

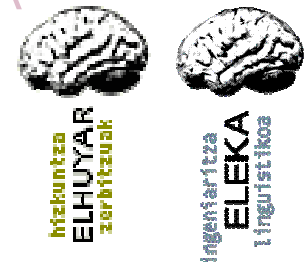


Eroski Consumer Corpusaren aurkezpena

Corpusgintza gaur egun

MINTEGIA

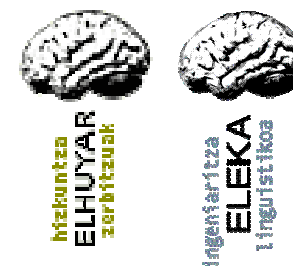
Igor Leturia - Eurne Martinez



2010eko urtarrilaren 21a



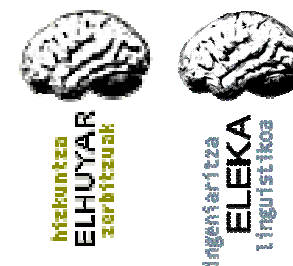
- Sarrera
- Corpusaren ezaugarriak
- Prozesua
 - Lerrokatzea
 - Analisia
 - Indexazioa
 - Kontsulta sistemaren garapena
- Kontsulta-sistema nola erabili
- Etorkizuna
- Estekak





Sarrera

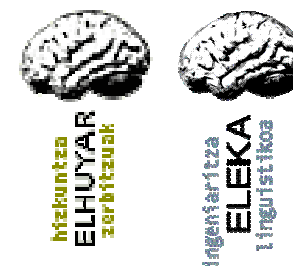
- Eroski Fundazioaren Consumer aldizkaria
 - Lau hizkuntzatan
 - Gaztelania (1998/01)
 - Euskara (1998/11)
 - Katalana (2000/01)
 - Galegoa (2000/04)
 - Kontsumo- eta gizarte-gaiak
 - Gaurkotasuna
- Hizkuntz erabileretarako interesgarria
- Eroski Consumer Corpusa
 - Elhuyar Fundazioa
 - Eleka Ingeniaritza Linguistikoa





Sarrera

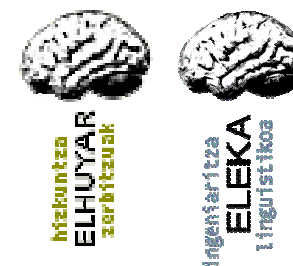
- Aurreko bertsioa
 - 2006/09 arte
 - Vigoko Unibertsitateko Seminario de Lingüística Informáticako webgunean
 - Beste corpus batzuekin batera
 - University of Southern Californiako Asier Alcazar-ek
- Berria
 - Eguneratuta
 - Consumeren webgunean
 - Beste bilaketa aukera eta emaitza mota batzuekin
 - TEI bertsioarekin ere





Corpusaren ezaugarriak

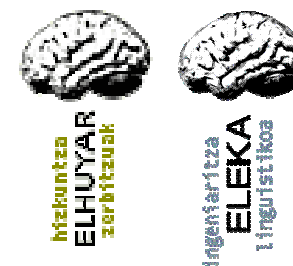
- Eleanitza, 4 hizkuntzatan
- Esaldi mailan lerrokatua, automatikoki
- Espezializatua, kontsumo-arlokoa
- Gaurkotasunekoa
 - Produktu berri asko
 - Teknologia berriak



