



# Tresna linguistikoak informazioa atzitzeko

Eneko Agirre, Iñaki Alegria

**Ixa taldea**

<http://ixa.si.ehu.es>



# Aurkezpena

- Sarrera
- IR (informazioaren bilaketa)
- Oinarrizko tresna linguistikoak
- Morfologian oinarritutako aplikazioak
- Sintaxian oinarritutako aplikazioak
- CLIR (eleaniztasuna)
- IR multimodala
- QA (galdera-erantzunak)

# Ixa taldea





# Ixa taldea

- 6 esplotazio-lizentzia (patente)
- Spin-off enpresa: Eleka
- Hainbat produktu merkatuan (lankidetzan)
  - *Eleka, Elhuyar, Elkar, Euskaltel, ...*
  - *Microsoft, Scansoft, Eaton, ...*
- Hainbat prototipo aurrekomertzial
- Adibideak:
  - *Xuxen* zuzentzaile ortografikoa: [www.xuxen.com](http://www.xuxen.com)
  - *Opentrad*-Matxin itzultzaile automatikoa: [www.opentrad.com](http://www.opentrad.com)
  - Beste asko

ixA2007Aurkezpena Ixa Taldea - Produktuak - Mozilla Firefox

Etxategia Editatu Ikusi Historia Laster-markak Tresnak Hiztegiak Laguntza

http://ixa.si.ehu.es/Ixa/Produktuak

ixA **Lengoaia Naturala**

Ixa > Produktuak

Bilatu

**Produktuak**

- [Etxera](#)
- [Aurkezpena](#)
- [Kideak](#)
- [Ikerlerroak](#)
- [Argitalpenak](#)
- [Produktuak](#)
- [Proiektuak](#)
- [Estekak](#)
- [Pil-Pilean](#)
- [Demoak](#)
- [Bestelakoak](#)
- [Pribatua](#)

Ixa Taldea  
649 Posta kutxa  
20080 Donostia  
Harremanetarako:  
[acpalloi@si.ehu.es](mailto:acpalloi@si.ehu.es)

Euskal Herriko Unibertsitatea

Informatika Fakultatea  
Lengoaia eta Sistema Informatikoak Saila

[Xuxen eskuragarri](#)

[ERREUS](#)

[Eulia](#)

[Xuxen-Mac eskuragarri](#)

[EusWN, euskarazko WordNet](#)

[Elhuyar-Word hiztegi-sistema](#)

[Multimeteo euskaraz](#)

[EDBL: Euskararen Datu-Base Lexikala](#)

[Diccionario Básico Escolar \(Ikaslearen oinarrizko hiztegia\)](#)

[BertsolariXa](#)

[Patenteak](#)

[Sariak](#)

Etxategia Editatu Ikusi Historia Laster-markak Tresnak Hiztegiak Laguntza

http://ixa.si.ehu.es/Ixa/Ikerlerroak

Ixa Taldea - Artikuluak

**Ixa Taldea - Ikerlerroak**

**Tresnak**

	Lehen	Orain	Gero
Corpus	Corpusak sortu eta lantzeko tresnak	Corpusak sortu eta lantzeko tresnak Lexiko-eskurapen automatikoa: Terminologia	Corpusak sortu eta lantzeko tresnak
Lexikoa	Hiztegien bertsio elektronikoak - Ing-Eusk Morris - Gazt-Eusk Elhuyar - EH Ibon Sarasola		Lexikografoarentzako lan-postua (workbench)
Morfologia	Analizatzaile/Sortzaile morfologikoa Lematizatzaile/Etiketatzailea	Hobetzen Hobetzen	
Sintaxia	Azaleko sintaxia: - Funtzio sintaktikoak - Zatiak ( <i>Chunks</i> )	Hobetzen - Anbiguotasun sintaktikoaren ebazpena - Esaldi-mugak - Postposizioak - Aditz-azpikategorizazioa - Menpekotasunak	Parserra Estaldura zabala Eraginkorra Hainbat formalismo <b>- Murritzapen Gram. (CG)</b> <b>- Baterakuntza</b> <b>- Estatistikoak</b>
Semantika	Adiera-desanbiguzioa (WSD)	WSD hobetzen Ezagutza eleanitza	WSD hobetzen Anlisi semantikoa
Integrazioa	Tresnen integrazioarako ingurune informatikoa - XML estandarra - TEI gida-lerroak	Tresna berriak integratzen - morfosintaxia - sintaxia	Hobetu

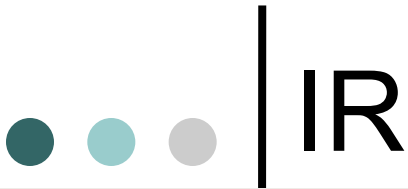
**Aplikazioak**

47619.pdf



# IR

- Bilaketen teknologia
  - Internet: *Google, Yahoo, Elebila, ...*
  - Intranet: Hemerotekak, liburutegi digitalak, ...
- Teknologia
  - Tresna orokorrak (plug-in/gehigarriak):
    - Jabedunak: *Autonomy, ...*
    - Libreak: *Lucene, Greenstone, ...*
  - Neurrira egindakoak
- Ebaluazioa:
  - *Precision* (doitasuna)
  - *Recall* (estaldura)
- Aldaerak
  - CLIR, IR multimodala, QA



leioa - fitxategi-arakatzaila

hemerote...

Fitxategia Editatu Ikusi Historia Laster-markak Tresnak Hiztegiak Laguntza

http://www.google.com/search?num=100&hl=eu&q=hemeroteques&btnG

bidaiak moodle eurOp kongresuak EHU TR tekniker udako ikast izaskun

Aukerak ES EU Euskalterm 3000 Elhuyar Hiztegi Batua Itzul

Webgunea Inudiak Maps Taldeak Gida

Google hemeroteques Bilatu Bilaketa aurreratu Hobespenak

Webgunea

**PROMOTORA de SOMNIS | Les pancartes i les hemeroteques**  
Les pancartes i les **hemeroteques**, promotora | 11 Març, 2007 22:51. "La política no es cosa de agitación, sino de ideas. No es cosa de pancartas, ...  
[promotoradesomnis.baleaerweb.net/post/31007](#) - 15k - [Katxean](#) - [Antzeko orriak](#)

**Hemeroteques. Estat actual i perspectives - Dialnet**  
Información del artículo **Hemeroteques. Estat actual i perspectives**.  
[dialnet.unirioja.es/servlet/articulo?codigo=1291113](#) - 11k - [Katxean](#) - [Antzeko orriak](#)

