El sistema de traducció automàtica espanyol↔català interNOSTRUM

R. Canals-Marote, A. Esteve-Guillén, A. Garrido-Alenda, P. Gilabert-Zarco, M.I. Guardiola-Savall, J. Herrero-Vicente, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, A. Pertusa-Ibáñez, P.M. Pérez-Antón, G. Ramírez-Sánchez, F. Ramos-Salas, M. Samper-Asensio i Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics

Universitat d'Alacant, E-03071 Alacant

Índex

- Què és interNOSTRUM?
- Plataforma
- Estratègia de traducció automàtica
- Integrabilitat i extensibilitat
- Problemes de traducció automàtica
- Altres aspectes lingüístics
- Treball pendent

Què és interNOSTRUM? [1]

- Sistema de TA bidireccional espanyol↔català
- Impulsat per la Caja de Ahorros del Mediterráneo i cofinançat per la Universitat d'Alacant des de l'any 1999
- Tradueix text sense format, RTF, HTML, correu electrònic, xat
- Velocitat (6000 par./s en AMD a 2600 Mhz): permet la *navegació traduïda*

Què és interNOSTRUM? [2]

- Disponible públicament en Internet (www.internostrum.com)
- I a través d'altres portals (www.softcatala.org, www.ociosfera.com)
- 500.000 accessos mensuals, en alça
- S'usa per a produir esborranys de traduccions al català...
- ...o per a llegir en *espanyol aproximat* pàgines d'Internet en català.

¿Què és interNOSTRUM? [3]

- Versió actual: català central
- Versions valenciana i balear analitzades
- Versió català→espanyol una mica menys desenvolupada
- Taxes d'error al voltant del 8%

Plataforma [1]

- El servidor públic s'executa sobre Linux (Apache, PHP amb suport XML)
- La configuració com a servidor permet oferir sempre la millor versió als usuaris
- InterNOSTRUM: cadena de muntatge de 8 mòduls
- 6 mòduls generats a partir de dades lingüístiques (format intel·ligible)

Estratègia de traducció automàtica [1]

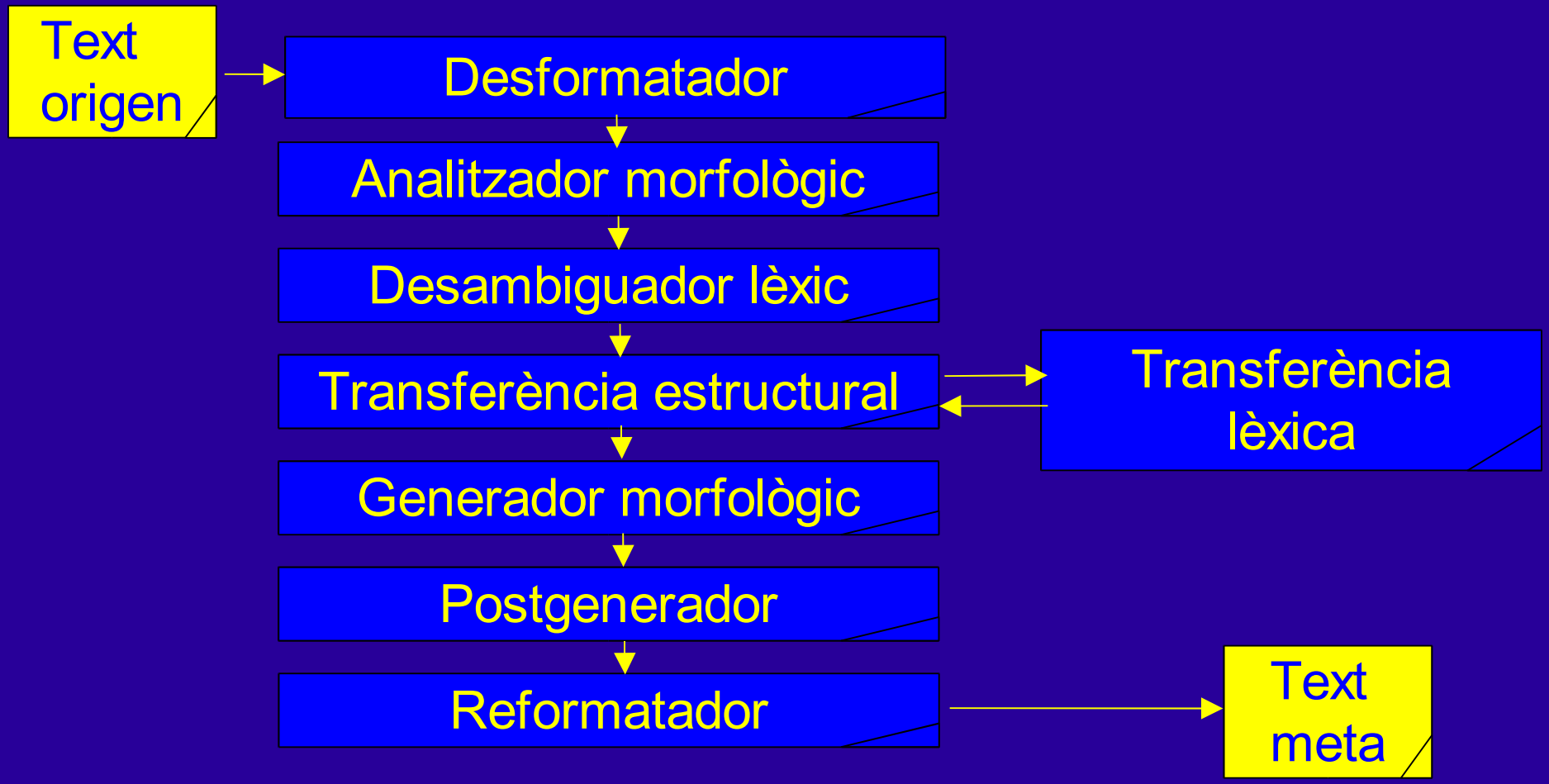
- InterNOSTRUM és un sistema de transferència morfològica avançada.
- Estratègia semblant a la de sistemes comercials per a PC: Transcend RT, Reverso i versions antigues de Power Translator.