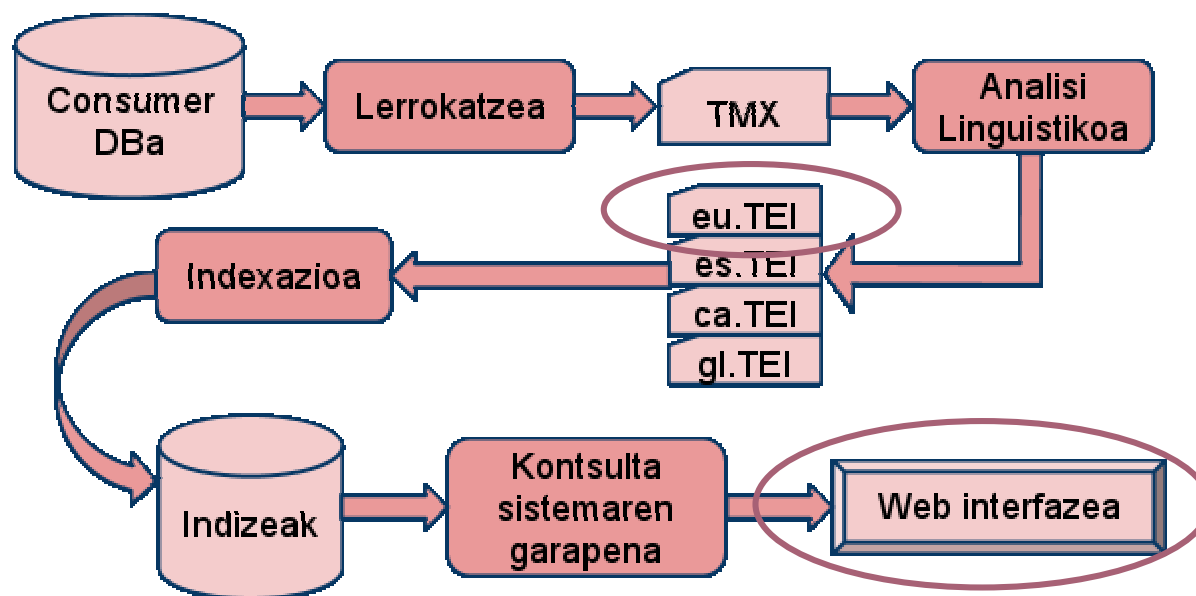


Corpusaren ezaugarriak

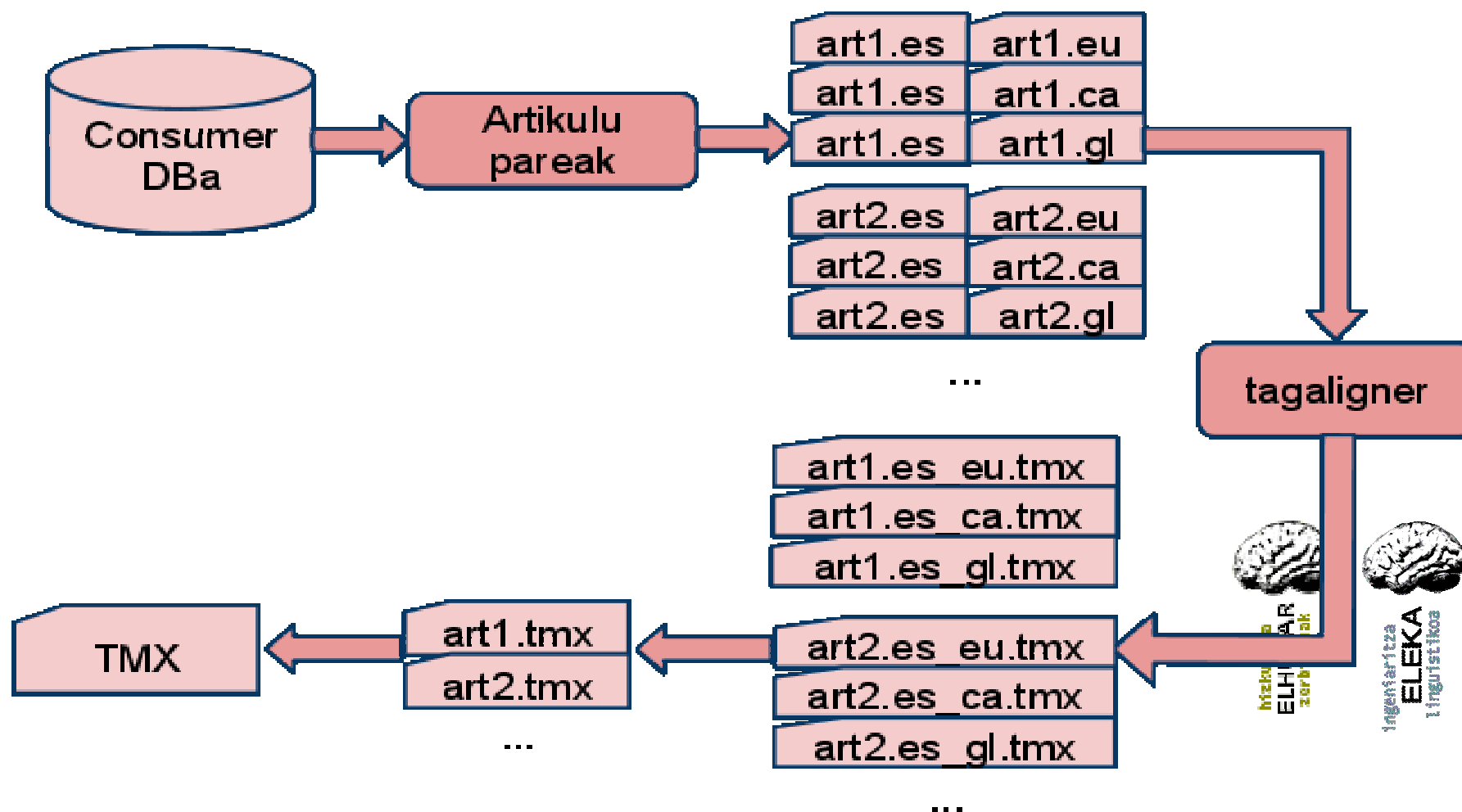
- Aldizkariaren 131 ale, 1998/01-2009/12 tartekoak
- 2.590 artikulu
- Tamaina
 - Euskaraz: 232.250 esaldi, 2.365.236 hitz
 - Gaztelaniaz: 292.274 esaldi, 3.758.454 hitz
 - Katalanez: 214.584 esaldi, 2.760.467 hitz
 - Galegoz: 208.652 esaldi, 2.548.978 hitz
- Lerrokatzearen zuzentasun-portzentajea:
 - Euskara eta beste hizkuntzen artean: %82 - %84
 - Beste hiruen artean: %89 - %93