**IV Col·loqui Aula d'Historia del Periodisme Diari de Barcelona ...**  
**Hemeroteques** digitals: experiències. Autores: Lluís Codina Bonilla; Localització: Treballs de comunicació, ISSN 1131-5687, N.º 16, 2001, pags. ...  
[dialnet.unirioja.es/servlet/articulo?codigo=271331](#) - 10k - [Katxean](#) - [Antzeko orriak](#)  
[dialnet.unirioja.es](#) domeinuan emaitza gehiago >

**MONTCADA CONFIDENCIAL: HEMEROTEQUES ...**  
**HEMEROTEQUES** ... (PER VEURE LA IMATGE MÉS GRAN, "CLICA" SOBRE LA MATEIXA). RECOMANEM LLEGIR ATENTAMENT AQUEST ARTICLE PUBLICAT A LA VEU (L'ANY 1999) I QUE ...  
[montcadaconfidencial.blogspot.com/2007/05/hemeroteques.html](#) - 182k - [Katxean](#) - [Antzeko orriak](#)

**MONTCADA CONFIDENCIAL: SEGUEIXEN ... LES HEMEROTEQUES ...**  
**LES HEMEROTEQUES** ... (per veure més gran la imatge, clica sobre la mateixa) CARTA ADREÇADA PER JOAN MARESMÀ ALS ELECTORS, ANY 1995. ...  
[montcadaconfidencial.blogspot.com/2007/05/seguixen-les-hemeroteques.html](#) - 167k - [Katxean](#) - [Antzeko orriak](#)  
[montcadaconfidencial.blogspot.com](#) domeinuan emaitza gehiago >

[PDF] **HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**  
Formatua: PDF/Adobe Acrobat - [HTML bertsiog](#)  
**HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**. PROYECTO DE DIGITALIZACIÓN DE, PrensA EN LA BIBLIOTECA NACIONAL, Dolores Rodríguez, ...  
[eprints.rclis.org/archive/00004465/01/nacional.pdf](#) - [Antzeko orriak](#)

[PDF] **HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**  
Formatua: PDF/Adobe Acrobat - [HTML bertsiog](#)  
**HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**. HEMEROTECA DE LA BIBLIOTECA PÚBLICA. DEL ESTADO AZORÍN DE ALICANTE. FICHA TÉCNICA: ...  
[eprints.rclis.org/archive/00004495/01/alicante.pdf](#) - [Antzeko orriak](#)  
[eprints.rclis.org](#) domeinuan emaitza gehiago >

[PDF] **HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**  
Formatua: PDF/Adobe Acrobat - [HTML bertsiog](#)  
**HEMEROTEQUES. ESTAT ACTUAL I PERSPECTIVES**. LA HEMEROTECA MUNICIPAL DE

Elebila - Mozilla Firefox

Fitxategia Editatu Ikusi Historia Laster-markak Tresnak Hiztegiak Laguntza

http://www.elebila.eu/search/?bilatu=hemerotekak&bot\_bilatu=Bilatu&li

bidaiak moodle eurOp kongresuak EHU TR tekniker udako ikast izaskun MAMT apache2 clef

Aukerak ES EU Euskalterm 3000 Elhuyar Hiztegi Batua Itzul Harluxet Mokoroa ZT Corpora

eu | es Txertatu nabigatzailearen tresna-barran **Berria!** | Laguntza | Honi buruz

**elebila** hemerotekak **BILATU** Bilaketa aurreratu Hobespenak

Euskarazko web orrietan Edozein hizkuntzatan

Emaizak: 36700 orri

Erabilitako analisia: *hemeroteka* izena.

**Untitled Document**  
**HEMEROTEKA** ... bat adostea ez da erraza, parte-hartze maila eta ... daudelako eta garapenerako estrategia desberdinak ere badirelako. Testuaren helburua, beraz, ez da ...  
[http://www.uztaro.com/berria/laburpena.cfm?labur=fitxategiak\Uztaro63\Uztaro63\\_6.Villalba.txt](#) Katxean

**Untitled Document**  
...  
[http://www.uztaro.com/berria/laburpena.cfm?labur=fitxategiak\Uztaro60\5\\_Lizaso.txt](#) Katxean

**Ameriketako euskaldunei buruzko webgunea - Website on Basques ...**  
**Hemeroteka:** Berri gehiago : Laburrak: Agurrak: d: dd ... Buffalon XX mende hasierakoa bada ere, kluba ez zen 1980 ... Bertan hautatzen da urterako lehendakaria eta zuzendaritza taldea.  
[http://www.basqueheritage.com/berria/eusk/index.php?id=fit&idf=56&PHPSESSID=bca39099488c1b0614d9273df93fc617](#) Katxean

**Kritikak (hemeroteka)**  
...  
[http://www.susa-literatura.com/kritikak/argia/krit0262.htm](#) Katxean

**Kritikak (hemeroteka)**  
...  
[http://www.susa-literatura.com/kritikak/krit0399.htm](#) Katxean

**UZEI - Terminologia eta Lexikografia Zentroa**  
**Hemeroteka** ... hizkuntzaren egungo bizkortasunaren seinalea da, eta etorkizunean irauteko bermea. Hori 1977an horrela esaten ez bazen ere ...  
[http://www.uzei.com/antbuspre.asp?nombre=1682&cod=1682&sesion=14](#) Katxean

**UZEI - Terminologia eta Lexikografia Zentroa**  
...  
[http://www.uzei.com/antcatalogo.asp?nombre=1701&hoja=0&sesion=14](#) Katxean

**Bide Ertzean: Hemeroteka**  
**Hemeroteka** Mondo Sonoro aldizkaria. 06 ... baita beste toki askotan ere. "Non dira"

Opentrad  
Itzulpen automatikoa kode irekian

**XUXENweb**

**Zure enpresa ager daiteke hemen**  
Zure enpresa hemen ikusi nahi baduzu idatzi [info@eleka.netera](#)  
[http://www.eleka.net](#)

**Zure enpresa ager daiteke hemen**  
Zure enpresa hemen ikusi nahi baduzu idatzi [info@eleka.netera](#)  
[http://www.eleka.net](#)

**Zure enpresa ager daiteke hemen**  
Zure enpresa hemen ikusi nahi baduzu idatzi [info@eleka.netera](#)  
[http://www.eleka.net](#)



