

CLiC-Centre de Llenguatge i Computació

SGR2005-00309

Universitat de Barcelona - Universitat Autònoma de Barcelona



Jornada del Processament Computacional del Català
26 DE MARÇ DE 2009, PALAU ROBERT (BARCELONA)

Línies de recerca (I)

- **Fonaments lingüístics i metodològics** en el desenvolupament de recursos i eines d'enginyeria lingüística.
 - **Transductors:** grafia - fonema
 - **Analitzadors:** morfològic, *chunker*
 - **Corpus:** anotació morfosintàctica, sintàctica, semàntica i pragmàtica
 - **Lexicons:** nominals i verbals
- Frase → Discurs
- Llengua estàndard → parla, textos no normatius



Línies de recerca (I)

- **Fonaments lingüístics i metodològics** en el desenvolupament de recursos i eines (i escrita) d'enginyeria lingüística.
 - **Transductors:** grafia - fonema
 - **Analitzadors:** morfològic, *chunker*
 - **Corpus:** anotació morfosintàctica, sintàctica, semàntica i pragmàtica
 - **Lexicons:** nominals i verbals
- **Frase → Discurs**
- **Llengua estàndard → parla, textos no normatius**



Línies de recerca (I)

- **Fonaments lingüístics i metodològics** en el desenvolupament de recursos i eines (i escrita) d'enginyeria lingüística.
 - **Transductors:** grafia - fonema
 - **Analitzadors:** morfològic, *chunker*
 - **Corpus:** anotació morfosintàctica, sintàctica, semàntica i pragmàtica
 - **Lexicons:** nominals i verbals
- **Frase → Discurs**
- **Llengua estàndard → parla, textos no normatius**



Línies de recerca (II)

- Eines per al **processament massiu** de textos
 - Anàlisi i desambiguació **morfològica** (HsMorfo)
 - Anàlisi **sintàctica** superficial (Gram-Cat: TaCat)
- **Corpus anotats: AnCora.**

Nivells: morfològic, sintàctic, semàntica lèxica, semàntica oracional, named entities, coreferència.
- Base per al desenvolupament d'analitzadors basats en ML (*corpus-driven approach*)
- Resolució de la **coreferència**: sistema basat en ML amb coneixement lingüístic.
- Extracció i identificació de **paràfrasis**: **coneixement lingüístic + tècniques automàtiques.**



Línies de recerca (II)

- Eines per al **processament massiu** de textos
 - Anàlisi i desambiguació **morfològica** (HsMorfo)
 - Anàlisi **sintàctica** superficial (Gram-Cat: TaCat)
- **Corpus anotats: AnCorà.**

Nivells: morfològic, sintàctic, semàntica lèxica, semàntica oracional, named entities, coreferència.
- Base per al desenvolupament d'analitzadors basats en ML
(corpus-driven approach)
- Resolució de la **coreferència**: sistema basat en ML amb coneixement lingüístic.
- Extracció i identificació de **paràfrasis**: **coneixement lingüístic + tècniques automàtiques.**



Línies de recerca (II)

- Eines per al **processament massiu** de textos
 - Anàlisi i desambiguació **morfològica** (HsMorfo)
 - Anàlisi **sintàctica** superficial (Gram-Cat: TaCat)
- **Corpus anotats: AnCora.**

Nivells: morfològic, sintàctic, semàntica lèxica, semàntica oracional, named entities, coreferència.
- Base per al desenvolupament d'analitzadors basats en ML
(corpus-driven approach)
- Resolució de la **coreferència**: sistema basat en ML amb coneixement lingüístic.
- Extracció i identificació de **paràfrasis: coneixement lingüístic + tècniques automàtiques.**



Projectes per al català

- **CEsCa: Corpus Escrit del Català Escolar** (2008ARIE-00053; 2007ARIE-00005; 2006ARIE-10058). Creació i processament lingüístic del corpus del català escolar escrit. Durada: 2007-2009.
- **Lang2World: Descubriendo el conocimiento del mundo codificado en la lengua** (TIN2006-15265-C06-06). Combinació de tècniques estadístiques i coneixement lingüístic per al tractament sintàctic i semàntic del català i el castellà. Desenvolupament de corpus anotats a diferents nivells d'anàlisi lingüística (corpus AnCora) per a l'aplicació de tècniques d'aprenentatge automàtic. Durada: 2006-2009.
- **CESS-ECE: Corpus Etiquetado Sintáctica y Semánticamente del Euskera, Catalán y Español.** (HUM2004-21127-E). Banc de dades sintàctic anotat amb constituents i funcions de 500.000 paraules per a cada llengua. Durada: 2005-2007.
- **Praxem, etiquetado semántico y pragmático del corpus CESS-ECE** (HUM2006-27378-E). Anotació de corpora amb informació semàntica i pragmàtica. Durada: 2007-2008.
- **AnCora-Nom: Anotación semántica del SN en los corpus AnCora** (FFI2008-02691-E/FILO). Extensió dels corpus AnCora amb l'anotació de l'estructura argumental dels noms. Durada: 2009.
- **Dialcat: Analizador morfosintáctico de corpus dialectales del catalán** (HUM2005-24445-E). Desenvolupament d'un analitzador morfològic de textos dialectals del català. Durada: 2006-2007.
- **Histocat: Analizador morfosintáctico de textos históricos del catalán** (HUM2005-24438-E). Descripció: Desenvolupament d'un analitzador morfològic de textos històrics del català. Durada: 2006-2007.