Prozesua



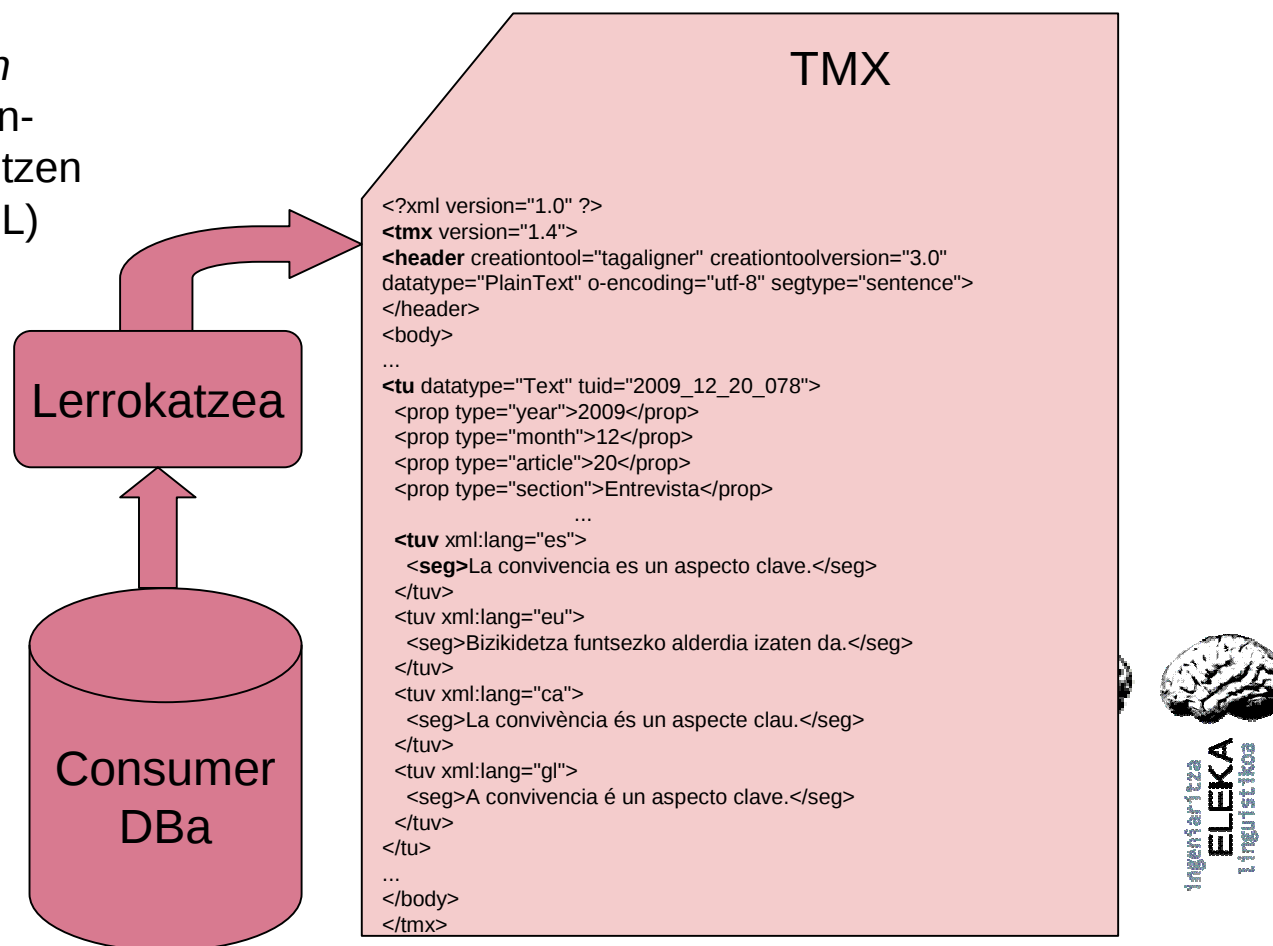
Prozesua: Lerrokatzea



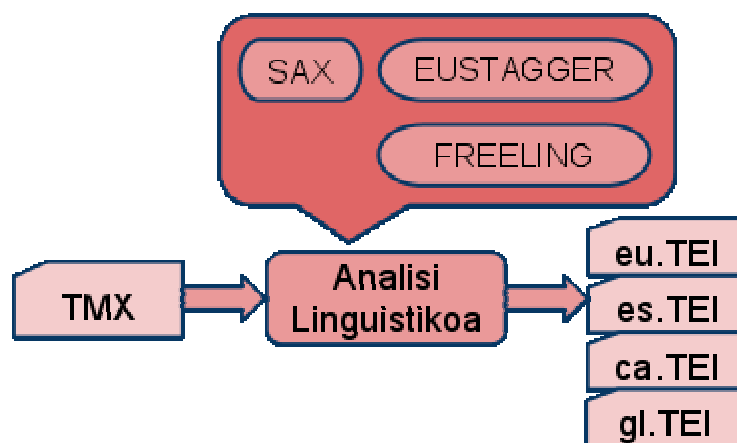


Prozesua: Lerrokatzea (TMX egitura)

TMX formatua (*Translation Memory eXchange*) itzulpen-memoriak trukatzeko erabiltzen den estandar irekia da (XML)



Prozesua: Analisia



- **SAX** parserra (XML)
- Analizatzaileak
 - Eletagger (eu)
 - Freeling (es,ca,gl)
- **TEI** formatua

Prozesua: Analisia (SAX)