# IR

- Bizitza *google* baino lehen
  - emaitza eskasagoak
  - abiadura motela
  - galdera zehatzagoak
  - ontologiaren erabilpena
- Egoera korapilatsuak
  - erantzunik ez/gutxi (estaldura handitu behar)
  - erantzun gehiegi (galdera findu behar)
    - *relevance-feedback*





# Hemerotekak

- Informazio egituratuagoa eta laburragoa
- Metadatuak:
  - bilaketa-estrategia matadatuaren arabera
- Informazio multimodala?
- Estandarrak informazio-trukerako
  - *Dublin Core, MARC* (liburutegi digitalak)
- IR ezaugarriak
  - galdera osoagoak?
    - bilaketa aurreratua
  - denbora gehiago?
  - estaldura oso garrantzitsua

# Hemeroteca

Fitxategia Editatu Ikusi Historia Laster-markak Irresnak Hiztegiak Laguntza

http://www.cervantesvirtual.com/busquedas/busqueda\_textos.jsp eca digital cervantes

bidaiak moodle eurOp kongresuak EHU TR tekniker udako ikast izaskun MAMT apache2 clef linux laura termino AhoTTS

Aukerak ES EU Euskalterm 3000 Elhuyar Hiztegi Batua Itzul Harluxet Mokoroa ZT Corpusa XUXENweb Opendrad


Consulte los contenidos más recientes de esta sección.

**Sugerencias**

**Efemérides**

**Noticias**

**Suscripción al Boletín**

 PRIMERA VISTA

**Búsqueda de**

Las PALABRAS

en **TEXTO DE LAS OBRAS**

**TEXTO DE LAS OBRAS**

SOLO EN PÁRRAFOS

SOLO EN VERSOS

SOLO EN CITAS

Las PALABRAS

en

Todos

los parlamentos del PERSONAJE

**Buscar**

Las PALABRAS

Todas las PALABRAS

en **CUALQUIER IDIOMA** distinto al de la obra

**Buscar**

**Restringir por**

TÍTULO de la obra

AUTOR de la obra

PERÍODO de

Escritura de la obra

Edición

entre  y

**Opciones**

Buscar frase exacta

Buscar palabras sueltas ordenadas

Buscar palabras sueltas no ordenadas

Buscar en el interior de las notas del ed.

Distinguir mayúsculas/minúsculas

Ordenar resultados

**Mostrar**

FUNDACIÓN BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES



# Tresna linguistikoak

- morfologia/lematizazioa/*stemming*
  - estaldura handitzeko (hizkuntzen arabera)
  - galderen akatsak zuzentzeko
- sintaxia + estatistika
  - terminologia, pertsona/toki/erakunde izenak
  - informazio interesgarriena eskaintzeko
- semantika
  - informazioa erlazionatzeko
  - estaldura handitzeko (epaile/magistratu)
- eleaniztasuna (hiztegiak, itzulpen automatikoa)
  - hemeroteka eleanitzak
  - erabiltzaile eleanitzak
  - hizkuntza ez-ezagunak
- modu multimodala (hizketaren ezagutza, irudien sailkapena)
  - irudi, hizketa, bideo gainean bilatzea

# Itzulpen automatikoa: *Matxin*

demo - OpenTrad - Mozilla Firefox

Eitxategia Editatu Ikusi Historia Laster-markak Irresnak Hiztegiak Laguntza

http://www.opentrad.org/demo/libs/nabigatzailea.php?language=eu

UPV-EHU Google http://www.euskaltz... Inscripción en la Que...

Aukerak ES EU Euskalterm 3000 Elhuyar Hiztegi Batua Itzul Harluxet Mokoroa ZT Corpusa XUXENweb Opentrad

demo - OpenTrad quijote - Buscar con Google CVC. Don Quijote de la Mancha. Capit... Google News España

Portadan Espainia Ir Generada automáticamente hace 15 minutos

**Portadan**

- Nazioartekoa
- Espainia
- Ekonomia
- Tecn zientzia .
- Kirolak
- Ikuskizunak
- Osasuna
- Herritar GEHIAGO

☑ Berrien alertatuak

Atom RSS | Feeds-ei buruz

Higikorrenzat berriak

**Portadan** Espainia Ir Generada automáticamente hace 15 minutos


**Ratzinger: beldurraren estrategia**

Egunean (Mexiko) **1 ordu Egin** - Haren 1 bisita da Eliza katolikoaren bastioira. Benedicto XVI Brasilen egongo da Kariberean landetxe Conferencia General De El Episcopado Latinoamericano inauguratzeko eta (CELAM), gertatuko den martxoaren 13an ekintza. Latinamerika-erangan ia bizi da ...

Sao Paulori blindatzen da eta edertzen da hartzeko Benedicto Xvi La Razón-i (Espainia)

Munduaren herri katoliko gehiagok Benedicto Xvi La Gaceta Tucumán-i zain dagokio

Merkataritza (Ekuatorea) Terra España - Univisión - EITB [eta 572 gai erlazionatu](#) - »




Ideal Digital

**Dick Cheney Irak-era iristen da ustekabe bisitan**

**2 ordu egin** Terra España - Lehendakariorde estatubatuarra, Dick Cheney, Asteazken honetan iritsi zen ustekabe bisita batean Bagdadera erreklamazioa egiteko irakiar liderretara bikoitz dezaten adiskidetzee nazionalaren alde haren esfortzuak, bonba kamioi batek utzi zuen mementoetan ...

Irak Clarín-i EE.UU. lehendakariordearen ustekabeko bisita Com-a Dick Cheney Bagdadera iristen da ustekabe bisita batean ABC

Télam - Swiss Info - Voz De América - Abangoardia [eta 58 gai erlazionatu](#) - »



Voz De América

**Recomendada berriak kontsulta egin ditzan**

Recomendada berriak kontsulta egin ahal edun-tzeko onets beza bilaketen

**Orrialde hau izena aipatzea**

**Bulegoa**  
Diario Palentino - [eta 212 gai erlazionatu](#) - »

**Google jarrera hartzearen kontra borrokatzen da bilatzaileetan**  
Herria (Espainia) [Eta 16 gai erlazionatu](#) - »

**Garbajosa-k soilqune uzten du ahal badu, Europarra jolastuko da**  
Marka - [eta 17 gai erlazionatu](#) - »

**José Coronado espainiar Grissom izango da Telecinco-n**  
ABC - [eta 23 gai erlazionatu](#) - »

**Hobetzeko plana lehen arretra eta estaldura landatarra**  
Europa Sur - [eta 14 gai erlazionatu](#) - »

**Berria da:**

<a href="#">Ian Paisley</a>	<a href="#">Isabel II</a>
<a href="#">Martin Mcguinness</a>	<a href="#">Dow Jones</a>
<a href="#">Diego Maradona</a>	<a href="#">Néstor Kirchner</a>
<a href="#">Roger Federer</a>	<a href="#">Auzitegi Gorena</a>
<a href="#">San Luis Potosí</a>	<a href="#">George W. Bush</a>

http://www.opentrad.org/demo/libs/nabigatzailea.php?markatu=&norantza=es-eu&inurl=http://www.lagaceta.com.ar/vernotae.asp?id\_nota=217363&titulo=El%20pa%C3%ADs%20m%C3%A1s%20...

