

# Seguint la traça de CUCWeb: IAC

Toni Badia, José R. Boullosa, Judith Domingo, David Garcia, Albert Miguel (Grup de Veu i Llenguatge, Barcelona Media Centre d'Innovació)  
 Gemma Boleda, Stefan Bott (Universitat Politècnica de Catalunya)  
 Rodrigo Meza, Carlos Castillo, Vicente López (Cátedra Telefónica – Universitat Pompeu Fabra)

## CUCWeb (Corpus d'ús de català a la web)

- Recopilació de totes les pàgines del domini .es
- Extracció dels textos (36,2 GB)
- Selecció de les pàgines en català mitjançant un classificador estadístic que distingeix entre 10 llengües
- Filtratge del resultat (qualitat lingüística, detecció de duplicats...)
- Anotació lingüística (CatCG)

Resultados de la búsqueda  
 [lem="haber"] & [pos="Verb"]

#	Contexto	Copia local	Original
1	30 01 2003 Introducció L'aplicació "Persona" es va concebre amb la finalitat d'incloure en la base de dades	<a href="#">html</a>	<a href="#">#</a>
2	de la vostra unitat organitzativa, premeu aci. Si voleu publicar més dades personals al directori LDAP, tal com	<a href="#">html</a>	<a href="#">#</a>
3	comptes vells a la aplicació "Persona" s'han pogut produir errors involuntaris al crear les dades. Si aquest	<a href="#">html</a>	<a href="#">#</a>
4	a un grup d'investigació d'aquesta universitat. També poden accedir a aquesta màquina els estudiants i estudiants de segon cicle	<a href="#">html</a>	<a href="#">#</a>
5	Grup Eroski Ideasana Hobbies, Creativitat Ser una persona creativa significa saber gaudir millor de la vida, treure 'n tot el	<a href="#">html</a>	<a href="#">#</a>
6	Eroski Ideasana Hobbies, Creativitat Ser una persona creativa significa saber gaudir millor de la vida, treure 'n tot el suc	<a href="#">html</a>	<a href="#">#</a>
7	els més vulnerables, marginats i exclosos dels països empobrits puguin satisfer per si mateixos i de manera sostenible les seves	<a href="#">html</a>	<a href="#">#</a>
8	més vulnerables, marginats i exclosos dels països empobrits puguin satisfer per si mateixos i de manera sostenible les seves necessitats	<a href="#">html</a>	<a href="#">#</a>
9	a Publicacions Hobbies, Creativitat Ser una persona creativa significa saber gaudir millor de la vida, treure 'n tot el	<a href="#">html</a>	<a href="#">#</a>
10	Publicacions Hobbies, Creativitat Ser una persona creativa significa saber gaudir millor de la vida, treure 'n tot el suc	<a href="#">html</a>	<a href="#">#</a>

Resultats de la cerca

Estadísticas para esta búsqueda

Relativa	Frecuencia Cumulativa	Absoluta	Position 1	Lema	Position 2	View
72,92%	72,92%	1344	parlar	de	Prep	▶
9,16%	82,09%	169	parlar	amb	Prep	▶
5,64%	87,73%	104	parlar	en	Prep	▶
4,88%	92,62%	90	parlar	a	Prep	▶
3,58%	96,20%	66	parlar	sobre	Prep	▶
2,17%	98,37%	40	parlar	per	Prep	▶
0,37%	98,75%	7	parlar	entre	Prep	▶
0,32%	99,07%	6	parlar	sense	Prep	▶
0,16%	99,24%	3	parlar	durant	Prep	▶
0,16%	99,40%	3	parlar	fins	Prep	▶
0,16%	99,56%	3	parlar	des	Prep	▶
0,10%	99,67%	2	parlar	després	Prep	▶
0,05%	99,72%	1	parlar	contra	Prep	▶
0,05%	99,78%	1	parlar	mitjançant	Prep	▶
0,05%	99,83%	1	parlar	dins	Prep	▶
0,05%	99,89%	1	parlar	via	Prep	▶
0,05%	99,94%	1	parlar	davant	Prep	▶
0,05%	100,00%	1	parlar	segons	Prep	▶

Estadístiques

## Interfície de cerca

Corpus de Uso del Catalán en la WEB

Selecció de corpus:  
 WebCat A: 15 milions de paraules - subconjunt del total de pàgines en català en .es  
 WebCat B: 208 milions de paraules - total de pàgines en català en .es

Búsqueda simple  
 Palabra exacta  
 Categoría morfológica (Cat.): (Cualquiera)  
 Sintaxis: (Hay que elegir primero una categoría morfológica)

Búsqueda experta  
 Nota: en función de la complejidad de la búsqueda, los resultados pueden tardar más de un minuto

	Posición 1	Posición 2	Posición 3
Palabra	<input type="checkbox"/> NEG	<input type="checkbox"/> NEG	<input type="checkbox"/> NEG
Lema	<input checked="" type="checkbox"/> haber	<input type="checkbox"/>	<input type="checkbox"/>
Cat.	<input type="checkbox"/> Verbo	<input type="checkbox"/> Verbo	<input type="checkbox"/>
Sintaxis	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Uno o más	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Opcional	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Opciones de resultados  
 Posiciones para búsqueda: 3 posiciones  
 Contexto: 10 palabras  
 Resultados por página: 10  
 Núm. máximo de resultados: 100

http://www.catedratelefonica.upf.es/cucweb

## De CUCWeb a IAC (Interfície d'accés a corpus)

IAC és una interfície d'accés a corpus:

- dinàmica (s'adapta als atributs de cada corpus)
- per a corpus monolingües i bilingües
- multilingüe (cada usuari pot triar la llengua de visualització de la interfície)

### Arxiu de definició del corpus (xml)

```

<?xml version="1.0" encoding="UTF-8"?>
<corpus name="BT_cata" xmlns:cw="http://paies.upf.es/cucweb"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://paies.upf.es/cucweb/xsd/test.xsd">
  <corpus origin corpus id="cata" type="original" cw:language="ca">
    <positional_atts>
      <cw:pos_att id="word" obligatory="simple">
        <cw:tag cw:language="ca">Palabra/cw:tag</cw:tag>
        <cw:tag cw:language="es">Palabra/cw:tag</cw:tag>
        <cw:info cw:language="ca">Cerca de forasac/cw:info</cw:info>
        <cw:info cw:language="es">Búsqueda de forasac/cw:info</cw:info>
      </cw:pos_att>
      <cw:pos_att id="lema">
        <cw:tag cw:language="ca">Lema/cw:tag</cw:tag>
        <cw:tag cw:language="es">Lema/cw:tag</cw:tag>
        <cw:info cw:language="ca">Cerca de lemas/cw:info</cw:info>
        <cw:info cw:language="es">Búsqueda de lemas/cw:info</cw:info>
      </cw:pos_att>
      <cw:pos_att id="etiq">
        <cw:tag cw:language="ca">Etiqueta/cw:tag</cw:tag>
        <cw:tag cw:language="es">Etiqueta/cw:tag</cw:tag>
        <cw:info cw:language="ca">Introdueix l'etiqueta morfològica/cw:info</cw:info>
        <cw:info cw:language="es">Introduce la etiqueta morfológica/cw:info</cw:info>
      </cw:pos_att>
      <cw:pos_att id="pos">
        <cw:tag cw:language="ca">Categoría/cw:tag</cw:tag>
        <cw:tag cw:language="es">Categoría/cw:tag</cw:tag>
        <cw:info cw:language="ca">Introdueix la categoria morfològica/cw:info</cw:info>
        <cw:info cw:language="es">Introduce la categoria morfológica/cw:info</cw:info>
        <cw:value id="Neg">
          <cw:tag cw:language="ca">No/cw:tag</cw:tag>
          <cw:tag cw:language="es">No/cw:tag</cw:tag>
        </cw:value>
      </cw:pos_att>
      <cw:pos_att id="fuso" valid_pos_att="pos">
        <cw:tag cw:language="ca">Funció/cw:tag</cw:tag>
        <cw:tag cw:language="es">Función/cw:tag</cw:tag>
        <cw:info cw:language="ca">Introdueix la funció morfològica/cw:info</cw:info>
        <cw:info cw:language="es">Introduce la función morfológica/cw:info</cw:info>
      </cw:pos_att>
      <cw:pos_att id="cfk">
        <cw:tag cw:language="ca">CFK/cw:tag</cw:tag>
      </cw:pos_att>
    </corpus origin>
  </corpus>
  
```

Atributs que afecten un sol mot (posicionals)  
 Representació de la cerca

Adapta la interfície al corpus

Área de configuración de CUCWeb

Subir corpus a CUCWeb

Elige el idioma de la aplicación:

Representació de la cerca

### Interfície de càrrega de corpus

L'usuari ha de proporcionar el corpus en el format llegible per CWB i l'arxiu de definició del corpus.

La interfície s'adapta, a partir de l'arxiu de definició, i el corpus s'indexa automàticament.

Opcions de presentació

## Interfície de cerca

Búsqueda simple | Búsqueda avanzada | Configuración

Selecciona un corpus:

Corpus origen

Condición 1	Condición 2
Palabra ? <input type="text"/>	Palabra ? <input type="text"/>
Lema ? <input type="text"/>	Lema ? <input type="text"/>
Categoría ? <input type="text"/>	Categoría ? <input type="text"/>
Opcionalidad ? <input type="text"/>	Opcionalidad ? <input type="text"/>

Corpus destino

Condición 1	Condición 2
Palabra ? <input type="text"/>	Palabra ? <input type="text"/>
Lema ? <input type="text"/>	Lema ? <input type="text"/>
Categoría ? <input type="text"/>	Categoría ? <input type="text"/>
Opcionalidad ? <input type="text"/>	Opcionalidad ? <input type="text"/>

Escoge estructural

Atrs. estructurales:

Tipo de error:

Condición 1 | Condición 2

Tipo de error: Error de sentido | Traducción literal

Metadatos

Corpus de origen	Corpus de destino
Tipo de texto: Cualquiera	Tipo de texto: Periódico
Curso: Cualquiera	Curso: Cualquiera
Tipo traducción: General	Tipo traducción: Cualquiera
Profesor/Corrector: Cualquiera	Profesor/Corrector: Cualquiera

Opciones de pantalla

Contexto ?

Resultados por página ?

Número máximo de resultados ?

Atributs que afecten un o més mots (estructurals)

Restricció per metadades