengües és una de les tasques que requereix més esforç de desenvolupament en els sistemes basats en regles. L'ús de tècniques estadístiques per extreure lèxics bilingües o llistes bilingües de termes a partir de corpus bilingües alineats pot ajudar a reduir dràsticament aquest esforç.

3) Gramàtiques d'anàlisi estocàstiques

En aquest cas, la idea és aplicar models de llengua basats en corpus d'arbres sintàctics de la llengua de partida. D'aquesta manera, es podran assignar probabilitats a les regles sintàctiques que fa servir el sistema RBMT per tal de construir l'arbre sintàctic corresponent a la frase d'entrada. Així es podria augmentar el percentatge d'arbres sintàctics correctes obtinguts i, per tant, la qualitat de la traducció.

4) Millora de la selecció lèxica amb tècniques estadístiques

Com s'ha esmentat abans, la selecció lèxica és un dels punts febles dels sistemes basats en regles. Fent ús de models de traducció estadístics es pot millorar la precisió d'aquesta selecció lèxica. La idea és afegir noves condicions de selecció lèxica a les entrades bilingües del sistema basat en regles, a més de les que ja hi pugui haver sobre informació morfològica, sintàctica o semàntica. Aquestes noves condicions en una entrada bilingüe $P_o \rightarrow P_d$ accedirien a un model de traducció estadístic derivat de corpus bilingües i retornaria la probabilitat que la paraula P_o tingui com a traducció P_d. Resta oberta la qüestió de sobre quines condicions lingüístiques han de ser prioritàries les condicions de selecció estadístiques.

5) Millora de la llegibilitat amb tècniques estadístiques

Com també s'ha dit abans, la llegibilitat de les traduccions generades pels sistemes basats en regles és un altre dels punts febles d'aquests sistemes. L'ús de models de llengua probabilístics derivats de corpus monolingües pot ajudar a millorar aquesta llegibilitat, per exemple en allò que té a veure amb l'ordre de les paraules.

Ampliació de sistemes SMT amb tècniques lingüístiques

1) Millora de la sortida d'un SMT amb tècniques lingüístiques

En aquest cas, la idea és fer servir un *parser* sintàctic i un conjunt de regles que conformin una gramàtica de la llengua d'arribada i que validi sintàcticament les diferents alternatives de sortida ofertes pel sistema estadístic.

2) Models de llengua basats en arbres sintàctics

Abans hem explicat què és el *model de llengua* d'un SMT. En principi, aquests models de llengua solen estar basats en seqüències de paraules (*n*-grames), però també poden estar basats en arbres sintàctics. Això significa fer servir corpus d'arbres sintàctics correctes per a una llengua en concret per tal de derivar-ne un conjunt de probabilitats sobre com és de probable que un arbre sintàctic corresponent a la frase d'entrada sigui o no correcte.

Vegeu també

Al subapartat 4.3.1 s'ha tractat breument el *model de llengua* d'un SMT.

3) Models de traducció basats en arbres sintàctics

De la mateixa manera que en el punt anterior, es poden fer construir models de traducció basats en arbres sintàctics. Aquests models computen la probabilitat que un arbre sintàctic A_o corresponent a una frase de la llengua de partida tingui com a "traducció" un arbre sintàctic A_d corresponent a la llengua d'arribada.

4.5. Sistemes basats en exemples (EBMT)

Els sistemes de TA basats en exemples[14] fan servir el principi de l'analogia com a factor clau del seu funcionament. Aquests sistemes intenten trobar patrons de similitud entre un gran nombre de frases de la llengua de partida amb les seves corresponents traduccions en la llengua d'arribada a partir dels quals es puguin derivar noves traduccions. Per exemple, a partir de dos exemples de traducció entre el català i l'alemany, com ara:

(14)En anglès, *Example Based Machine Translation* o EBMT.

Això és un llibre ← → Das ist ein Buch
El professor llegeix un diari ← → Der Lehrer liest eine Zeitung

Un sistema EBMT hauria de poder generar una traducció correcta en alemany per a la frase catalana (nova per al sistema):

El professor llegeix un llibre → Der Lehrer liest ein Buch

Es podria pensar que, en el fons, els sistemes EBMT són un subtipus de sistema de TA basat en tècniques estadístiques (SMT), ja que, tal com passa en aquests sistemes:

- Fan servir corpus bilingües extensos.
- Hi ha una fase prèvia d'entrenament a partir d'aquests corpus.
- El sistema deriva les traduccions a partir d'aquests corpus i sense fer servir informació lingüística.

Tanmateix, els sistemes basats en exemples (EBMT) no fan servir tècniques estadístiques i això els diferencia radicalment dels SMT. Com hem dit abans, els sistemes EBMT tenen com a base el principi de l'analogia com a factor clau en el procés de traducció. Segons el pare d'aquest tipus de sistemes, el japonès Makoto Nagao:

> Nagao (1984): "A student memorizes the elementary English sentences with the corresponding Japanese sentences. The first stage is completely a drill of memorizing lots of similar sentences and words in English, and the corresponding Japanese. Here we have no translation theory at all to give to the student. He has to get the translation mechanism through his own instinct. He has to compare several different English sentences with the corresponding Japanese. He has to guess, make inferences about the structure of sentences from a lot of examples".
>
> **Traducció**
>
> "Un estudiant que intenta aprendre japonès memoritza frases senzilles en la seva llengua materna juntament amb les corresponents frases en japonès. La primera etapa és, doncs, un exercici de memoritzar moltes frases i paraules similars en la seva llengua i en japonès. Aquí no li estem donant a l'estudiant absolutament cap teoria de la traducció. Aquest ha d'obtenir el mecanisme de la traducció a través del seu propi instint. Ha de comparar moltes frases diferents de la seva llengua amb les corresponents frases en japonès. Ha d'endevinar, ha de fer inferències sobre l'estructura de les frases a partir d'un munt d'exemples".

Malgrat que la teoria que hi ha darrere els sistemes de TA basats en exemples sigui racionalment impecable, és obvi que els problemes per dur-la a la pràctica són, sovint, insuperables. Certament, els sistemes EBMT són capaços d'oferir traduccions correctes per a frases curtes o amb estructures senzilles, però tenen problemes greus per derivar correspondències estructurals amb frases llargues o estructures morfosintàctiques complexes.

4.6. Sistemes basats en memòries de traducció

Les memòries de traducció no són, estrictament parlant, programes de traducció automàtica, ja que no realitzen un procés autèntic de traducció.

El nucli d'un programa d'aquest tipus és un corpus (anomenat precisament *memòria de traducció*), com més gran millor, de frases en la llengua de partida i d'un tema específic, juntament amb la seva traducció en la llengua d'arribada. El programa va mirant frase a frase el text que es vol traduir i intenta trobar a la memòria de traducció una frase igual, o si més no, que s'hi assembli en un grau predeterminat. Quan el resultat de la cerca és positiu, el programa simplement canvia la frase de l'idioma original per la seva traducció emmagatzemada.

Un cop aclarit aquest punt, podem extreure certes conclusions sobre aquests programes:

Les memòries de traducció funcionen bé quan s'han de traduir molts textos d'una mateixa temàtica molt semblants entre ells. L'exemple típic és la traducció de versions successives de manuals tècnics. En aquest cas, pot donar-se el cas que planes senceres de dues versions diferents del mateix manual siguin iguals. La idea és que si ja s'han traduït un cop, no cal traduir-les un altre cop.

Com més extenses siguin les memòries de traducció (com més frases hi hagi emmagatzemades juntament amb la seva traducció), millor serà el rendiment del programa.

Òbviament, un programa d'aquest tipus necessita una memòria de traducció inicial amb un nombre mínim de frases emmagatzemades, abans de no ser operatiu.

El grau de similitud que es fa servir en el procés de comparació entre les frases del text i les frases emmagatzemades és variable i pot ser definit per l'usuari. Normalment, el seu valor mínim sol ser d'un 80% (és a dir, almenys un 80% de les paraules d'ambdues frases han de ser les mateixes i ser al mateix lloc de la frase).