Inicio Microsoft Mis doc... demo - ... sisx04... Posta E... Bandeja... matxin... 14:12

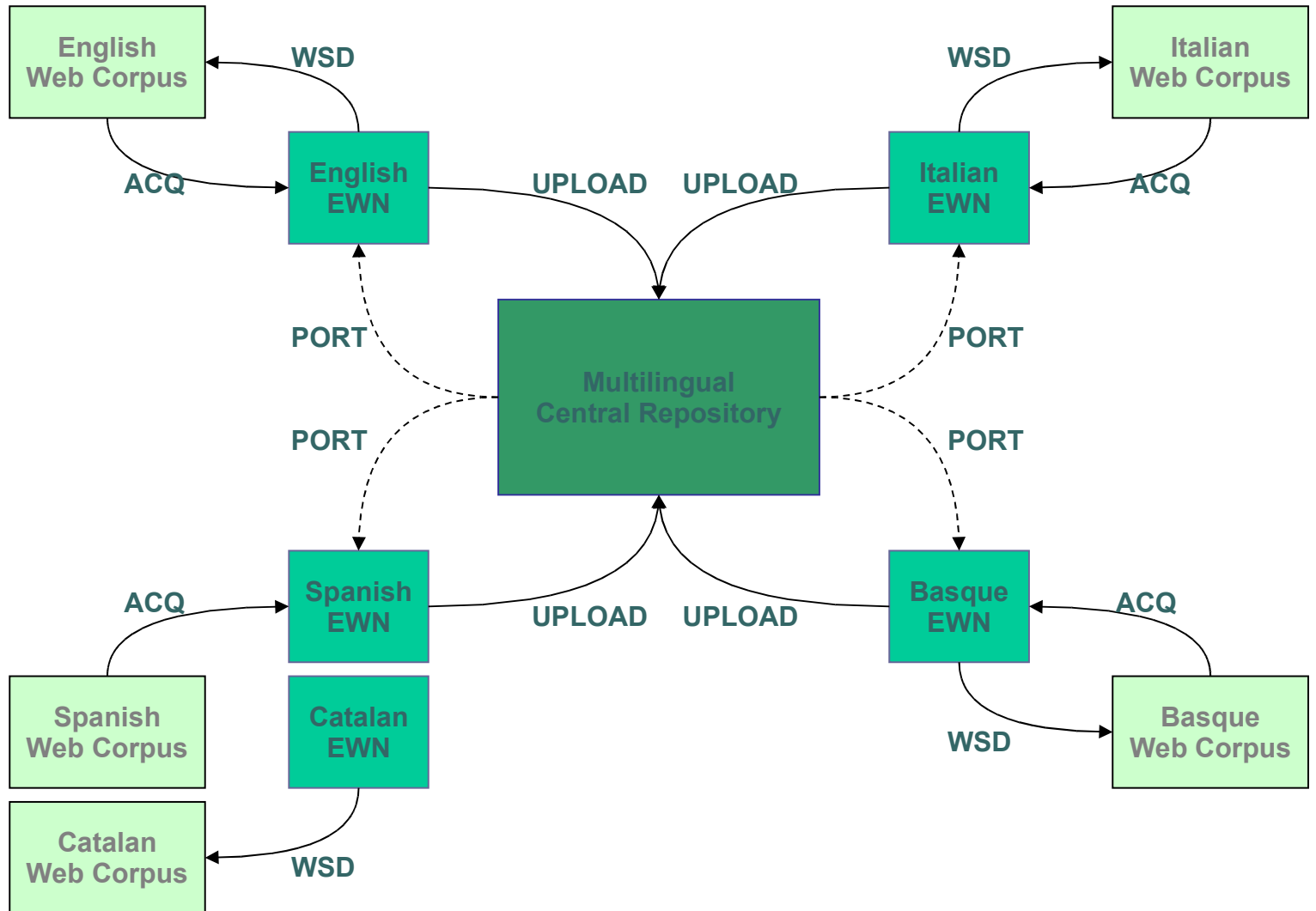


# Semantika: MCR

*(Multilingual Central Repository)*

- Hizkuntza desberdinetako kontzeptuak MCRn daude gordeta
- Erlazioa hizkuntzen artean eta kontzeptuen artean
- Kontzeptuen artean erlazio konplexuak lortzea edo inferitzea
  - *Zer egiten da kafearekin? Edan*

# MCR



# WordNet

Web MCR Interface MEANING - Microsoft Internet Explorer

Fitxategia Editatu Ikusi Gogokoak Tresnak Laguntza

Atzera Bilatu Gogokoak

http://nipadio.lsi.upc.edu/cgi-bin/wei4/public/wei.consult.perl

Google eneko agirre

Go Bookmarks 0 blocked Check AutoLink AutoFill Send to

java

Word Nouns English\_1.6

Synonyms near\_synonym English\_1.6

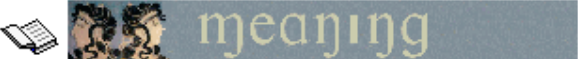
Gloss  English\_1.6  English\_1.7

Score  Spanish\_1.6  English\_1.7.1


Rels  Catalan\_1.6  English\_2.0

Full  Basque\_1.6  Catalan\_1.5


Italian\_1.6  Spanish\_1.5


 **meaning** Multilingual Central Repository


---


05948884n  11 [coffee\\_1](#) [java\\_2](#)


[alimentation](#) food a beverage consisting of an infusion of ground coffee beans: *he ordered a cup of coffee;*

[Beverage+](#) 05948884n  17 [café\\_1](#) Bebida estimulante que se obtiene de los granos del cafeto

[Comestible+](#) 05948884n  17 [café\\_1](#) Beguda estimulant que s'obté dels grans de l'arbre del café

[Liquid+](#) 05948884n  12 [kafe\\_2](#)

[Natural+](#) 05948884n  12 [caffè\\_1](#)

