

EXPLOTACIÓ DIDÀCTICA DE CORPUS. EINES D'ANÀLISI LINGÜÍSTICA

Ma. Antònia Martí

Universitat de Barcelona
CLiC-Centre de Llenguatge i Computació

JORNADES DE DIFUSIÓ DE RECERCA UNIVERSITÀRIA – ICE
Octubre 2009

Explotació didàctica de corpus

- Lingüística
- Lingüística Computacional
- Lingüística de corpus

<http://clic.ub.edu>

Tipus de corpus

Oral:

senyal sonor

transcripció fonètica

transcripció ortogràfica

transcripció ortogràfica enriquida

...

Escrit:

paper → OCR

digital

Tipus de corpus

CLINT: oral, llengua espontània
→ transcripció ortogràfica

AnCora: digital, notícies de diari

Cesca: escrit, llengua espontània (?)
→ digitalització (transcripció manual)

Corpus CLINT

Corpus CLINT

Projecte en procés de desenvolupament

- 40 entrevistes
- 4 metges diferents
- 10 entrevistes/metge
- Servei Pulmorespiratori dels Germans Trias i Pujol

<http://clic.ub.edu/clint>

NIVELES DE TRANSCRIPCIÓN

Transcripción ortográfica.

Transcripción ortográfica enriquecida.

Transcripción fonológica.

Transcripción fonética ancha.

Transcripción fonética estrecha.

un pulmón lesionado?

<eee> u(n) un pulmó:n lesionado/

un pulmón lesjonáDo

um pulmón lesjonáDo

um pulmón lesjonáo

GRABACIÓN

↑
Transcriber

Transcripción codificada (en XML)

Corpus CLINT

Quin interès té un corpus d'aquestes característiques?

Quines línies de treball permet obrir?

Quin interès tenen les diferents transcripcions?

Corpus AnCora

Corpus AnCora

AnCora-CAT	500,000
EFE:	75,000
ACN:	225,000
El Periódico:	200,000

AnCora-ESP	500,000
EFE	225,000
Lexesp	75,000
El Periódico	200,000

Corpus AnCora

Corpus anotats: (projecte acabat)

- **Manualment:** Sintaxi, Entitats, Coreferència, Semàntica lèxica
- **Automàticament:** Morfologia
- **Semiautomàticament:** Papers temàtics

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


AnCora - Mozilla Firefox

Eitxer Edita Visualitza Història Adreces d'interès Eines Ajuda

http://clic.ub.edu/ancora/

Més visitades Primers passos Darreres notícies

AnCora | CLIC



AnCora

🇪🇸 🇩🇪 🇪🇸 🇬🇧

- Introducció
- Participants
- Cerques
- Lèxics
- Descàrregues
- Documentació
- Publicacions
- Competicions

Cerques

Selecciona els corpora:
selecciona / deselecciona tots els corpus
Veure resultats en grups de 5

Corpus	
<input type="checkbox"/>	CESS_EU
<input checked="" type="checkbox"/>	AnCora_CA
<input type="checkbox"/>	AnCora_ES

Cerca totes les frases que continguin la paraula: Cerca!

Cerca totes les frases que continguin el lema: Cerca!

Quins contenen Cerca!

Quins fan funció de Cerca!

Quins sintagmes fan funció de Cerca!

Quines frases contenen i (opcional) Cerca!

Buscar estructura sintàctico-semàntica del verb Cerca!

Quins papers temàtics té la funció sintàctica Cerca!

Quines funcions té el paper temàtic Cerca!

Quines frases contenen el paper temàtic i (opcional) Cerca!

Fet

Inicio I IBERIAN SL 2009 - ... I. JORNADES DE DIF... Jornades jaen AnCora - Mozilla Firefox ES 12:58

Corpus CLINT

Quin interès té un corpus d'aquestes característiques?

Quines línies de treball permet obrir?

Quina utilitat li veieu com a material didàctic o per als mestres?

Corpus CesCa

Corpus CesCa

S'han recollit i processat

2.426 textos

Nens i nenes → últim curs d'educació infantil (P5)
fins a l'últim curs d'educació obligatòria (4t
d' ESO)

31 centres educatius

Diferents comarques de Catalunya

Corpus CesCa

El corpus conté:

- 1) El lèxic produït per cinc camps lèxics:
 - noms d'aliments
 - peces de roba
 - fenòmens de la natura
 - activitats de lleure
 - trets de personalitat
- 2) Acudits
- 3) Definicions
- 4) Narració:
 - a. Títol
 - b. El contingut
- 5) Text augmentatiu

Corpus CesCa

Lèxic

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

Dispersió lèxica → procés de lematització manual

Concepte molt laxe de 'lema'

Corpus CesCa

Textos

- Segmentació
- Anàlisi morfològica

Corpus CesCa

Ilavia un para cavulia veura sarvesa pro
das pres vadi no ara ja no tin set

(Text original)

L avia un para ca vulia veura sarvesa
pro das_pres va di no ara ja no tin set

(Text corregit)