De vegades, es fan servir aplicacions híbrides entre memòries de traducció i tècniques de traducció automàtica "real". En el cas de frases que no siguin un 100% iguals a les que són a la memòria de traducció, les paraules que no coincideixen s'envien a traduir a un motor de traducció automàtica, mentre que la traducció de la resta de la frase (la part coincident) s'agafa de la traducció emmagatzemada.

4.7. Altres aproximacions al problema de la traducció automàtica

A banda dels tipus de sistemes que hem vist fins ara podem fer esment d'altres tipus d'aproximació a la traducció automàtica, com ara les xarxes neuronals. Tot i que la idea que hi darrera de les xarxes neuronals és la d'apropar-se més a la manera com funciona el nostre cervell, la simplicitat relativa d'aquestes simulacions –comparades amb la complexitat extrema de funcionament del cervell humà– i el fet que encara es trobin en una fase d'investigació fa que amb aquests tipus de sistemes, de moment, només sigui possible la traducció d'algunes frases en contextos molt restringits.

5. Els components d'un sistema de TA basat en coneixement lingüístic

A continuació presentarem els components que es troben a la majoria de sistemes de TA basats en regles lingüístiques i que fan servir el model de transferència. Hem agafat aquest tipus de sistema com a referència per explicar els mòduls components i el procés de la traducció automàtica perquè, ara per ara, és el que més qualitat ofereix en aplicacions de traducció automàtica comercials.

Vegeu també

El model de transferència s'ha tractat al subapartat 4.2.3.

Com a la majoria d'aplicacions informàtiques, els mòduls components d'un sistema de TA es poden classificar en dos grans grups: **dades** i **programes**.

És fonamental entendre la diferència entre dades i programes. Els **programes** són conjunts d'instruccions d'ordinador que **fan** coses (processen una informació d'entrada segons l'algorisme que tinguin programat i donen una informació de sortida com a resultat). Per exemple, un **segmentador** segmenta un text en frases (entrada: un text continu; sortida: la llista de frases que componen el text); un **analitzador sintàctic** construeix un arbre sintàctic a partir de la llista de paraules d'una frase (entrada: llista de paraules lematitzades i etiquetades morfològicament; sortida: un o més arbres sintàctics); un **generador** genera una frase a partir d'un arbre sintàctic (entrada: arbre sintàctic; sortida: frase corresponent).

Per poder fer aquestes coses, els programes necessiten la informació que hi ha emmagatzemada a les **dades:** un segmentador fa servir les **regles de segmentació de frases** d'una llengua determinada; un analitzador sintàctic fa servir les regles de la **gramàtica d'anàlisi** per saber com ha d'anar construint l'arbre sintàctic a partir de la llista de paraules d'una frase; un generador fa servir les regles de la gramàtica de generació per generar una frase a partir d'un arbre sintàctic.

5.1. Les dades

Les dades són la informació de la qual disposa el programa per dur a terme el procés de traducció. És a les dades on hi ha tot el coneixement lingüístic del sistema i, per tant, són aquestes les que bàsicament en determinen el grau de sofisticació.

Seguidament veurem amb una mica de detall les característiques fonamentals dels dos tipus principals de dades d'un sistema de TA: els lèxics i les gramàtiques.

5.1.1. Els lèxics

Implementació i característiques dels lèxics

Podem considerar que un sistema de TA típic entre una llengua de partida i una llengua d'arribada ha de tenir tres components lèxics:

- Un lèxic monolingüe de la llengua de partida.
- Un lèxic monolingüe de la llengua d'arribada.
- Un lèxic bilingüe entre la llengua de partida i la llengua d'arribada.

Cal deixar ben clara la diferència que hi ha entre un lèxic computacional (els usats als sistemes de TA) i un diccionari tradicional. A un diccionari monolingüe tradicional (en paper) cada entrada té una petita quantitat d'informació gramatical (la categoria gramatical, el gènere dels noms i, de vegades, indicacions sobre el model de flexió de noms, adjectius i verbs). La resta d'informació consisteix en una o més definicions de la paraula. Vegem, per exemple, l'entrada de la paraula *traducció* a un diccionari "tradicional" català:

> **traducció f.** ***1*** 1. Acció de traduir. 2 Obra traduïda. ***2*** ling 1 Reproducció del contingut d'un text o d'un enunciat oral, formulat en una llengua, en formes pròpies d'una altra llengua. 2 **traducció simultània** Traducció oral efectuada a mesura que és pronunciat un text en la llengua original. ***3***gen Procés mitjançant el qual es tradueix una determinada seqüència de nucleòtids de l'ARN missatger en una determinada seqüència d'aminoàcids gràcies al concurs de ribosomes. ***4*** inform Transformació de sentències des d'un llenguatge, origen, etc, a un altre, resultant.

Pel contrari, com ja veurem, els lèxics computacionals tenen molta més informació gramatical, però, en canvi, no tenen definicions per a les paraules. La raó és evident: els programes de traducció automàtica no en poden fer res, d'una definició en el sentit tradicional, però necessiten tota la informació gramatical possible, generalment formalitzada en forma de trets i valors.

Els lèxics monolingües

Típicament, un lèxic monolingüe d'un sistema de TA conté informació morfològica, sintàctica i semàntica per a cada paraula de la llengua de partida o de la llengua d'arribada. Aquesta informació està emmagatzemada en forma d'un conjunt de trets i valors. Cada tret pot tenir un o més valors d'un tipus predeterminat (numèric, booleà, cadena de caràcters, un element d'una llista). Tant el conjunt de trets lèxics com el tipus de valors que puguin tenir aquests trets solen estar especificats en un fitxer de definició d'estructura lèxica. Per exemple:

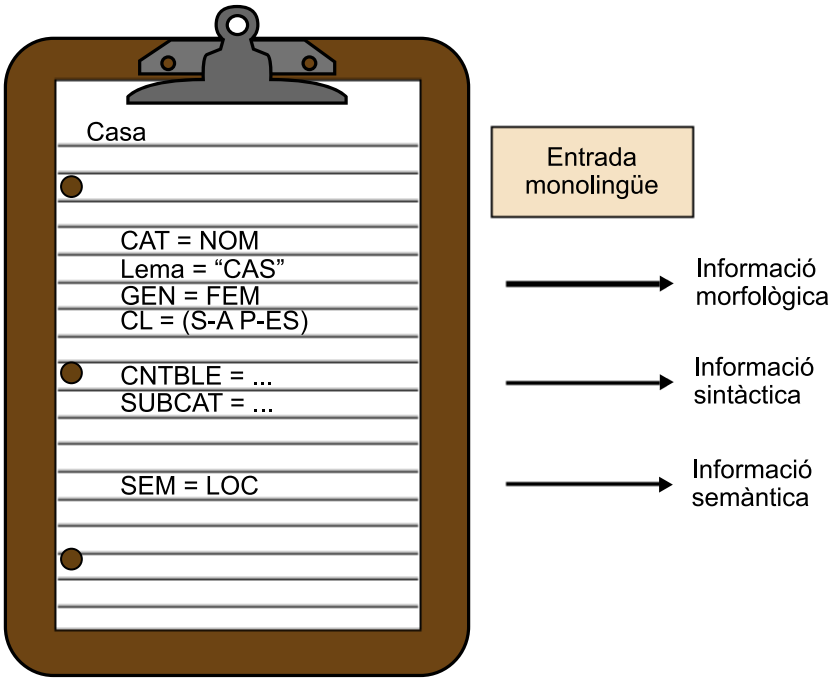
Taula 2

Tret	Categoria	Tipus	Valors
LEMA	NOM, ADJ, VERB...	Cadena	qualsevol cadena de caràcters
ARREL	NOM, ADJ, VERB...	Cadena	qualsevol cadena de caràcters

Tret	Categoria	Tipus	Valors
GN	NOM	Element	(M F)
NB	NOM	Element	(SG PL)
TERM	NOM, ADJ	Booleà	
CL	NOM, ADJ	Element	(SM-0, SF-0, PM-S, PF-ES...)