[Substance+](#) 05948884n  12 [caffè\\_1](#)

5 gloss 5 role agent 2 has mero madeof 38 rgloss 1 has hyperonym 44 role patient 11 has hyponym

2 has mero madeof 1 has hyperonym 16 has hyponym

2 has mero madeof 1 has hyperonym 16 has hyponym

3 role agent 2 has mero madeof 1 has hyperonym 33 role patient 11 has hyponym

5 role agent 2 has mero madeof 1 has hyperonym 20 role patient 11 has hyponym

# Morfologian oinarritutako aplikazioak

- Lematizazioa/*stemming*
  - funtzioak:
    - hemerotekaren lematizazioa datu-basera eraman aurretik
    - galderan oinarritutako sorkuntza
  - galderaren lematizazioa
- Zuzenketa
  - hitz arraroen (erantzun gabe/gutxi) aurrean proposamenak

eu | es

Txertatu nabigatzailearen tresna-barran **Berria!** | Laguntza | Honi buruz

eLebila

zuhaitsetikan

BILATU

Bilaketa aurreratua  
Hobespenak



inguruetan  
EUSKALTZAINDIA  
inguruetan

Euskarazko web orrietan Edozein hizkuntzatan

Emaitzak: 12300 orri

Erabilitako analisia: [zuhaitz izena](#).

Sartutako testuan aldaerak daude: [zuhaitzetik](#)

Aldaerak: [zuhaitzetik](#), [zuhaitzetio](#)

Google  
Scholar BETA

diachronic linguistics

Search

Scholar All articles - [Recent articles](#)

Did you mean: [diachronic linguistics](#)

[Arabic Phonology](#)

[F. Corriente - Jewish Quarterly Review 1972 - JSTOR](#)





zientziaren  
ELHUYAR  
komunikazioa

# zientzia.net

Eguneratze-data 2004/6/23



## Bilaketaren emaitza

284 kasu aurkitu dira **saguarekin** hitzarekin



< Aurrekoa Hurrengoa >

- Artikuluak
- Argazkiak
- Sarean
- Agenda
- Hileko zerua
- Dosierak
- Elhuyar aldizkaria
- Zientzia-hiztegiak
- Euskara teknikoa
- Dibulgazio-liburuak
- CAF-Elhuyar sariak

[asteko laburpena]  
[gaiak]

## Artikuluak euskaraz (284)



**6.- Saguaren obesitatea murriztu...**  
Massachusetsko Millennium Pharmaco...  
obesitatea kontrolatzeko erabili daite...

**Osasuna** > Osasuna



**7.- Saguaren hiru dimentsioko a...**  
Erresonantzia magnetikoaren bidezko  
atlasa aurkeztu dute Kaliforniako iker...

**Biziaren zientziak** > Biologia

**Biziaren zientziak** > Genetika



**8.- Etxe-sagua, gizakiaren apro...**  
Munduan 1.500etik gora karraskari-es...  
Orden honen barruan, muridoen fami...  
eta familia honetakoak diren arratoia...  
daitezkeelako.

**Biziaren zientziak** > Zoologia

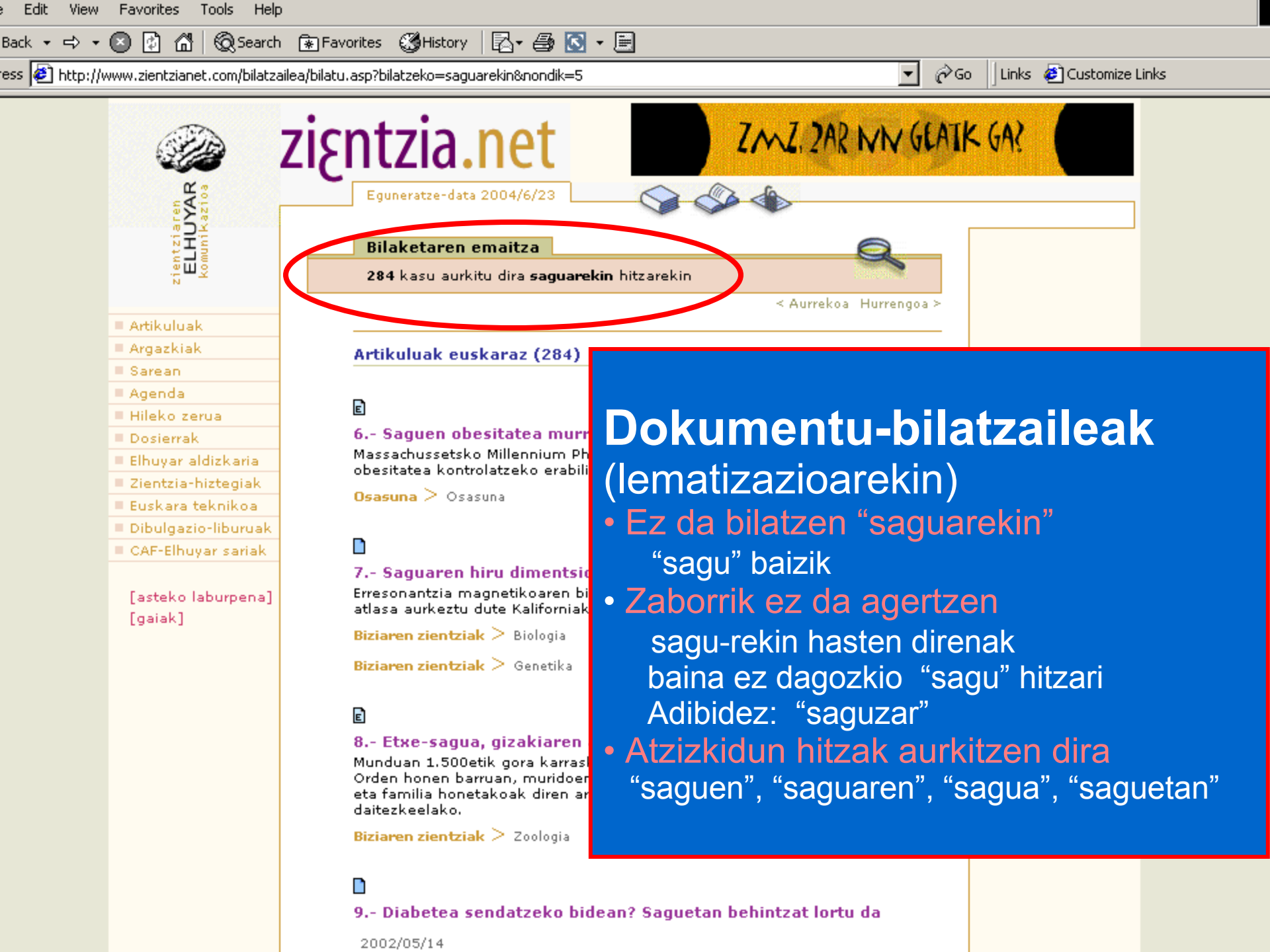


**9.- Diabetea sendatzeko bidean? Saguetan behintzat lortu da**

2002/05/14

## Dokumentu-bilatzaileak (lematizazioarekin)

- Ez da bilatzen "saguarekin"
- "sagu" baizik
- Zaborrik ez da agertzen
- sagu-rekin hasten direnak baina ez dagozkio "sagu" hitzari Adibidez: "saguzar"
- Atzizkidun hitzak aurkitzen dira "saguen", "saguaren", "sagua", "saguetan"