Projectes per al català

- **CEsCa: Corpus Escrit del Català Escolar** (2008ARIE-00053; 2007ARIE-00005; 2006ARIE-10058). Creació i processament lingüístic del corpus del català escolar escrit. Durada: 2007-2009.
- **Lang2World: Descubriendo el conocimiento del mundo codificado en la lengua** (TIN2006-15265-C06-06). Combinació de tècniques estadístiques i coneixement lingüístic per al tractament sintàctic i semàntic del català i el castellà. Desenvolupament de corpus anotats a diferents nivells d'anàlisi lingüística (corpus AnCora) per a l'aplicació de tècniques d'aprenentatge automàtic. Durada: 2006-2009.
- **CESS-ECE: Corpus Etiquetado Sintáctica y Semánticamente del Euskera, Catalán y Español.** (HUM2004-21127-E). Banc de dades sintàctic anotat amb constituents i funcions de 500.000 paraules per a cada llengua. Durada: 2005-2007.
- **Praxem, etiquetado semántico y pragmático del corpus CESS-ECE** (HUM2006-27378-E). Anotació de corpora amb informació semàntica i pragmàtica. Durada: 2007-2008.
- **AnCora-Nom: Anotación semántica del SN en los corpus AnCora** (FFI2008-02691-E/FILO). Extensió dels corpus AnCora amb l'anotació de l'estructura argumental dels noms. Durada: 2009.
- **Dialcat: Analizador morfosintáctico de corpus dialectales del catalán** (HUM2005-24445-E). Desenvolupament d'un analitzador morfològic de textos dialectals del català. Durada: 2006-2007.
- **Histocat: Analizador morfosintáctico de textos históricos del catalán** (HUM2005-24438-E). Descripció: Desenvolupament d'un analitzador morfològic de textos històrics del català. Durada: 2006-2007.



Resultats: eines i recursos

- **Corpus:**

- **CesCa:** Corpus Escrit del Català escolar. 2.396 textos produïts per nens des de P5 fins a 4t d'ESO de 31 centres educatius de diferents comarques de Catalunya.

<http://clic.ub.edu/ca/cesca>

- **AnCora-Ca:** corpus del català de 500.000 paraules. Textos periodístics, anotat a diferents nivells lingüístics: morfologia, sintaxi (constituents i funcions), semàntica (estructura argumental, synsets nominals de WordNet i Named Entities) i pragmàtica (coreferència definida i anafòrica).

<http://clic.ub.edu/ca/ancora>

- **Lèxics:**

- **AnCora-Verb:** lèxic de 4.520 sentits que inclou la correspondència entre les funcions sintàctiques, els arguments i rols semàntics de cada predicat verbal, tenint en compte la classe semàntica del verb i les alternances de diàtesis en què pot participar.

- **AnCora-Nom:** lèxic de 820 noms deverbals que inclou la correspondència entre les funcions sintàctiques, els arguments i rols semàntics (en procés)

- **Eines:**

- **AnCora-Pipe:** Eina d'anotació de corpus que permet anotar a diferents nivells lingüístics de manera simultània i eficient.

- **Analitzador morfològic:**

- **Dial-Cat** i **Histo-Cat:** Ampliació de l'analitzador morfològic per al tractament de les variants dialectals i variants diacròniques.

<http://stel.ub.edu/dialcat/>, <http://stel.ub.edu/histocat/>



Tots els recursos i eines són de lliure distribució

Futur

Sempre hem treballat per al català

- Projecte (convocatòria 2009): **TEXT-MESS 2.0: Las Tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital.**

Llengües implicades: **català**, espanyol, rus, polonès, anglès, àrab, ...

- **Tractament del text en el marc de la web 2.0:**

- Recopilació de **corpus representatius de la llengua oral i escrita (informal)**
- Definició de **mètodes** i desenvolupament **tècniques** per a la detecció **d'estructures lingüístiques no estàndard pròpies de la llengua espontània**, no normativa, tant oral com escrita
- Anàlisi de la **polaritat**, actitud **emocional**, ...
- **Coreferència**
- **Paràfrasi**

- **Difondre** la recerca **en/per al català**: inclusió del català en competicions internacionals:

Semeval 2010, <http://stel.ub.edu/semeval2010-coref/>

CoNLL 2009, <http://ufal.mff.cuni.cz/conll2009-st/task-description.html>

ARE 2009, <http://www.anaphora-and-coreference.info/ARE2009>