Podríem imaginar-nos cada entrada d'un lèxic monolingüe com una fitxa, el nom de la qual és la paraula a la qual correspon l'entrada. A la fitxa de cada paraula hi ha escrita diversa informació morfològica, sintàctica i semàntica:

Figura 7. Exemple de fitxa



A la taula següent podem veure aquesta informació amb més detall:

Taula 3

Tipus d'informació	Tret	Descripció
Morfològica	Categoria	Categoria gramatical: nom, adjectiu, verb...
	Classe flexiva	Model de flexió per a noms, adjectius o verbs
	Gènere	Gènere gramatical (noms)
	Nombre	Nombre gramatical (noms, adjectius, verbs)
	Persona	Persona gramatical (verbs, pronoms)
	Temps	Temps gramatical: present, passat, futur (verbs)
	Mode	Mode gramatical: indicatiu, subjuntiu... (verbs)

Tipus d'informació	Tret	Descripció
	Forma del predicat	Forma del predicat: finit, infinitiu, participi, gerundi (verbs)
	Derivació adverbial	Possibilitat de derivació adverbial per a adjectius: normal → normalment
Sintàctica	Subcategorització	Informació sobre els arguments (subjecte, objecte directe, objecte indirecte, objectes preposicionals, etc.) subcategoritzats per als verbs, i per a certs noms i adjectius.
	Comptable	Informació sobre si un nom es refereix a una entitat comptable (llibres, cases) o no comptables (aigua, fred).
	Cas governat	Cas governat per les preposicions
	Mode verbal exigit	Mode verbal exigit per certes conjuncions
	Ser/estar	Informació sobre si un adjectiu va sempre amb el verb *ser* (*ser japonès*), amb el verb *estar* (*estar cansat*) o amb tots dos (*ser/estar blanc*).
	Posició atributiva	Informació sobre si l'adjectiu en posició atributiva sol anar davant o darrere del nom.
Semàntica	Tipus de nom	Tipus del nom segons una tipologia semàntica específica (humà, animal, planta, temporal, lloc, procés, matèria, entitat concreta, etc.)
	Tipus d'adjectiu	Tipus d'adjectiu segons una tipologia semàntica específica (temporal, locatiu, de color, de propietat objectiva, etc.)
	Tipus de verb	Tipus de verb segons una tipologia semàntica específica (verb de moviment, de dicció, d'intenció, etc.)
	Tipus d'adverbi	Tipus d'adverbi segons una tipologia semàntica específica (adverbi de temps, de manera, de lloc, de direcció, etc.)

Així, l'entrada lèxica per a la paraula *llibre* podria tenir l'aspecte següent:

Taula 4

Tret	Valor
Lema	*llibre*
Arrel	*llibre*
Categoria	NOM
Classe flexiva	SM-0, PM-S
Gènere	MASC
Tipus semàntic	Semiòtic (entitats concretes que poden ser subjectes de verbs de dicció: *el llibre diu, indica, afirma...*)
Comptable	Sí

Vegem també un exemple d'entrada lèxica monolingüe per a la paraula *cantar:*

Taula 5

TRET	VALOR
Lema	*cantar*
Arrel	*cant*
Categoria	VERB
Classe flexiva	AR
Subcategorització	Subjecte nominal humà, objecte directe nominal abstracte opcional objecte indirecte nominal humà opcional "algú canta (alguna cosa) (a algú)"

Els lèxics bilingües

El lèxic bilingüe (també anomenat *de transferència*) conté entrades lèxiques que relacionen entrades del lèxic monolingüe d'origen amb entrades del lèxic monolingüe de destinació. En definitiva, és aquí on són emmagatzemades les traduccions de cada paraula coneguda pel sistema.

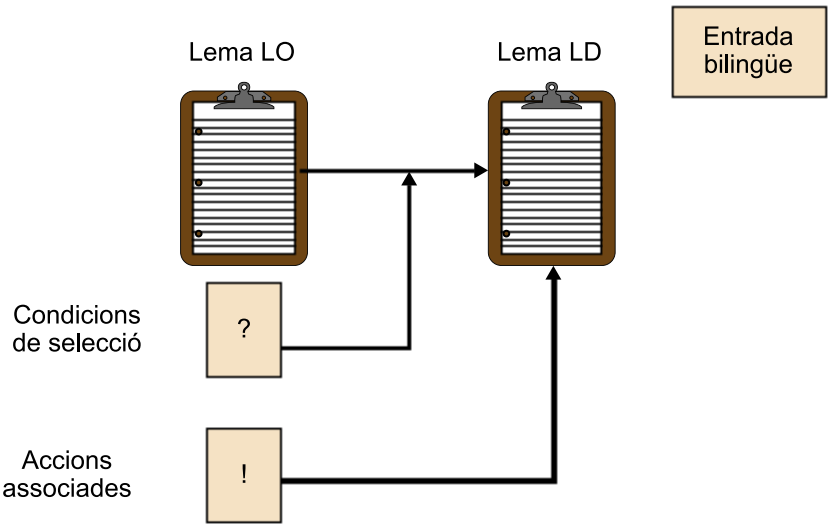
El lèxic bilingüe seria una simple llista de correspondències "paraula-origen" → "paraula-destinació" si no fos per la coneguda tendència de les paraules de tota llengua humana a tenir més d'una traducció en altres llengües: només cal pensar en les diverses traduccions possibles de paraules catalanes com ara *cap, porta, volen, pot* o *tens*.

La selecció de la traducció adequada per a cada paraula pot dependre de certes condicions sobre el context morfosintacticosemàntic de la paraula a la frase en què apareix. Algunes d'aquestes condicions fan referència a informació lingüística que pot estar present a l'arbre d'anàlisi i es poden formalitzar en forma de crides a funcions que s'activen des de les entrades bilingües.

A més, quan se selecciona una traducció determinada pot ser necessari realitzar un canvi en l'estructura sintàctica de la frase (pensem en l'exemple del verb anglès *to love* quan la seva traducció en català és *agradar*). Aquestes transformacions o accions associades també es poden formalitzar en forma de crides a funcions que s'activen des de les entrades bilingües.

En resum, podem representar una entrada del lèxic bilingüe com una fitxa amb l'estructura següent:

Figura 8. Exemple de fitxa per a una entrada de lèxic bilingüe



Hem de tenir en compte que no totes les entrades lèxiques tenen condicions d'aplicació o accions associades. Hi ha moltes entrades bilingües que no necessiten cap condició de selecció (perquè només hi ha una traducció possible) ni cap acció associada (perquè no cal realitzar cap transformació morfosintàctica). Aquest seria el cas de, per exemple, les entrades bilingües ***cervesa → cerveza, groc → amarillo*** o ***desaparèixer → desaparecer.***

5.1.2. Les gramàtiques

Implementació i característiques de les gramàtiques

Tal com hem vist que passava amb els lèxics, les gramàtiques que es fan servir als sistemes de TA (gramàtiques computacionals) tenen poc a veure amb les gramàtiques tradicionals en format de llibre. Encara que l'objectiu dels dos tipus de gramàtiques és el mateix (oferir una descripció tan acurada com sigui possible de com funciona una llengua determinada), la forma és radicalment distinta.

Una gramàtica computacional és una descripció formal dels diferents fenòmens morfosintàctics d'una llengua, expressada amb l'ajuda d'un formalisme i d'un llenguatge de programació associat. Cada sistema fa servir el seu propi formalisme i el seu propi llenguatge d'implementació. Aquest formalisme pot coincidir en alguns casos amb el de teories lingüístiques conegudes (per exemple, GPSG, HPSG, GB, LFG, etc.) o pot ser un formalisme *ad hoc* dissenyat expressament per al sistema en qüestió.

En qualsevol cas, el tipus de sistemes que estem considerant (basats en transferència) han de tenir una gramàtica d'anàlisi, per poder dur a terme l'anàlisi sintàctica de les frases en la llengua de partida, i una gramàtica de generació, per generar les frases traduïdes en la llengua d'arribada.

Gramàtica d'anàlisi

La gramàtica d'anàlisi és la part del sistema on resideix tota la informació disponible sobre com funciona la morfologia i la sintaxi de la llengua de partida. Com més sofisticat sigui un sistema de TA (és a dir, com més "alt" es pugui col·locar al triangle dels tipus de sistemes que hem vist anteriorment), més complexa serà la seva gramàtica d'anàlisi.