Artikuluak euskaraz (284)

6.- Saguaren obesitatea murr  
Massachussetsko Millennium Ph  
obesitatea kontrolatzeko erabili

Osasuna > Osasuna

7.- Saguaren hiru dimentsio  
Erresonantzia magnetikoaren bi  
atlasa aurkeztu dute Kaliforniak

Biziaren zientziak > Biologia

Biziaren zientziak > Genetika

8.- Etxe-sagua, gizakiaren  
Munduan 1.500etik gora karras  
Orden honen barruan, muridoer  
eta familia honetakoak diren ar  
daitezkeelako.

Biziaren zientziak > Zoologia

9.- Diabetea sendatzeko bidean? Sagueta behintzat lortu da

2002/05/14

## Dokumentu-bilatzaileak (lematizazioarekin)

- Ez da bilatzen “saguarekin”  
“sagu” baizik
- Zaborrik ez da agertzen  
sagu-rekin hasten direnak  
baina ez dagozkie “sagu” hitzari  
Adibidez: “saguzar”
- Atzizkidun hitzak aurkitzen dira  
“saguen”, “saguaren”, “sagua”, “sagueta”



# Sintaxian oinarritutako aplikazioak

- Informazio esanguratsua:
  - izen-sintagmak, terminoak eta izen propioak
- Funtzioa:
  - bilatzea
  - estekatzea
  - multzokatzea (*clustering*)
  - galdera fintzea
- Aplikazioak:
  - bistaratze bereziak
  - dokumentuen multzokatzea (Eleka)

# IR aurreratua

The image displays three overlapping browser windows illustrating advanced search results for the query "Clinton Obama".

- Left Window (Clusty Search):** Shows search results categorized into clusters. The top cluster is "Hillary Clinton and Barack Obama" (28 results), with sub-clusters for "Debate" (28), "Pennsylvania" (22), "McCain" (24), "Bill Clinton" (12), "Obama, Edwards" (12), "Blog" (9), "Bush" (6), "Clinton-Obama ticket" (5), and "Clinton Answers" (4). A "remix" button and a "find in clusters" search box are also visible.
- Middle Window (Quintura):** Displays a word cloud of terms related to the search, including "mocks", "feud", "clash", "bill", "complain", "campaign", "comparing", "ticket", "barack", "clinton", "obama", "rodham", "lead", "returns", "attacks", "senator", "camp", "questions", "debate", and "trade". A list of search results is visible on the right side of the window.
- Right Window (KartOO Metamotor):** Shows a semantic map of the search results. The map features nodes for various topics and sources, such as "report", "blog.washingtonpost.com", "times", "democr", "www.usnews.com", "lead", "news", "news.aol.com", "mccain", "vote", "barack", "points", "shows", "chart", "www.factcheck.org", "election", "www.cnn.com", "www.politifact.com", "blog", "ticker", "www.truthorig.com", "contest", "politics", "washington", "www.nytimes.com", "www.usnews.com", "lead", "news", "news.aol.com", "mccain", "vote", "barack", "points", "shows", "chart", "www.factcheck.org", "election", "www.cnn.com", "www.politifact.com", "blog", "ticker", "www.truthorig.com", "contest", "politics", "washington", "www.nytimes.com". A "Topics" sidebar lists related terms like "hillary clinton", "rival barack obama", "clinton and barack obama", "john mccain", "percentage points", "hillary rodham clinton", "york times", "hillary", "points", "rival", "barack", "chart", "election", "politics", "news", "contest", and "ticker". A "Sponsor" section highlights "Democrats Face Off" from The New York Times.



# CLIR

- IR hainbat hizkuntzetan
- Aukerak
  - galderak itzultzea
  - dokumentuak itzultzea
  - semantika bidez proiektatzea
- Gure aukera:
  - MCR
  - Etorkizunean itzulpen automatikoa
- Adibideak: *EFE (Meaning), ArgazkiPress*


# CLIR (EFE)

http://efe.irion.nl/efe\_D/web/init.do?queryLg=en - Microsoft Internet Explorer

File Edit View Favorites Tools Help


Back Forward Stop Home Search Favorites Media Print Copy Paste


Address http://efe.irion.nl/efe\_D/web/init.do?queryLg=en Go Lin


English  fire chemical plant Best phrase OK Reset all New task ? 21

Search in results Results per page 10 [Show advanced options](#)

Sort on : ===== OK 1 | 2 | 3 | Next » Result(s): 1975 hit(s)  
25 hit(s) processed

75.0% 20040521 [CATEGORÍAS SUPLEMENTARIAS: JUSTICIA-INTERIOR-SUCESOS/SUCESOS PALABRAS](#)  
[CLAVE: JUSTICE,ACCIDENTS CRIME INCENDIOS / INCENDIO EN FÁBRICA QUÍMICA,](#)  
[VALENCIA 2004. FUEGO / HUMO NEGRO / CARRETERA / COCHES CT](#)   
ACCIDENTS CRIME INCENDIOS/ INCENDIO EN F BRICA QU MICA , VALENCIA 2004 . **FUEGO** / HUMO NEGRO/ CARRETERA/ COCHES CT INCENDIO FABRICA : V. 11 ..... 2004 . **Una inmensa columna de humo sale del incendio que se ha declarado esta tarde en una fábrica química dedicada al tratamiento del mármol en la localidad de San Antonio de Benagéber , a**

