

Informàtica Aplicada a la Traducció  
Examen de prova  
13 de gener de 2001  
Segona part (2 hores)

**Instruccions:**

- En aquesta part de l'examen podeu consultar qualsevol material que hàgeu portat però no podeu compartir material ni parlar amb ningú.
- Si teniu algun dubte, alceu la mà i aniré a atendre-us.

**Qüestions:**

1. Indiqueu quina classe d'ambigüitat presenten aquestes frases (justifiqueu breument la vostra resposta):
  - (a) “Van fer una manifestació pel parc científic” (1: Van desfilars a través del parc científic; 2: La manifestació era a favor del parc científic)
  - (b) “Els assassins són policies o milicians rebels” (Els policies eren també rebels?)
  - (c) “Com penses que han arribat al cim?” (1: Com és possible que penses això? 2: Saps de quina manera hi han arribat?)
  - (d) “Ocultó el rollo para que nadie conociese su verdadera orientación sexual” (1: Les fotos, si les revelaven, podien descobrir les seues preferències sexuals; 2: aquella relació deixaria clares les seues orientacions sexuals)
  - (e) “Els professors acompanyaven els xiquets quan els policies els van cridar” (A qui van cridar?)
  - (f) “És difícil però no impossible” (Què és difícil?)
  - (g) “Será mejor que verse sobre la nueva construcción” (1: Més val que tracte sobre el mateix tema; 2: és millor citar-se en aquell altre lloc).
2. (!) Imaginem que tenim textos en la llengua  $L_1$  i les traduccions (correctes) corresponents en la llengua  $L_2$ , fetes per una persona de manera que s'hi respecte tant com siga possible l'estructura de les frases del text original. Indiqueu com es podria usar la informació present en aquestes traduccions per a resoldre les ambigüitats que es troba un sistema de traducció automàtica quan tradueix *els mateixos textos* a una altra llengua  $L_3$  (Martin Kay, de Xerox Palo Alto, Califòrnia, anomena aquest procés *triangulació en la traducció*). Si voleu, fixeu-vos només en les ambigüitats lèxiques.

3. Elegiu dues llengües,  $L_1$  i  $L_2$ , i poseu dos exemples de traduccions errònies de  $L_1$  a  $L_2$ , una deguda a la traducció incorrecta d'un homògraf i l'altra deguda a la traducció incorrecta d'un mot polisèmic, indicant la frase original, la frase mal traduïda i la frase correcta.
4. El programari que porten instal·lat les naus de la confederació galàctica inclou un programa que tradueix la una de les llengües majoritàries del planeta Zkanagg, el tazkannwat, al català. El sistema és un sistema de transferència morfològica avançada estàndard, que llegeix els textos d'esquerra a dreta, mot a mot, busca en l'entrada els patrons de categories lèxiques que conté en el seu catàleg, selecciona el més llarg, reordena i concorda els mots del patró, els escriu, i continua després de la zona reordenada. Algunes traduccions són errònies perquè el sistema no té un catàleg massa complet de regles. Fixeu-vos en els exemples i digueu quins són els patrons que detecta i quines les regles de reordenament associades.
- (1) *Thlong u knaar uw phlagyw.*  
Va adquirir el navegant el-OBJ control-OBJ  
TA: El navegant va adquirir el control (correcta).
- (2) *Thlong u knaar qimratt uw phlagyw.*  
Va adquirir el navegant estelar el-OBJ control-OBJ  
TA: El navegant estelar va adquirir el control (correcta).
- (3) *Thlong u knaar na Zkannag uw phlagyw.*  
Va adquirir el navegant de Zkannag el-OBJ control-OBJ  
TA: El navegant de Zkannag va adquirir el control (correcta).
- (4) *Thlong u knaar qimratt na Zkannag uw phlagyw.*  
Va adquirir el navegant estelar de Zkannag el-OBJ control-OBJ  
TA: \*El navegant estelar va adquirir de Zkannag el control.  
Correcta: El navegant estelar de Zkannag va adquirir el control.
5. Les estructures sintàctiques usades per diverses llengües a l'hora d'expressar el fet que alguna acció agrada a algú són molt diverses. La frase "m'agrada nadar" té una estructura sintàctica molt diferent en anglés (5) i en alemany (6).
- (5) *Peter likes swimming.*  
Peter s'estima nadant.

- (6) *Peter schwimmt gern.*  
Peter nada amb gust.

Altre exemple el tenim en la frase “li diuen Joan”:

- (7) *He is called Joan.*  
Ell és anomenat Joan.

- (8) *Er heißt Joan.*  
Ell s'anomena Joan.

En canvi, les estructures que serveixen per a expressar altres conceptes poden ser completament paral·leles. Quines conseqüències té això per al disseny d'un sistema de traducció automàtica de transferència sintàctica entre llengües amb aquestes característiques? Com es podrien evitar els problemes observats?

6. (!) Els mots no són tots igualment freqüents en els textos. De fet, si ordenem els mots d'un gran corpus de text real (de qualsevol tipus i de qualsevol idioma) pel nombre de vegades que hi apareixen, començant pel més freqüent, el nombre d'aparicions es redueix dramàticament segons que anem baixant per la llista. Típicament, el mot més freqüent pot arribar a constituir el 10% de tot el text, però el segon només cobreix al voltant del 5%, el tercer al voltant del 3%, etc.; quan arribem al 100é mot més freqüent ja hem de parlar del 0,1% (una vegada cada 1.000 mots), i si arribem a la posició 1000, del 0,01% (una vegada cada 10.000 mots). En resum, la distribució no és gens homogènia: uns pocs mots són els més freqüents i la majoria són moltíssim menys freqüents. De fet, és típic que la majoria dels mots siguin *hapax legomena*, és a dir, mots que han aparegut només una vegada en tot el corpus. Si haguésseu de supervisar la construcció dels diccionaris d'un sistema de traducció automàtica, per a què us podrien servir aquestes constatacions estadístiques?