Corpus CesCa

Anàlisi morfològica

Format intern: XML

```
<?xml version="1.0" encoding="UTF-8" ?>
< article lng="ca" morpho="automatic:27/07/09 16:55">
< sentence>
< p gen="c" lem="hi" name="n" num="c" person="3" pos="pp3cn000" postype="personal"
  wd="II" />
< v lem="haver" mood="indicative" name="v" num="s" person="3" pos="viii3s0"
  postype="auxiliary" tense="imperfect" wd="avia" />
< d gen="m" lem="un" name="d" num="s" pos="di0ms0" postype="indefinite" wd="un" />
< n gen="m" lem="pare" name="v" num="s" pos="ncms000" postype="common" wd="para" />
< p complex="no" gen="c" lem="que" name="s" num="c" pos="pr0cn000" postype="relative"
  wd="ca" />
< v lem="voler" mood="indicative" name="n" num="s" person="3" pos="vmii3s0"
  postype="main" tense="imperfect" wd="vulia" />
< v lem="veure" mood="infinitive" name="n" pos="vmn0000" postype="main" wd="veura" />
< n gen="f" lem="cervesa" name="n" num="s" pos="ncfs000" postype="common" wd="sarvesa"
  />
< c complex="no" lem="però" name="s" pos="cs" postype="coordinating" wd="pro" />
< r lem="després" name="n" pos="rg" wd="das_pres" />
< v lem="anar" mood="indicative" name="v" num="s" person="3" pos="vaip3s0"
  postype="auxiliary" tense="present" wd="va" />
< v lem="dir" mood="infinitive" name="v" pos="vmn0000" postype="main" wd="di" />
[...] </sentence> </article>
```

wd="II" lem="hi" pos="pp3cn000" postype="personal" person="3"
gen="c" num="c"

wd="avia" lem="haver" pos="viii3s0" mood="indicative" num="s"
person="3" postype="auxiliary" tense="imperfect"

wd="un" lem="un" pos="di0ms0" gen="m" postype="indefinite"
num="s"

wd="para" lem="pare" pos="ncms000" postype="common" gen="m"
num="s"

wd="ca" lem="que" pos="pr0cn000" postype="relative" complex="no"
gen="c" num="c"

wd="vulia" lem="voler" pos="vmii3s0" postype="main"
mood="indicative" num="s" person="3" tense="imperfect"

wd="veura" lem="veure" pos="vmn0000" postype="main"
mood="infinitive"

wd="sarvesa" gen="f" lem="cervesa" num="s" pos="ncfs000"
postype="common"

Resultats del processament morfològic

Variants (orto)gràfiques:

ll → hi

ca → que

para → pare

pare → pare

vulia → volia (volar)

avia → havia (haver)

di → dir

Navigator

- 11281_pel.licula_te
- 11291_pel.licula_te
- 11296_pel.licula_te
- 11301_pel.licula_te
- 11306_pel.licula_te
- 1131_pel.licula_tex
- 11311_pel.licula_te
- 11316_pel.licula_te
- 11321_pel.licula_te
- 11326_pel.licula_te
- 11331_pel.licula_te
- 11336_pel.licula_te
- 11341_pel.licula_te
- 11346_pel.licula_te
- 11351_pel.licula_te
- 11356_pel.licula_te
- 1136_pel.licula_tex
- 11361_pel.licula_te
- 11366_pel.licula_te
- 11371_pel.licula_te
- 11376_pel.licula_te
- 11381_pel.licula_te
- 11386_pel.licula_te
- 11391_pel.licula_te
- 11396_pel.licula_te
- 11401_pel.licula_te
- 11406_pel.licula_te
- 1141_pel.licula_tex
- 11411_pel.licula_te
- 11416_pel.licula_te
- 11421_pel.licula_te
- 11426_pel.licula_te
- 11436_pel.licula_te
- 11441_pel.licula_te
- 11451_pel.licula_te
- 11456_pel.licula_te
- 1146_pel.licula_tex

Properties Function/Arg/T

Property	Value
Coreference	
homophoric	no
Functions synta	
arg	NOT PRESENT
func	NOT PRESENT
tem	NOT PRESENT
Lexic	
lexicalized	no
Lexical	
lexicalid	
Misc	
name	n
toreview	NOT PRESENT
toreviewcor	
Morphology	
case	NOT PRESENT
gen	f
multiword	no
num	p
postype	common
Named entities	
ne	NOT PRESENT
Syntax	
discid	NOT PRESENT
discontinuo	no
elliptic	no
missing	no
Text	
lem	cosa
title	no
titlelevel	NOT PRESENT
unknown	no
wd	coses
wording	NOT PRESENT
Unknown	
pos	ncfp000
WordNet	
sense	NOT PRESENT

*11376_pel.licula_text.tbf.xml

id content

spiderman3 es po de espide-man coses increibe es un aventura mol divertida

Node name	Func/Arg/Tem/...	Lemma	Contents
[- sentence			
n		spiderman3	spiderman3
p		es	es
n		po	po
s		de	de
n		espide-man	espide-man
n		cosa	coses
n		increibe	increibe
p		es	es
d		un	un
n		aventuara	aventuara
n		mol	mol
a		divertit	divertida
c		i	i
p		es	es
n		vai	vai
s		a	a
d		el	la
n		vania	vania
p		es	es
n		mol	mol
n		difcil	difcil
n		mol	mol
n		goapa	goapa
c		i	i
n		dvertida	dvertida
f		.	.

gen="f"
lem="cosa"
name="n"
num="p"
pos="ncfp000"
postype="common"
wd="coses"

Tree Text Data

Search Synchro morphol

Info

Word form: coses
Word lemma: cosa

Equivalent PAROLE tag: ncfp
Existing PAROLE tag (obs): ncfp000

All node names: a c d f i n p r s v w z

Node principal category: vm vs va nc np rg rn da dp di dd pr pp pi pd aq ao sps spc cc cs fp fc

Gender and Number: ms mp fs fp gen:c num:c nogen nonum

Attributes in node

Property	Value
case	NOT PRESENT
gen	f
lem	cosa
multiword	no
name	n
num	p
postype	common
title	no
titlelevel	NOT PRESENT
unknown	no
wd	coses
wording	NOT PRESENT

Corpus CLINT

Quin interès té un corpus d'aquestes característiques?

Quines línies de treball permet obrir?

Quina utilitat li veieu com a material de treball per als mestres/ professors?