75.0% 20040521 [CATEGORÍAS SUPLEMENTARIAS: JUSTICIA-INTERIOR-SUCESOS/SUCESOS PALABRAS](#)  
[CLAVE: JUSTICE,ACCIDENTS CRIME INCENDIOS / INCENDIO EN FÁBRICA QUÍMICA,](#)  
[VALENCIA 2004. FUEGO / HUMO NEGRO / TENDIDO ELÉCTRICO / CURIOSOS CT](#)   
ACCIDENTS CRIME INCENDIOS/ INCENDIO EN F BRICA QU MICA , VALENCIA 2004 . **FUEGO** / HUMO NEGRO/ TENDIDO EL CTRICO/ CURIOSOS CT INCENDIO FABRICA ..... 2004 . **Varios residentes de la urbanización adyacente al incendio que se ha declarado esta tarde en una fábrica química dedicada al tratamiento del mármol en la localidad de San Antonio de Benagéber , a**

58.0% 20040428 [CATEGORÍAS SUPLEMENTARIAS: EMERGENCY PLANNING TERRORISMO SIMULACRO DE](#)  
[ATAQUE TERRORISTA CON ARMAS QUIMICAS EN NEWCASTLE BOMBEROS POLICIA JGB NO](#)   
VENDED EN REINO UNIDO MIT IRLANDA

Internet

# CLIR (EFE)

## News Article 10

CONTEXT = Sigue la violencia en Colombia y especialmente en Medellín.

GOAL = Un entierro en Medellín.

QUERY = entierro medellín

TEXT = sepelio medellín

RESULT = FH\_1205173 20040524

RESULT = FH\_1205172 20040524

<entierro #35, sepelio #14, enterramiento #7> = <burial, funeral>





# CLIR: ebaluazioa

---

	Hitza	Hobetua
Actions	295	168
Pictures	20	24

---

MCR+desanbiguazio semantikoa:

- Ekintza gutxiago
- Argazki gehiago

SemEval-2007 / CLEF-2008 exercise

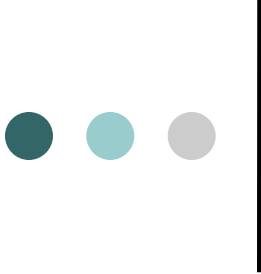
- Semantikaren eragina CLIR eta Q&A) arlotan





# IR multimodala

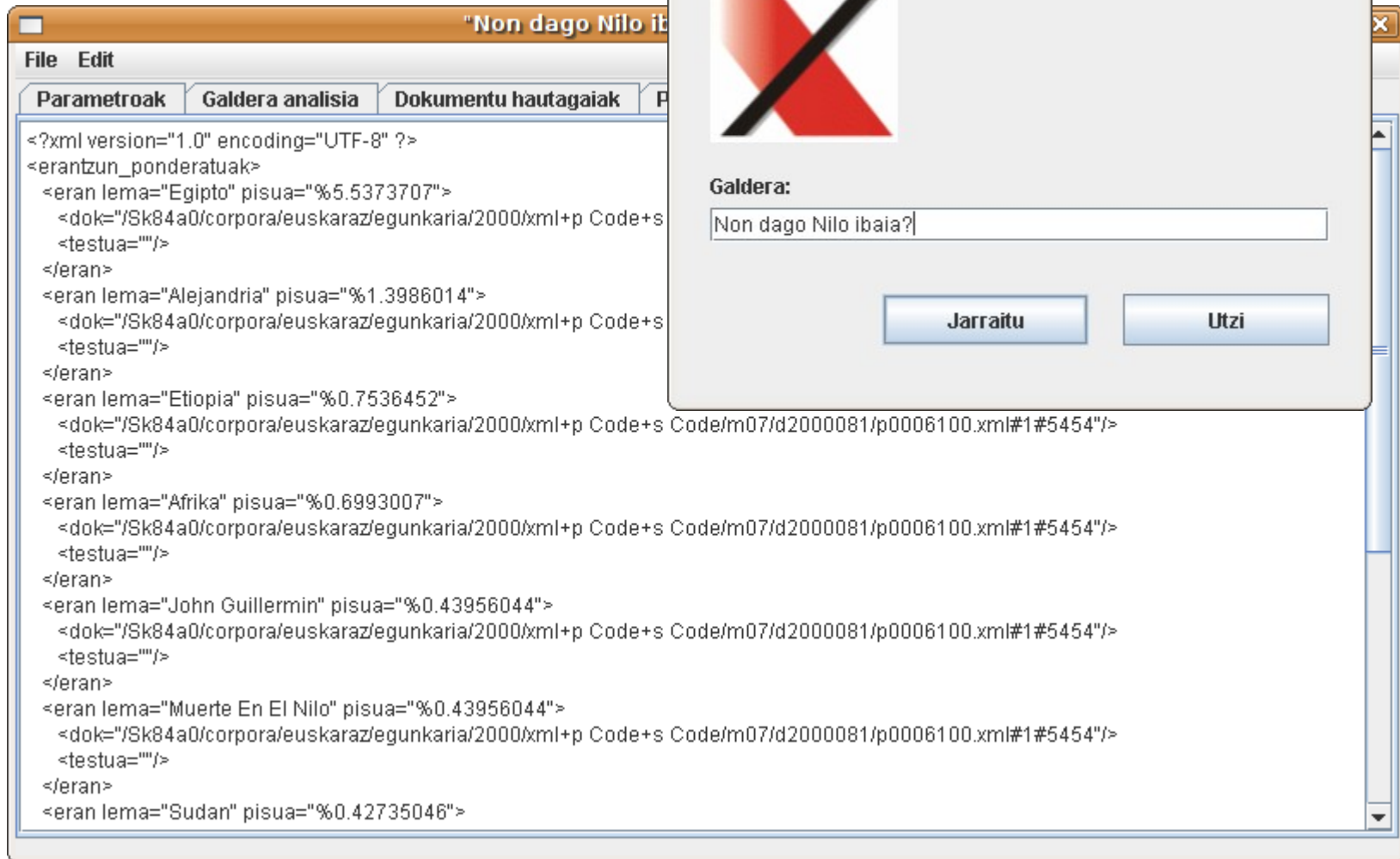
- Soinuan, irudietan eta bideotan bilaketak egitea
- Oinarrizko tresnak
  - hizketa-ezagutzaileak
    - doitasuna: %60tik gora
  - irudien sailkapen automatikoa
- Oinarrizko estrategia:
  - sailkapenak --> metadatuak
  - hizketa --> testua
  - elementuak: estekak
  - ohiko bilatzailea