Una bona gramàtica d'anàlisi ha de tenir quatre característiques fonamentals:

a) Ser exhaustiva: ha de ser capaç d'analitzar tantes estructures sintàctiques com sigui possible.

b) Ser precisa: els arbres sintàctics construïts i la informació lingüística calculada per a cada frase n'han de representar fidelment l'estructura lingüística.

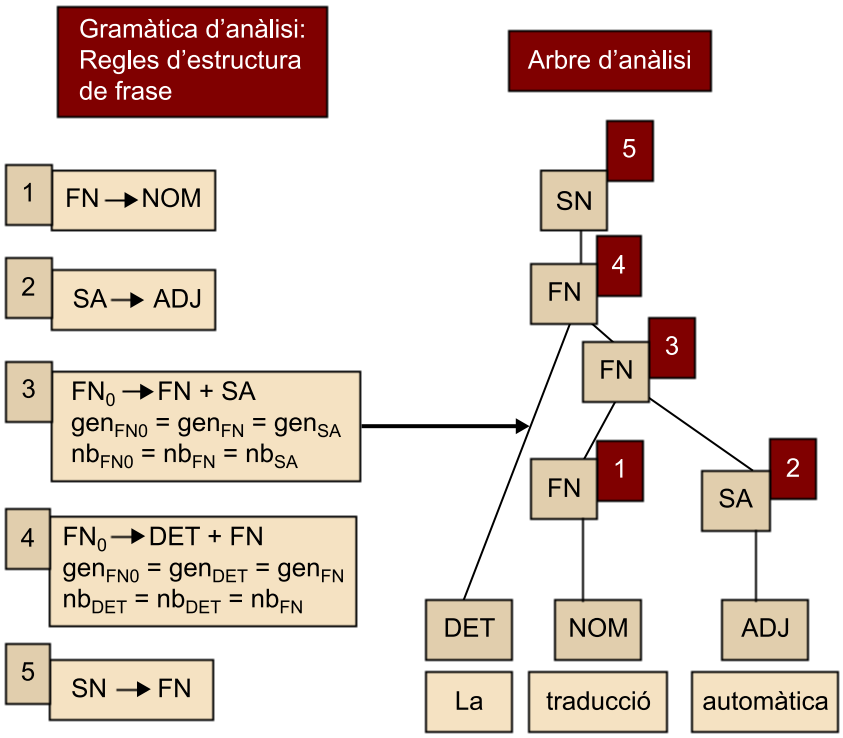
c) Ser robusta: atès que no és possible cobrir totes les estructures sintàctiques que ens poden arribar (sobretot en el cas de frases molt llargues, molt "enrevessades" o, simplement, mal escrites), la gramàtica ha de tenir mecanismes per assegurar que, en el cas de no poder arribar a una interpretació correcta per a una frase, es puguin construir almenys arbres parcials o alternatius, per tal que les fases de transferència i generació tinguin algunes dades amb què treballar.

Robustesa

Un sistema de TA no es pot aturar quan "no entén" una frase!

d) Ser flexible: les gramàtiques han d'estar dissenyades i implementades de tal manera que la seva actualització, el manteniment i la millora sigui tan fàcil com sigui possible.

Figura 9. Exemple de gramàtica



Gramàtica de generació

Les gramàtiques de generació són un reflex de les gramàtiques d'anàlisi (de fet, com ja hem esmentat, en alguns sistemes es fa servir la mateixa gramàtica per analitzar i per generar). Aquestes gramàtiques han de contenir la informació necessària per, a partir dels arbres sintàctics obtinguts a la fase d'anàlisi, i un cop realitzades la transferència lèxica i la transferència estructural, poder generar correctament la frase en la llengua d'arribada.

5.2. Els programes

5.2.1. Els analitzadors morfològics

Un analitzador morfològic agafa com a entrada les paraules del text, una a una i tal com apareixen al text d'entrada, i treu com a sortida una llista de paraules lematitzades i etiquetades morfològicament. És a dir, per a cada paraula obté un conjunt d'interpretacions morfològiques que contenen un lema i la informació morfològica que se'n pugui deduir de la forma. Normalment és un programa que fa servir dades emmagatzemades al lèxic monolingüe de la llengua de partida.

5.2.2. Els analitzadors sintàctics

Els analitzadors sintàctics (també coneguts pel terme anglès *parser*) són programes que agafen com a entrada la llista de paraules lematitzades i etiquetades morfològicament que componen la frase d'entrada (és a dir, la sortida de l'analitzador morfològic) i, amb la informació del lèxic monolingüe de la llengua de partida i, sobretot, amb les dades de la gramàtica d'anàlisi, construeixen una estructura en forma d'arbre que reflecteix l'estructura de constituents de la frase.

6. El procés de la TA basada en coneixement lingüístic

6.1. Adquisició i preparació del text

El primer pas per poder traduir un document amb un sistema de TA és tenir-lo en suport magnètic i en un format que sigui llegible pel sistema en qüestió. La primera condició és cada cop més habitual, tot i que molts documents encara existeixen només en paper. En aquest cas, s'han d'escanejar i passar per un programa OCR, per tal d'obtenir-ne la versió en suport magnètic (òbviament, hi ha una alternativa al pas "escaneig + OCR", que és picar el text a un processador de textos).

Hi ha tres formats de document que són acceptats per gairebé la majoria de sistemes de TA: ASCII, RTF (Microsoft Word) i HTML (pàgines web).

6.2. Segmentació de frases

Normalment, els sistemes de TA duen a terme la traducció frase a frase. És per això que un cop s'ha enviat a traduir el document en un dels formats acceptats pel sistema de traducció, el pas següent és dividir el text en frases. Encara que no ho sembli, aquesta tasca no és trivial. No tots els punts delimiten una frase, com podem veure en aquest exemple:

El Dr. Pons li va receptar 3 mg de preparat cada hora.

Hem de tenir en compte que els punts poden seguir abreviatures (Dr., m., s., etc.), xifres (1.345.213 PTA) o números de capítol o secció d'un llibre (2.3.1), entre d'altres.

És molt important que la segmentació de frases sigui correcta. En cas contrari, l'anàlisi sintàctica i, per tant, la traducció poden ser incorrectes.

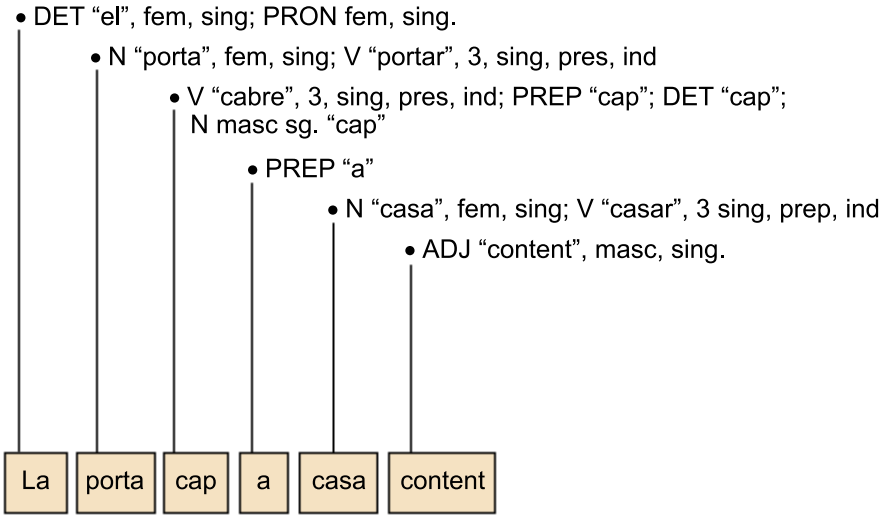
6.3. Anàlisi morfològica

A partir d'aquest punt treballarem amb la frase com a unitat de traducció. Com a exemple en els diferents passos que veurem a continuació farem servir la frase següent, considerant la seva traducció del català al castellà:

La porta cap a casa content.

