El Centre de Referència en Enginyeria Lingüística

M. Teresa Cabré
Membre de la Secció Filològica de l'IEC, directora del CREL
i directora del projecte *El català en la societat de la informació*

L'objectiu d'aquesta comunicació és informar sobre el Centre de Referència en Enginyeria Lingüística de Catalunya (CREL) i sobre els treballs que s'hi duen a terme.

El CREL, Centre de Referència en Enginyeria Lingüística, aprovat per la Generalitat de Catalunya el 1996, és un centre creat en el marc del II Pla de Recerca de la Generalitat de Catalunya, que aplega grups d'investigació en enginyeria lingüística, lingüística computacional i lingüística aplicada. El propòsit fonamental dels treballs és de dotar la llengua catalana dels recursos i les eines que la facin apta per als usos digitals i la incorporin a les noves tecnologies que la societat de la informació requereix.

Formen part del CREL els grups de recerca següents:

a) Grup de Recerca de l'Institut d'Estudis Catalans:

— Equip del *Diccionari del català contemporani* de l'IEC.

— Equip de les Oficines Lexicogràfiques de l'IEC.

b) Grup de Recerca en Tractament del Llenguatge Natural de la Universitat Politècnica de Catalunya i de la Universitat de Barcelona.

c) Grup de Tractament de la Parla de la Universitat Politècnica de Catalunya.

d) Grup de Recerca en Lingüística Aplicada al Tractament del Llenguatge de la Universitat Pompeu Fabra.

e) Seminari de Filologia i Informàtica de la Universitat Autònoma de Barcelona.

f) Grup de Recerca en Variació en el Llenguatge de la Universitat de Barcelona.

g) Grup de Lingüística Aplicada de la Universitat de Girona.

h) Grup de Gramàtica Teòrica de la Universitat Autònoma de Barcelona.

En total, el CREL està constituït per nou equips pertanyents a sis institucions diferents. El CREL s'organitza internament en el Consell Científic, el Consell de Direcció i la Direcció. La gerència del centre l'exerceix l'Institut d'Estudis Catalans.

El contracte programa del CREL amb la Generalitat de Catalunya especifica els treballs que s'han de realitzar en un període de quatre anys i estableix les seves tres funcions bàsiques:

1) Realitzar recerca i desenvolupament en enginyeria lingüística, lingüística computacional i lingüística aplicada en llengua catalana.

2) Desenvolupar projectes comuns en aquests àmbits.

3) Potenciar les infraestructures de R+D en enginyeria lingüística.

Al costat d'aquests treballs específics, el contracte programa també preveu la confecció d'un llibre blanc dels recursos informatitzats en català, la interconnexió entre els grups per mitjà d'una intranet i el disseny d'un centre de recursos ubicat a l'IEC a través del portal del qual s'oferiran els recursos constituïts.

Els camps de treball previstos s'organitzen en tres eixos:

1) La constitució de recursos lingüístics (textuals, lèxics i gramaticals).

2) L'elaboració d'eines de tractament de les dades.

3) La creació de prototips.

Vegeu, com a mostra, els treballs específics corresponents al bienni 1999-2000:

Llenguatge oral

1) *Constitució de corpus orals*
   - 1.1) Acabament de la constitució del corpus Speechdat.
   - 1.2) Anotació i compleció del corpus prosòdic, ja constituït, amb vista a facilitar la conversió de text a parla.
   - 1.3) Creació d'un corpus d'interacció oral simulada persona-màquina.
   - 1.4) Transcripció i anàlisi del corpus de diàlegs persona-persona.

2) *Eines de tractament de la parla*
   - 2.1) Transcripció automàtica a partir de regles externes.
   - 2.2) Elaboració del *Diccionari fonètic de suport a la transcripció automàtica*.
   - 2.3) Segmentació de corpus orals i paral·lelització.
   - 2.4) Convertidor de text a parla amb mòduls controlables externament.

3) *Eines de conversió de text a parla*
   - 3.1) Extensió i millora del preprocessador lingüístic.
   - 3.2) Construcció d'una eina de selecció òptima del conjunt d'unitats fonètiques a partir de corpus.
   - 3.3) Elaboració d'un model d'assignació de pauses.
   - 3.4) Elaboració d'un model d'assignació dels patrons locals d'entonació.

4) *Eines de reconeixement de la parla*
   - 4.1) Modelatge estadístic basat en semifonemes.

4.2) Entrenament discriminatori de les unitats sublèxiques.
4.3) Creació de models robustos a les condicions de l'entorn.
4.4) Consideració de la coarticulació entre paraules.

5) *Prototip d'accés oral*
5.1) Increment del vocabulari de l'aplicació a un miler de paraules.
5.2) Adaptació del modelatge prosòdic del convertidor de text a parla a la situació de diàleg.
5.3) Modelatge del diàleg per a l'aplicació de sol·licitud d'informació i reserva de bitllets de tren.

LLENGUATGE ESCRIT

1) *Constitució de corpus escrits*
1.1) Compleció del corpus de llenguatges d'especialitat.
1.2) Actualització del corpus de català contemporani.
1.3) Constitució d'un corpus paral·lel català-castellà.
1.4) Compleció del corpus de premsa.

2) *Eines de tractament de l'escrit*
2.1) Documentació final del preprocessador.
2.2) Documentació final dels analitzadors morfològics.
2.3) Documentació final del desambiguador lingüístic.
2.4) Documentació final del desambiguador estocàstic.
2.5) Finalització de l'analitzador sintàctic de primer nivell.
2.6) Eina de gestió de diccionaris.
2.7) Actualització dels diccionaris disponibles.
2.8) Tancament del *Diccionari de locucions i frases fetes*.
2.9) Incorporació del *Diccionari de locucions i frases fetes* a la cadena de processament.

3) *Eines de detecció d'informació*
3.1) Programa de detecció de neologismes.
3.2) Programa de detecció de terminologia (primera fase).
3.3) Elaboració de l'Eurowordnet català.

4) *Recerca per a l'elaboració d'eines*
4.1) Disseny de l'analitzador sintàctic de segon nivell.
4.2) Preparació de l'anàlisi i marcatge semàntics.
4.3) Anàlisi semàntica.

4.4) Sistema de recuperació d'informació.

4.5) Conversió del *Diccionari de la llengua catalana* de l'Institut d'Estudis Catalans en un diccionari computacional.

5) *Prototips*

5.1) Acabament de l'estació de treball lexicogràfic.

L'any 2001 acabarà el contracte programa signat per l'IEC, en representació dels grups científics, i la Generalitat de Catalunya. Cal esperar que aleshores ja disposarem del Centre de Distribució de Recursos Lingüístics situat a l'IEC, que oferirà a la comunitat investigadora les dades del CREL, juntament amb les eines que permetin de tractar-les automàticament.