# QA (galdera-erantzunak)

- Eraitza ez da dokumentua, erantzuna baizik
- Galdera motak
  - Faktoideak (nork, zer, non, noiz)
  - Definizioak
- IR + patroien erauzketa
  - galderen analisisa
  - pasarten berreskurapena ohiko IRz
  - erantzunaren bilaketa
- Arrakasta mugatua
  - *Ihardetsi* prototipoa

# Ihardetsi



The image shows two overlapping windows from the Ihardetsi application. The background window, titled "Non dago Nilo ibaia", displays XML data in a text editor. The XML contains several entries for different regions, each with a lemma, a pisua value, and a dok path. The foreground window, titled "IHARDETSI (en sisx04)", features the Ixa logo (a red 'X' with 'ixa' text) and the word "Ihardetsi". Below the logo, there is a "Galdera:" label, a text input field containing "Non dago Nilo ibaia?", and two buttons labeled "Jarraitu" and "Utzi".

```
<?xml version="1.0" encoding="UTF-8" ?>
<erantzun_ponderatuak>
  <eran lema="Egipto" pisua="%5.5373707">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s
    <testua=""/>
  </eran>
  <eran lema="Alejandria" pisua="%1.3986014">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s
    <testua=""/>
  </eran>
  <eran lema="Etiopia" pisua="%0.7536452">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s Code/m07/d2000081/p0006100.xml#1#5454"/>
    <testua=""/>
  </eran>
  <eran lema="Afrika" pisua="%0.6993007">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s Code/m07/d2000081/p0006100.xml#1#5454"/>
    <testua=""/>
  </eran>
  <eran lema="John Guillermin" pisua="%0.43956044">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s Code/m07/d2000081/p0006100.xml#1#5454"/>
    <testua=""/>
  </eran>
  <eran lema="Muerte En El Nilo" pisua="%0.43956044">
    <dok="/Sk84a0/corpora/euskaraz/egunkaria/2000/xml+p Code+s Code/m07/d2000081/p0006100.xml#1#5454"/>
    <testua=""/>
  </eran>
  <eran lema="Sudan" pisua="%0.42735046">
```

# IR semantikoa



- **Izenburua:** Yielding Ontologies for Transition-Based Organization
- **Helburuak:**
  - Ezagutza partekatzea hizkuntzen eta kulturen artean
  - Bilaketa semantiko sakona eta testu-erazketa intentsiboa
  - Wiki-ingurunea adituak haien ezagutza sartzeko informatikarien laguntzarik gabe
- **Iraupena:**
  - 2008ko martxoa – 2011ko martxoa
- **Ekimena:**
  - 30 pertsona-urte
  - Partaideak
    - Herbehereak, Italia, Alemania, Euskal Herria
    - Unibertsitateak, enpresak, administrazioa eta ingurumen-taldeak
    - Taiwan and Japonia bertatik finantziatuta

# Kyoto (ICT-211423)



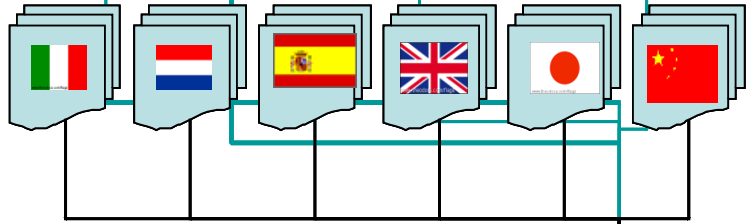
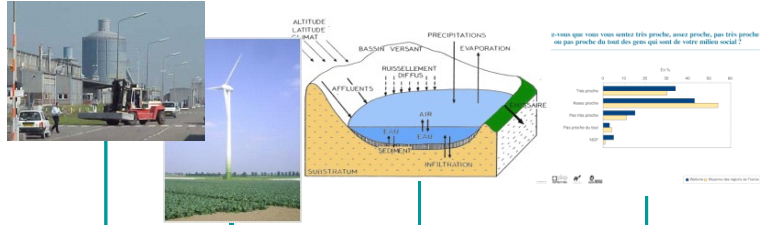
- **Hizkuntzak:**
  - English, Dutch, Italian, Spanish, Basque, Chinese, Japanese
- **Domeinua:**
  - Environmental domain, BUT usable in any domain
- **Esparrua:**
  - Both European and non-European languages
- **Eskuragarritasuna:**
  - Free: as open source system and data
- **Etorkizunerako pentsatua:**
  - Content standardization that supports world wide communication
  - Global Wordnet Grid



GREENPEACE



Environmental organizations

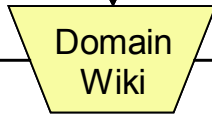


Governors Companies

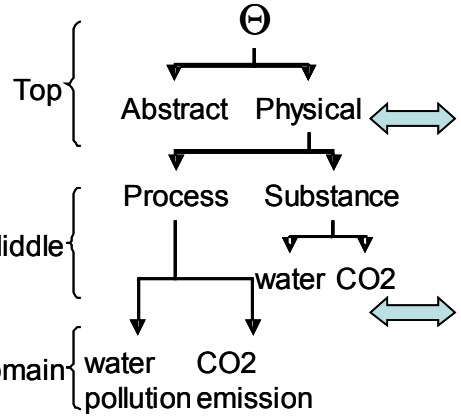
Environmental organizations



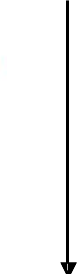
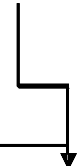
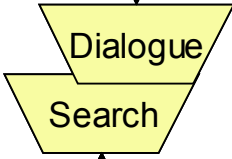
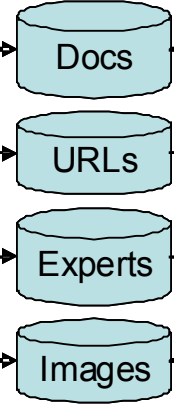
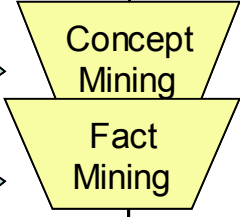
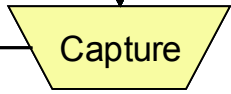
GREENPEACE



Universal Ontology



Wordnets



# Informazio-beharrak



- *What companies produce a lot of damaging substances?*
  - Zein enpresek sortzen dituzte substantzia kaltegarriak?
- *Air pollution by traffic*
  - Trafikoak eragindako aire poluzioa
- *Sick because of air pollution*
  - Aire poluzioarengatik gaixotutakoak
- *Cause of air pollution*
  - Aire-poluzioaren kausak
- *Air pollution from the Ruhr area*
  - Ruhr inguruko aire poluzioa