L'anàlisi morfològica és el primer procés lingüístic pel que passa la frase que s'ha de traduir. Per a cada paraula, l'analitzador morfològic dóna totes les possibles segmentacions i interpretacions. Per dur a terme aquesta tasca, l'analitzador morfològic fa servir la informació del lèxic monolingüe de la llengua de partida (en aquest cas, del lèxic monolingüe català).

Figura 10. Exemple d'anàlisi morfològica



Segons podem veure a l'exemple de la figura, l'anàlisi morfològica ofereix diverses possibilitats per algunes paraules de la nostra frase:

- ***la*** pot ser un article determinat o un pronom personal.

- ***porta*** pot ser una forma del verb *portar* (*ell/ella porta*) o el nom femení singular *porta*.

- ***cap*** (una d'aquelles parauletes catalanes que "porta de cap" els sistemes de TA!) pot ser un munt de coses: un nom (*el **cap***, que, per cert, té fins a tres traduccions diferents en castellà: *cabeza, jefe* i *cabo*), una forma del verb *cabre* (*ell/ella **cap***), un determinant nominal (***cap** llibre*) que pot funcionar com un pronom (*no n'hem trobat **cap***) i una preposició (*anem **cap** allà*).

- casa pot ser un nom (*la **casa** on vivim*) o una forma del verb casar (*En Pere es **casa** demà*).

És important tenir present que, en aquesta fase, l'anàlisi morfològica ens dóna **totes** les possibilitats per a cada paraula **vista individualment**, encara que moltes de les combinacions que en resulten quan es mira la frase sencera no siguin possibles (per exemple, la combinació *la* (*pronom*) *porta* (*nom*) *cap* (*nom*) *a casa* (*verb*) és impossible en català).

En resum, el resultat de l'anàlisi morfològica és una llista d'interpretacions per a cada paraula de la frase. Cada interpretació és encapçalada per un lema (p. ex. *portar* per a la interpretació verbal de *porta*), seguit de la categoria gramatical

(nom, verb, adjectiu, etc.) i de la informació morfològica que es pugui deduir de la forma de la paraula analitzada (gènere, nombre, temps, persona, mode, etc.).

6.4. Anàlisi sintàctica

L'anàlisi sintàctica té com a entrada la llista de paraules de la frase amb totes les interpretacions resultants de l'anàlisi morfològica. A partir d'aquesta informació, l'analitzador sintàctic, guiat per les regles de la gramàtica d'anàlisi, intenta construir un arbre que representi adequadament l'estructura de constituents de la frase i la informació sintàctica associada (dependències nucli-complements, funcions sintàctiques, dependències anafòriques, etc.).

Sens dubte, l'anàlisi sintàctica és la tasca més difícil en el procés de traducció. La gramàtica d'anàlisi ha de ser prou potent i robusta per fer front a les ambigüitats provinents de l'anàlisi morfològica i a les ambigüitats sintàctiques que es presenten segons es van analitzant parts de la frase. Pensem què passa quan l'analitzador intenta construir arbres vàlids per a la frase

La porta cap a casa

que és un fragment de la nostra frase exemple, sense la darrera paraula, *content*. Aquesta frase és sintàcticament ambigua en català, i pot tenir dues interpretacions:

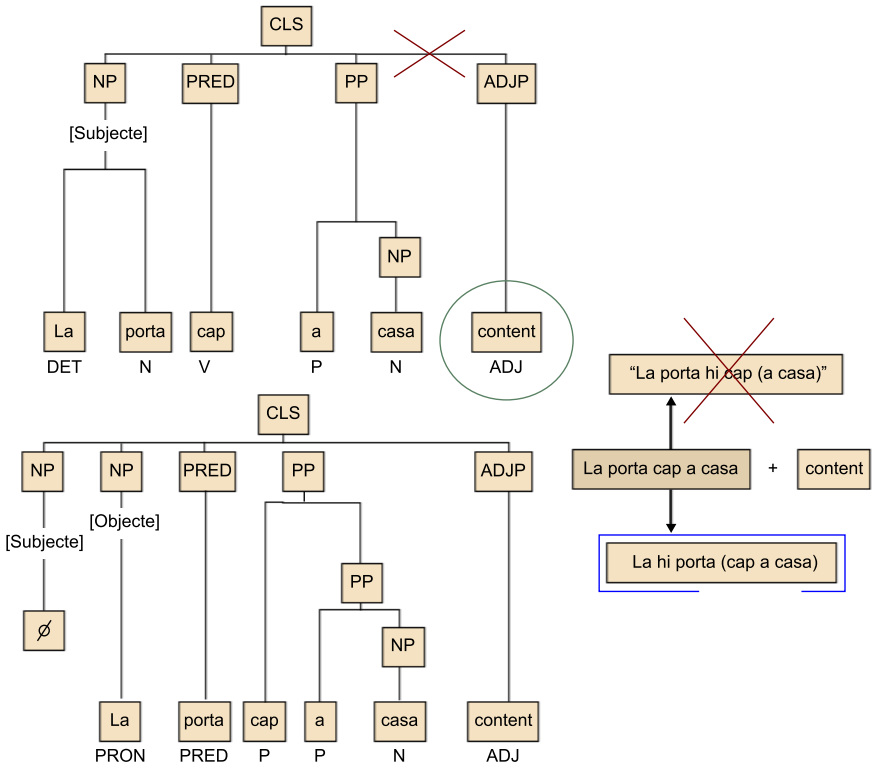
a) *La porta hi cap* (*a casa*) (*La puerta cabe en casa*).

b) *La hi porta* (*cap a casa*) (*la lleva hacia casa*).

En aquest moment l'analitzador sintàctic ha de tenir en compte totes dues interpretacions i ha de construir un arbre sintàctic per a cadascuna d'elles.

És quan l'analitzador consumeix la paraula següent de la nostra frase, *content*, que la gramàtica d'anàlisi catalana ha de ser capaç de rebutjar la interpretació (**a**) i seguir només amb la (**b**):

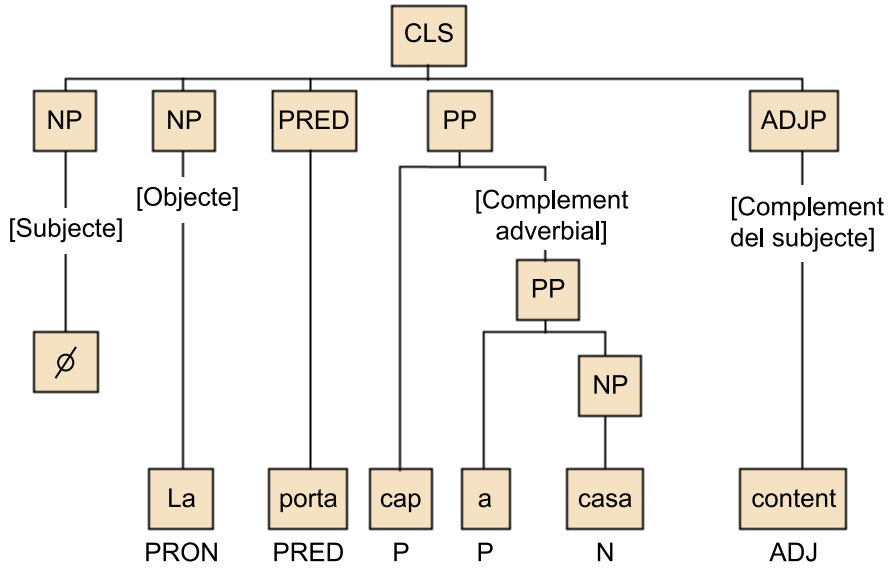
Figura 11



La raó és que l'adjectiu masculí *content* no pot modificar el nom femení que té a la seva esquerra, *casa*, i només pot funcionar com a complement del subjecte de la frase (que, de retruc, sabem que ha de ser masculí, ja que ha de concordar amb *content*): (*Ell*) *la porta cap a casa content.*

Així, doncs, l'arbre final resultant de l'anàlisi sintàctica és el següent:

Figura 12



Cada node d'aquest arbre conté informació lèxica o sintàctica, part de la qual ha heretat dels seus nodes fills, i part de la qual s'ha calculat durant el procés d'anàlisi sintàctica. Així, per exemple, el node PP (*prepositional phrase*, o sintag-

ma preposicional) que està directament sota del node superior (CLS) té informació sobre el gènere del seu nucli nominal *casa:* femení. Aquesta informació prové de l'entrada lèxica per al nom *casa* al lèxic monolingüe català, i durant la construcció de l'arbre sintàctic, s'ha anat copiant des del node terminal N *casa*, fins al node PP, a través dels nodes intermedis NP, PP. En aquest cas, es parlaria d'informació heretada des dels nodes fills. D'altra banda, la informació sobre la funció sintàctica d'aquest mateix node PP (*cap a casa* es tracta d'un complement adverbial de direcció del verb *portar*) es calcula des del node superior de l'arbre segons la informació sobre subcategorització argumental present al node verbal PRED. Aquest seria un exemple d'informació calculada.