SAX parserraren bidez TMX segmentuz-segmentu irakurri eta analizatu

TMX

```

<?xml version="1.0" ?>
<tmx version="1.4">
<header creationtool="tagaligner"
creationtoolversion="3.0"
datatype="PlainText" o-encoding="utf-8"
segtype="sentence">
</header>
<body>
...
<tu datatype="Text" tuid="2009_12_20_078">
<prop type="year">2009</prop>
<prop type="month">12</prop>
<prop type="article">20</prop>
<prop type="section">Entrevista</prop>
...
<tuv xml:lang="es">
<seg>La convivencia es un aspecto clave.</seg>
</tuv>
<tuv xml:lang="eu">
<seg>Bizikidetza funtsezko alderdia izaten da.</seg>
</tuv>
<tuv xml:lang="ca">
<seg>La convivència és un aspecte clau.</seg>
</tuv>
<tuv xml:lang="gl">
<seg>A convivencia é un aspecto clave.</seg>
</tuv>
</tu>
...
</body>
</tmx>
    
```

SAX

La convivencia es un aspecto clave

Bizikidetza funtsezko alderdia izaten da.

La convivència és un aspecte clau.

A convivencia é un aspecto clave.

es.TEI

eu.TEI

ca.TEI

gl.TEI

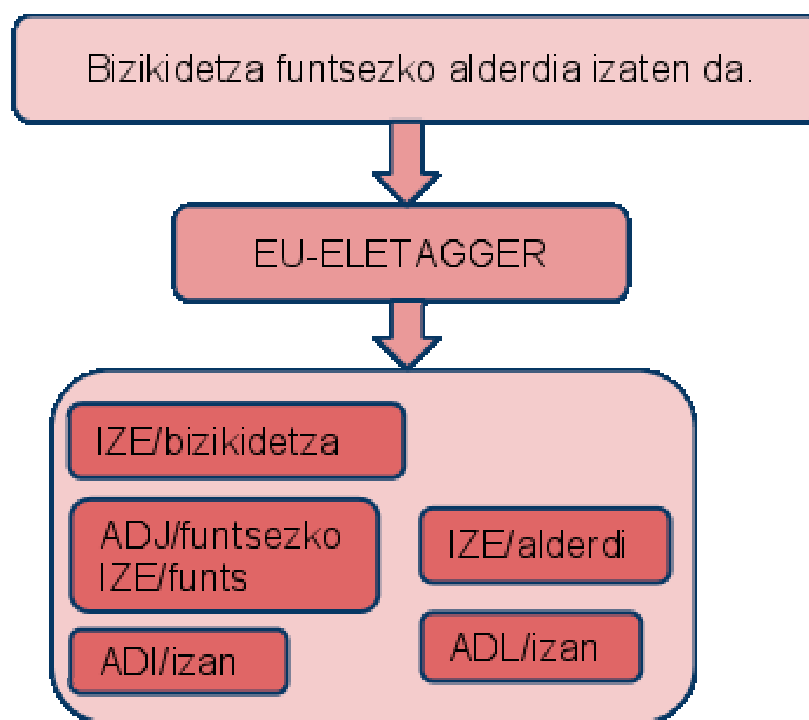


ELHUYAR
zerebitzuak

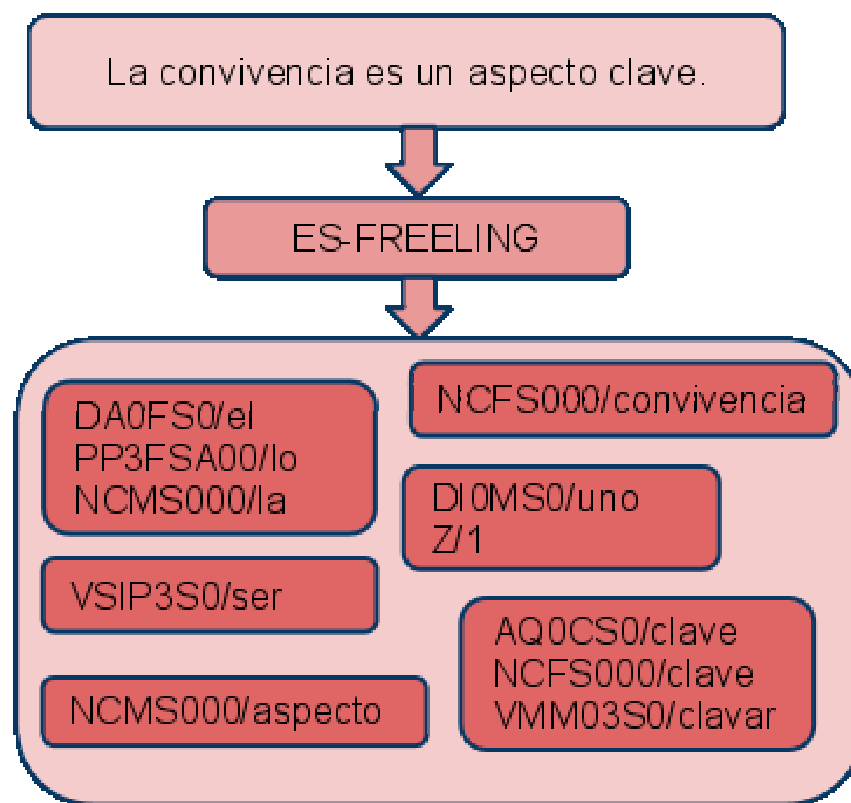
ELEKA
linguistikoa

Analisi linguistikoa

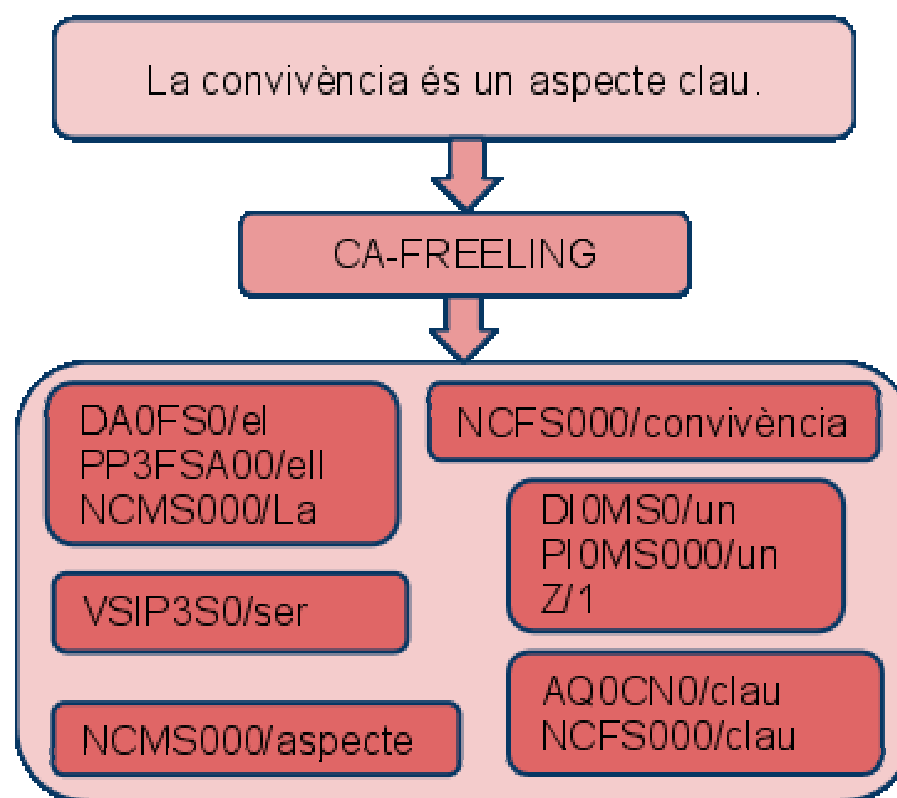
Prozesua: Analisia (EU)