Estratègia de traducció automàtica [2]

Vuit mòduls bàsics (cadena de muntatge de text):

- Desformatador (HTML, RTF)
- Analitzador morfològic
- Desambigüador lèxic (categorial, estadístic)
- Transferència estructural (invoca la transferència lèxica)
- Generador morfològic
- Post-generador (guionatge, apostrofació, etc.)
- Reformatador

Estratègia de traducció automàtica [3]



Integrabilitat

- InterNOSTRUM està programat de manera que es pot *integrar* fàcilment dins d'altres productes informàtics (especialment els relacionats amb Internet)
- El disseny permet integrar cadascun dels mòduls per separat (analitzador morfològic, etiquetador morfosintàctic, etc.)

Extensibilitat

- InterNOSTRUM es relativament fàcil d'aplicar a altres llengües:
  - Possibilitat d'aprofitar per separat la tecnologia desenvolupada i les dades lingüístiques existents
  - Admet l'adquisició i la implementació de noves dades lingüístiques d'altres llengües
  - En progrés, sistema de TA espanyol↔portugués amb un 85% de cobertura i taxes d'error de 10% (6 mesos)

Problemes de la traducció automàtica

- Ambigüitat lèxica:
  - Homografia
  - Polisèmia
- Ambigüitat sintàctica

Ambigüitat lèxica [Homografia]

- Homografia

"En el libro hay tres direcciones de la Dirección General. Tendré que elegir una."

- [libro: verb / nom]: allibero / llibre
- [la: pronom / nom /determinant]: la (*Passi gratuït*)
- [Una: determinant / verb]: una / uneixi

- Alguns homògrafs són *molt freqüents (entre 4 i 7 vegades / 1000 paraules)*:

- *Como*: com / com a / menjo
- *Para*: per a / per / atura

Ambigüitat lèxica [Polisèmia]

- Polisèmia

"En el libro hay tres direcciones de la Dirección General. Tendré que elegir una."

- [direcciones: lloc d'ubicació]: adreces
- [Dirección General: departament ministerial]: direcció

- Altres exemples:
  - *Set* = set / sed
  - *Cap* = cabeza / cabo / jefe

Ambigüitat lèxica [Resolució]

"En el libro hay tres direcciones de la Dirección General. Tendré que elegir una."

- Mòdul de desambiguació lèxica categorial
- Tractament d'unitats multimot:

  Direcció general (sense flexió)

  Haver de (amb flexió)
- Sacrifici de mots amb baixa freqüència (analitzem *menjo,* però no el generem perquè és molt més infreqüent que *com*)

Ambigüitat sintàctica

En la frase : "Se comió el postre enfadado"

Qui està empipat, la persona o les postres?

- Se comió [el postre] [enfadado]:

  "Es va menjar les postres empipat"
- Se comió [el postre enfadado]:

  "Es va menjar les postres empipades"

Altres aspectes lingüístics

- Sinonímia

"En el libro hay tres direcciones de la Dirección General. Tendré que elegir una."

- [elegir = triar / elegir]
- Indicacions de sentit en diccionaris bilingües

Castellà		Català	
elegir	↔	elegir	(ambdós sentits)
elegir	←	triar	(unidireccional)

Treball pendent

- Millora dels vocabularis (mots simples i unitats multimot), semiautomàticament mitjançant l'ús de bitextos alineats.
- Ampliació de les regles del mòdul de transferència estructural.
- Assistents de preedició i de postedició
- Gestió dels diccionaris a través de bases de dades i formularis *web*

Sistema de traducció automàtica
espanyol↔català

www.internostrum.com