6.5. Transferència lèxica

Un cop tenim l'arbre sintàctic, passem a la fase de transferència lèxica. En aquesta fase el sistema accedeix al lèxic bilingüe català-castellà per tal de seleccionar la traducció adequada per a cada paraula.

Com ja hem vist al subapartat dedicat als lèxics, les entrades del lèxic bilingüe poden tenir condicions d'aplicació i, de vegades, accions associades. Aquestes condicions de selecció fan referència a informació que ha d'estar present a l'arbre d'anàlisi. Per exemple, el verb *portar*, que surt a la nostra frase exemple, pot tenir diverses traduccions en castellà, entre d'altres, *llevar* i *traer.* Les condicions de selecció de l'una o de l'altra no són fàcils d'establir, però podríem donar-ne dues:

- Si a la frase el verb principal de la qual és *portar*, hi apareix un complement adverbial de direcció que conté l'adverbi *aquí* (p. ex. *cal portar-lo aquí*) o un pronom personal o possessiu de primera persona (p. ex. *cal portar-lo cap a nosaltres*), el programa selecciona la traducció castellana *traer.*

- Si la frase té un complement adverbial de direcció que no compleixi les condicions del punt anterior, el programa selecciona la traducció castellana *llevar.*

En aquest punt, resulta evident que la selecció de la traducció adequada depèn, de manera absoluta, que l'arbre sintàctic hagi estat correctament construït i que la informació que tingui emmagatzemada als seus nodes sigui igualment correcta.

Així, doncs, és en aquesta fase quan es fa la traducció de les paraules de la frase. En el nostre cas:

Taula 6

ell	⇒	él
portar	⇒	llevar

cap a	⇒	hacia
casa	⇒	casa
content	⇒	contento

Com veiem, el que es tradueix en aquesta fase són els lemes de les paraules. Serà més endavant, a la fase de generació quan s'obtindrà la forma correcta per a cada paraula, segons la informació que hi hagi als nodes corresponents sobre gènere, nombre, cas, persona, mode, temps, etc.

6.6. Transferència estructural

Algunes entrades lèxiques del lèxic bilingüe o de transferència tenen, a més de condicions de selecció, accions associades a la selecció de l'entrada. Aquestes accions poden estar referides a canvis de gènere o nombre (a l'entrada lèxica que relaciona el nom català ***diner*** amb el nom castellà ***dinero***, s'especifica que, si el nom apareix en plural en català, en castellà s'ha de posar en singular), però també poden referir-se a canvis estructurals activats per l'entrada. Així, una entrada bilingüe per al verb ***fer*** amb la condició de selecció que el verb porti un complement directe nominal amb el nom ***esternut*** té com a traducció castellana el verb ***estornudar*** (***fer un esternut → estornudar***). Aquesta entrada porta una acció associada consistent a esborrar de l'arbre el sintagma nominal que conté ***esternut***. Si no ho féssim així, la traducció resultant seria ****estornudar un estornudo.***

6.7. Generació de la frase traduïda

Entrem ara en l'última fase de la part lingüística del procés de traducció. Partim de l'arbre d'anàlisi amb les paraules dels nodes terminals traduïdes al castellà (transferència lèxica) i amb certes modificacions estructurals activades per algunes entrades del lèxic bilingüe (transferència estructural). A partir d'aquí, s'activen les tasques que condueixen a la generació de la frase final en la llengua d'arribada (en el nostre cas, el castellà).

Podem repetir les tasques més importants realitzades durant la fase de generació:

- Col·locar les paraules de la frase segons les regles de l'ordre de constituents a la llengua d'arribada.

- La inserció o esborrament de material lèxic (per exemple, la inserció de la preposició *a* davant dels objectes directes animats (*vaig veure el teu pare → vi **a** tu padre*) o l'esborrament en certs casos dels pronoms febles catalans *en* i *hi* quan estem traduint del català al castellà).

- La generació de les formes flexives adequades de les paraules de la frase de sortida, segons la informació present a l'arbre (p. ex. ***Llevar*** {3a. persona,

singular, present d'indicatiu} ⇒ ***lleva***, o el pronom ***él*** {femení, singular, acusatiu} ⇒ ***la***). Aquesta tasca és la que passa dels lemes a les formes correctes de cada paraula en la llengua d'arribada.

Els ajustos necessaris en cas de canvi de gènere en un nom de la llengua d'arribada respecte de la llengua de partida. Per exemple, la frase catalana ***L'anàlisi sintàctica és molt complicada*** es tradueix en castellà per ***El análisis sintáctico es muy complicado.*** Com podem veure, el canvi de gènere femení → masculí no només afecta les paraules que són al voltant del nucli nominal *anàlisi*/*análisis* (l'article *l'*/*el* i l'adjectiu *sintàctic*/*sintáctico*) sinó també un altre adjectiu (*complicat*/*complicado*) que, encara que sigui al final de la frase, realitza funcions de complement del subjecte i, per tant, ha de concordar amb el subjecte de la frase, és a dir, amb *anàlisi*.

La combinació i contracció d'elements lèxics (p. ex. en castellà, ***da le lo*** ⇒ ***dáselo***, o ***de el*** ⇒ ***del***).

La fase de generació accedeix al lèxic monolingüe de la llengua d'arribada.

En el nostre cas exemple, la frase final generada en castellà seria:

La lleva hacia casa.

6.8. Reposició del text original

Un cop tenim traduïdes totes les frases del text original, només resta tornar a posar aquestes frases al seu lloc (respectant la disposició de paràgrafs, taules, columnes, peus de pàgina, etc. del document original) i crear-ne un fitxer amb el format del document d'entrada (ASCII, RTF, HTML, etc.).

6.9. Correcció del text traduït

Aquesta és l'última tasca en el procés de la traducció automàtica, i és precisament l'única tasca que no és (i no pot ser!) automàtica i que ha de ser feta forçosament per correctors humans.

L'humà mana

El corrector humà té l'última paraula en el procés de traducció!

Tornem a repetir que un sistema de TA és només una eina que ens pot ajudar a fer traduccions més ràpidament. Tanmateix, atès que la qualitat de traducció que donen aquests sistemes és, en el millor dels casos (traducció entre llengües molt properes), no superior a un 90%, és el corrector humà l'últim responsable de la qualitat final del text traduït.

7. Aplicacions de la TA

Els sistemes de TA es fan servir normalment per a dos propòsits ben diferenciats:

- Aplicacions orientades a la comprensió de textos: l'objectiu és entendre el contingut d'un text escrit en una llengua que no coneixem.

- Aplicacions orientades a agilitar el procés de traducció massiva de textos: l'objectiu és traduir més ràpidament grans volums de textos entre dues llengües.

7.1. Aplicacions de comprensió

En aquest cas es tracta de fer servir el sistema de TA com una eina que ens permeti saber de què va un text escrit en una llengua que no comprenem. Els usos de sistemes de TA per a aplicacions de comprensió tenen alguns punts en comú:

- Normalment es tradueixen textos curts (entre 1 i 10 planes).

- Les llengües entre les quals es fa la traducció no són llengües properes entre elles (a poca gent se li acudiria fer servir un sistema de TA per traduir un text del castellà al català per tal de saber el que diu el text, però sí que ho faríem si el text fos escrit en alemany, rus o japonès).