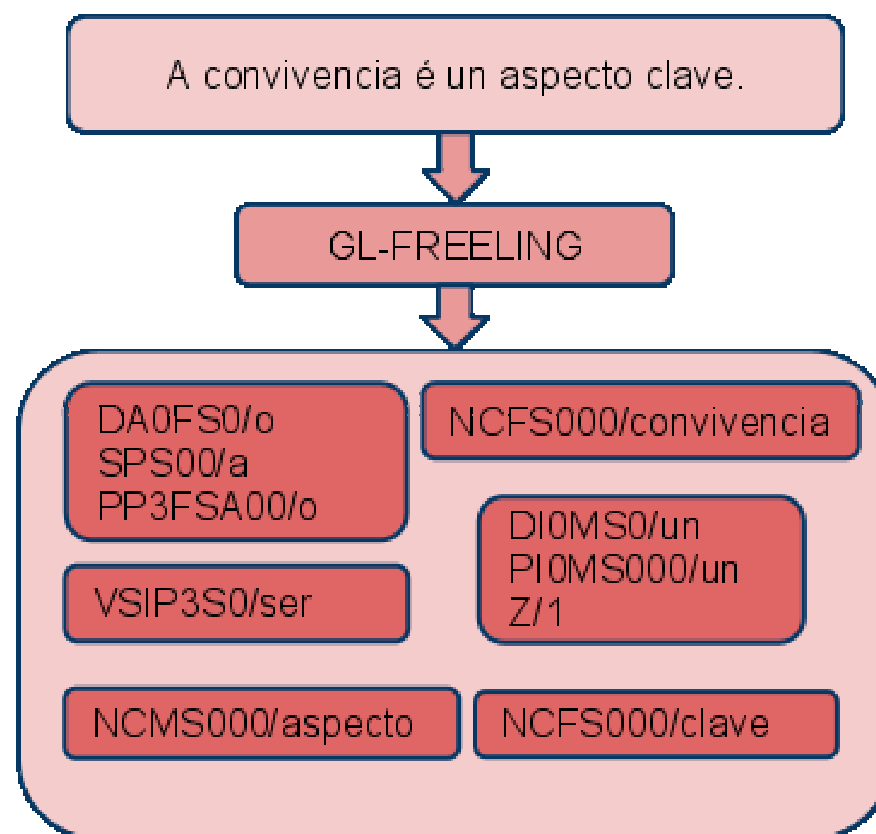
Prozesua: Analisia (ES)



Prozesua: Analisia (CA)