- La traducció es fa mitjançant programes per a ordinadors personals, o bé programes de TA que funcionen a Internet. Fins ara, aquests sistemes es poden situar dins el triangle dels tipus de sistemes a la part de traducció directa o, com a molt, a la part baixa de la traducció basada en transferència.

- La qualitat de la traducció resultant (que normalment és baixa) no sol ser crítica, ja que l'objectiu és més aviat tenir una idea general del que diu el text original.

- Normalment no hi ha un procés de correcció de la traducció oferta pel sistema.

7.2. Aplicacions de traducció massiva

En aquest altre tipus d'aplicació el sistema es fa servir com a eina industrial per obtenir traduccions de grans volums de textos tan ràpidament com sigui possible. Un exemple típic seria una agència de premsa que vulgui oferir les notícies en dues llengües simultàniament. Les aplicacions de traducció massiva també tenen certs punts en comú:

- La rapidesa i la qualitat són factors crítics. Com més ràpidament es faci la traducció i menys s'hagi de corregir el resultat, millor.

- Les llengües entre les quals es fa el procés de traducció han de ser molt properes (català-castellà, castellà-gallec, alemany-holandès). La raó és que en l'estat actual de la tecnologia de la TA no és possible oferir una qualitat acceptable (econòmicament rendible) per a aplicacions de traducció massiva entre llengües no properes (el límit acceptable podrien ser parells de llengües com ara el català-anglès o el castellà-francès).

- Els programes de TA que es fan servir solen estar basats en tècniques de transferència i sovint s'han d'adaptar a entorns informàtics i de gestió de documents específics.

7.3. Tipus de sistemes de TA i aplicacions

Hi ha tres factors principals que determinen el marc d'aplicacions "raonables" per a un sistema de TA:

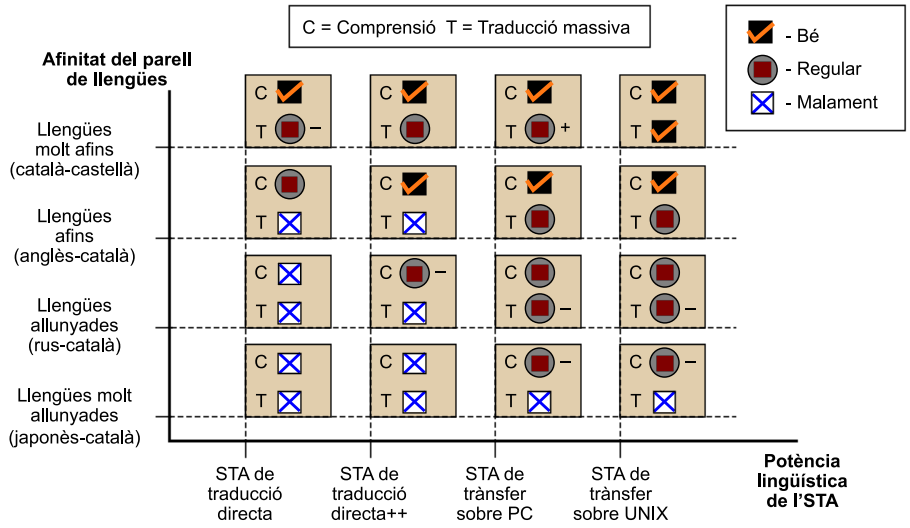
- La potència lingüística del sistema.

- El grau de proximitat lingüística de les llengües entre les quals s'està traduint.

- El tipus de textos que es vol traduir.

La figura següent dóna una idea de l'adequació de sistemes i tipus d'aplicacions segons aquests factors:

Figura 13


C = Comprensió T = Traducció massiva
- Bé
- Regular
- Malament
Afinitat del parell de llengües
Llengües molt afins (català-castellà)
Llengües afins (anglès-català)
Llengües allunyades (rus-català)
Llengües molt allunyades (japonès-català)
STA de traducció directa
STA de traducció directa++
STA de transfer sobre PC
STA de transfer sobre UNIX
Potència lingüística de l'STA

8. La TA a Internet

En els darrers anys el fenomen d'Internet s'ha convertit en una eina de treball omnipresent i gairebé imprescindible, no només a universitats, laboratoris o empreses, sinó també en entorns domèstics. Pràcticament tot es pot buscar, trobar, comprar o vendre a Internet, i la traducció automàtica no n'és una excepció.

Actualment, Google Translate és el sistema de traducció automàtica a Internet per excel·lència. Tot i així, hi ha força més sistemes que es poden consultar, provar i fer servir a la Xarxa. A Hutchins (2009), hi trobareu una relació molt extensa i bastant actualitzada de tots aquests sistemes.

A continuació, donem algunes adreces interessants:

Taula 7

Adreça	Descripció
http://traductor.gencat.cat/	Portal de traducció automàtica de la Generalitat de Catalunya
http://www.lucysoftware.com/	Lucy Software
http://www.apertium.org/	Apertium/OpenTrad
http://www.cervantes.es/	Instituto Cervantes
http://translate.google.cat/	Google Translate
http://www.automatictrans.es/	AutomaticTrans
http://www.freetranslation.com/	SDL FreeTranslation
http://www.internostrum.com/	InterNostrum (Universitat d'Alacant)
http://www.online-translator.com/	Prompt
http://www.reverso.com/	Reverso
http://www.systranet.com/	Systran
http://babelfish.yahoo.com/	Babelfish (Yahoo)

En alguns casos no només s'ofereix la traducció en línia a través d'un navegador d'Internet, sinó també la possibilitat d'accedir a la traducció automàtica a través de crides a Web Services des d'aplicacions externes programades per l'usuari. Aquest és el cas de Google Translate i d'OmegaT, una aplicació que fa servir aquesta interfície. Un altre exemple és l'API AJAX del traductor automàtic de Microsoft.

Resum

En aquest mòdul hem vist breument els aspectes principals que tenen a veure amb la traducció automàtica: què és i què no és, la seva història, els seus límits, els tipus de sistemes de TA, els components típics d'un sistema de TA, el procés de la TA, les aplicacions de la TA i la presència de la TA a Internet.

Activitats

1. Des d'un cercador d'Internet (p. ex. Yahoo!, Lycos, Olé, AltaVista, etc.) proveu a buscar pàgines amb les paraules clau *machine translation*.

2. Penseu com es podria simular el procés de la TA amb una cadena de persones, cadascuna de les quals s'encarregués d'una tasca en particular (p. ex., una persona buscaria paraules a un diccionari català, una altra consultaria la gramàtica per dibuixar l'arbre sintàctic, etc.). Penseu, també, en quina mena de material necessitarien (diccionaris, gramàtiques, etc.).

3. Considereu els avantatges i els inconvenients que comporta la traducció massiva amb sistemes de TA de textos entre el castellà i el català.

Exercicis d'autoavaluació

1. Què tenen en comú la traducció automàtica, la traducció simultània i un corrector gramatical per a processadors de textos?

2. Quina és la diferència fonamental entre els lèxics computacionals usats als sistemes de TA i els diccionaris tradicionals en paper?

3. Per què es pot dir que un sistema de TA basat en interlingua és un cas particular dels sistemes basats en transferència?

4. Quin problema de traducció al castellà veieu a la frase *Quan hi vam arribar, l'hi vam donar, malgrat que ella no hi comptava*?

5. Per què un sistema de TA actual té tantes dificultats per traduir bé al castellà una frase catalana com ara *Veus com els pares criden els seus fills quan volen*?

6. Quines traduccions podrien sortir de la frase anterior si es fa servir un sistema de traducció directa (traducció paraula per paraula)?

7. Quines són les diferències fonamentals entre els sistemes de TA basats en regles (RBMT) i els sistemes de TA estadístics (SMT)?

8. Quins són els punts forts i els punts febles dels sistemes basats en regles i dels sistemes estadístics?

Solucionari

Exercicis d'autoavaluació

1. El punt fonamental que tenen en comú és que les tres coses treballen amb el llenguatge humà. Més concretament, la traducció automàtica i la traducció simultània tenen la traducció com a activitat comuna, tot i que els objectius, la forma i els mitjans són radicalment diferents. Quant a la traducció automàtica i als correctors gramaticals, el seu punt comú és la necessitat de fer una anàlisi morfològica i sintàctica de les frases del text per poder-ne aconseguir els objectius.

2. El diccionaris tradicionals en paper contenen molt poca informació gramatical i molta informació sobre el significat de cada paraula. Els lèxics computacionals, al contrari, contenen molta informació gramatical i formalitzada sobre cada paraula, però gairebé gens d'informació sobre el seu significat.

3. Un sistema basat en interlingua és, en realitat, un sistema basat en transferència en què la fase d'anàlisi va més enllà de la representació morfosintàctica i arriba a una representació del significat de la frase, a partir de la qual es pot generar directament la frase de sortida, de manera que, idealment, la fase de transferència desapareix.

4. A la frase proposada, el pronom feble *hi* té tres funcions diferents: complement de lloc (*quan* ***hi*** *vam arribar* (*allà*)), complement indirecte (***l'hi*** *vam donar* (*a ella*)) i complement preposicional (*ella no* ***hi*** *comptava* (*amb allò*)). La traducció del pronom a la frase castellana canvia en cada cas, i és evident que, si no es fa una anàlisi adequada de la frase, no es podrà traduir la frase correctament.

5. La dificultat principal rau en el fet que moltes de les paraules de la frase presenten fenòmens d'homografia: ***veus*** (nom *veu*, forma del verb *veure*), ***pares*** (plural del nom *pare*, forma del verb *parar*), ***seus*** (adjectiu possessiu, plural del nom *seu*, forma del verb *seure*), ***volen*** (forma dels verbs *voler* i *volar*). El problema és fer que el sistema de TA agafi l'alternativa correcta.

6. Per exemple:

Voces como los paras llaman los sientas hijos cuando vuelan.
Voces como los padres gritan sus hijos cuando quieren.
Ves como los paras gritan los sedes hijos cuando vuelan.
Etc.

7. Els sistemes basats en regles fan servir coneixement lingüístic (lèxic, regles morfològiques, regles sintàctiques, regles semàntiques) per fer una anàlisi estructural de cada frase de la llengua de partida, a partir de la qual derivar l'estructura corresponent en la llengua d'arribada i, finalment, generar-ne la traducció. Per contra, els sistemes estadístics (purs) no fan servir cap coneixement lingüístic, sinó models probabilístics extrets a partir de corpus bilingües molt extensos i algorismes estadístics que intenten relacionar paraules, frases o estructures de la llengua de partida amb paraules, frases o estructures de la llengua d'arribada.

8. En general, els sistemes basats en regles lingüístiques són bons en el tractament de fenòmens morfològics (concordança de gènere, nombre i/o cas entre noms, adjectius i determinants, entre el verb i els seus arguments, generació de les formes flexives correctes, etc.) i de fenòmens sintàctics complexos (dependències a llarga distància, clàusules subordinades o de relatiu, negació, etc.), justament on els sistemes estadístics presenten més errors. Per contra, els sistemes estadístics són bons en el tractament de la selecció lèxica i en la generació de frases "naturals" en la llengua d'arribada, dos dels punts febles dels sistemes basats en regles.

Glossari

anàlisi morfològica *f* Procés que dóna com a resultat les possibles interpretacions morfològiques d'una paraula.

anàlisi sintàctica *f* Procés que dóna com a resultat un arbre sintàctic que representa l'estructura de constituents d'una frase.

arbre sintàctic *m* Resultat de l'anàlisi sintàctica i representa l'estructura de constituents d'una frase.

corrector gramatical *m* Eina que s'utilitza per detectar errades gramaticals en un text escrit.

corrector ortogràfic *m* Eina que s'utilitza per detectar errades ortogràfiques en un text escrit.

dada *f* En general, informació que fa servir un programa per dur a terme una tasca determinada.

descodificador *m* Algorisme dels sistemes de TA estadístics que intenta maximitzar la probabilitat que una frase de la llengua d'arribada sigui la traducció d'una frase de la llengua de partida i, a més, que sigui una frase correcta en la llengua d'arribada.

gramàtica d'anàlisi *f* Conjunt de regles formals que descriuen el comportament morfosintàctic d'una llengua i que és utilitzat per un sistema de TA per analitzar les frases d'entrada.

gramàtica de generació *f* Conjunt de regles formals que descriuen el comportament morfosintàctic d'una llengua i que és utilitzat per un sistema de TA per generar les frases de sortida.

informació lèxica *f* Informació lingüística present als lèxics d'un sistema de TA.

informació morfològica *f* Informació lingüística referida a la morfologia de les paraules d'una llengua.

informació sintàctica *f* Informació lingüística referida a la sintaxi de les frases d'una llengua.

informe ALPAC *m* Informe publicat el 1966 als Estats Units que va suposar una aturada de la inversió en recerca sobre TA.

interlingua *f* Llenguatge de representació semàntica que permet representar el significat d'una frase.

lema *m* Forma de citació d'una paraula (p. ex., el lema de *llegíem* és *llegir*).

lèxic monolingüe *m* Conjunt d'entrades amb informació morfològica, sintàctica i semàntica sobre les paraules d'una llengua determinada.

lèxic bilingüe *m* Conjunt d'entrades que relacionen entrades monolingües de la llengua de partida amb entrades monolingües de la llengua d'arribada.

llengua d'arribada *f* Llengua a la qual es tradueix.

llengua de partida *f* Llengua de la qual es tradueix.

llenguatge de descripció gramatical *m* Llenguatge formal que permet als lingüistes especificar les regles gramaticals que constitueixen les gramàtiques d'anàlisi o de generació d'un sistema de TA.

memòria de traducció *m* Conjunt de frases emmagatzemades amb la seva traducció a una altra llengua. Per extensió, el tipus de programa de TA que les fa servir.

model de llengua *m* Model probabilístic que reflecteix la probabilitat que una paraula, una frase o un arbre sintàctic pertanyi al conjunt de paraules, frases o arbres sintàctics correctes en una llengua específica.

model de traducció *m* Model probabilístic que reflecteix la probabilitat que una paraula, una frase o un arbre sintàctic X_d de la llengua d'arribada sigui la traducció correcta d'una paraula, una frase o un arbre sintàctic X_o de la llengua de partida.

parell tret-valor *m* Forma de representar informació lingüística als nodes d'un arbre sintàctic.

parser *m* Analitzador sintàctic.

programa *m* Conjunt d'algorismes que duen a terme tasques específiques a partir de la informació present a les *dades*.

Bibliografia

Copeland, C.; Durand, J.; Krauwer, S.; Maegaard, B. (1991). *The Eurotra Linguistic Specifications*. Luxemburg: Commission of the European Communities.

Deacon, T. (1997). *The Symbolic Species*. Londres: Penguin Books.

Giménez, J. (2009). "Empirical Machine Translation and its Evaluation". *Colección de Monografías de la SEPLN* (núm. 8).

Hawkins, J.; Blakeslee, S. (2004). *On Intelligence*. Nova York: Times Books. Traducció al castellà: *Sobre la inteligencia* (2005). Espasa Calpe.

Hutchins, W. J.; Somers, H. L. (1992). *An Introducction to Machine Translation*. Londres: Academic Press.

Hutchins, W. J. (comp.) (2009). *Compendium of Translation Software*. European Association for Machine Translation.

Nagao, M. (1984). "A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle". A: A. Elithorn i R. Barnerji (editors). *Artificial and Human Intelligence*. Elsevier Science Publishers, B.V.

Nagao, M. (1989). *Machine Translation: how far can it go?* Oxford University Press.

Niremburg, S. (1987). *Machine Translation: theoretical and methodological issues*. Cambridge University Press.

Pinker, S. (1994). *The Language Instinct*. Londres: Penguin Books. Traducció al castellà: *El instinto del lenguaje*. Alianza Ensayo.

Pinker, S. (2007). *The Stuff of Thought*. Londres: Penguin Books. Traducció al castellà: *El mundo de las palabras*. Paidós.

Slocum, J. (1988). *Machine Translation Systems*. Cambridge: Cambridge University Press.