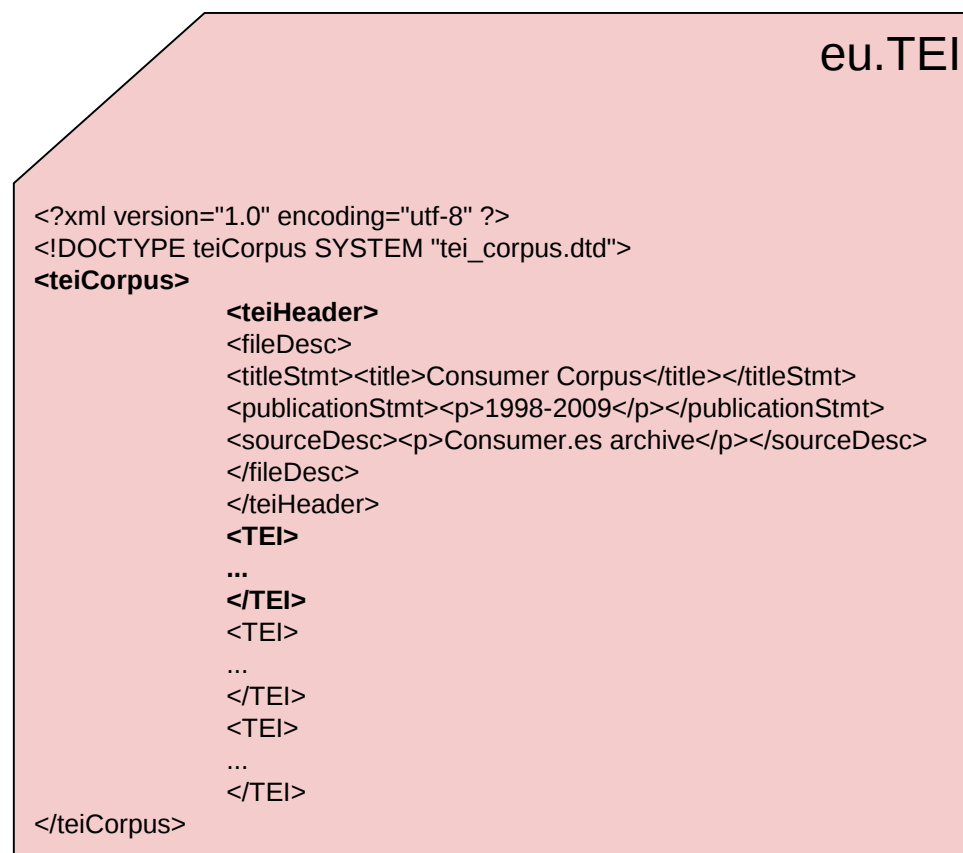


Prozesua: Analisia (GL)



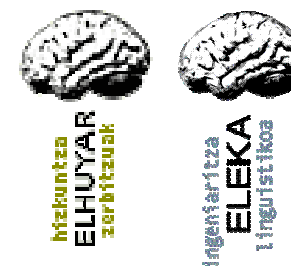


Prozesua: Analisia (TEI egitura)



TEI (*Text Encoding Initiative*)

Testuaren ezaugarriak
markatzeko XML
formatu estandarra da.

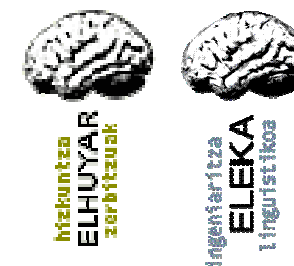




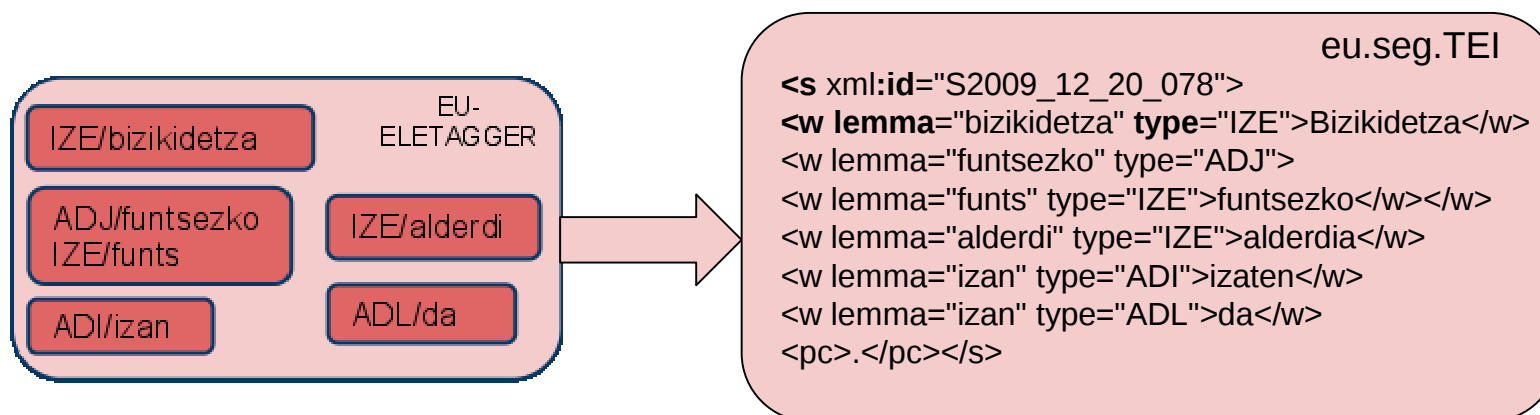
Prozesua: Analisia (TEI-artikulua)

eu.art.TEI

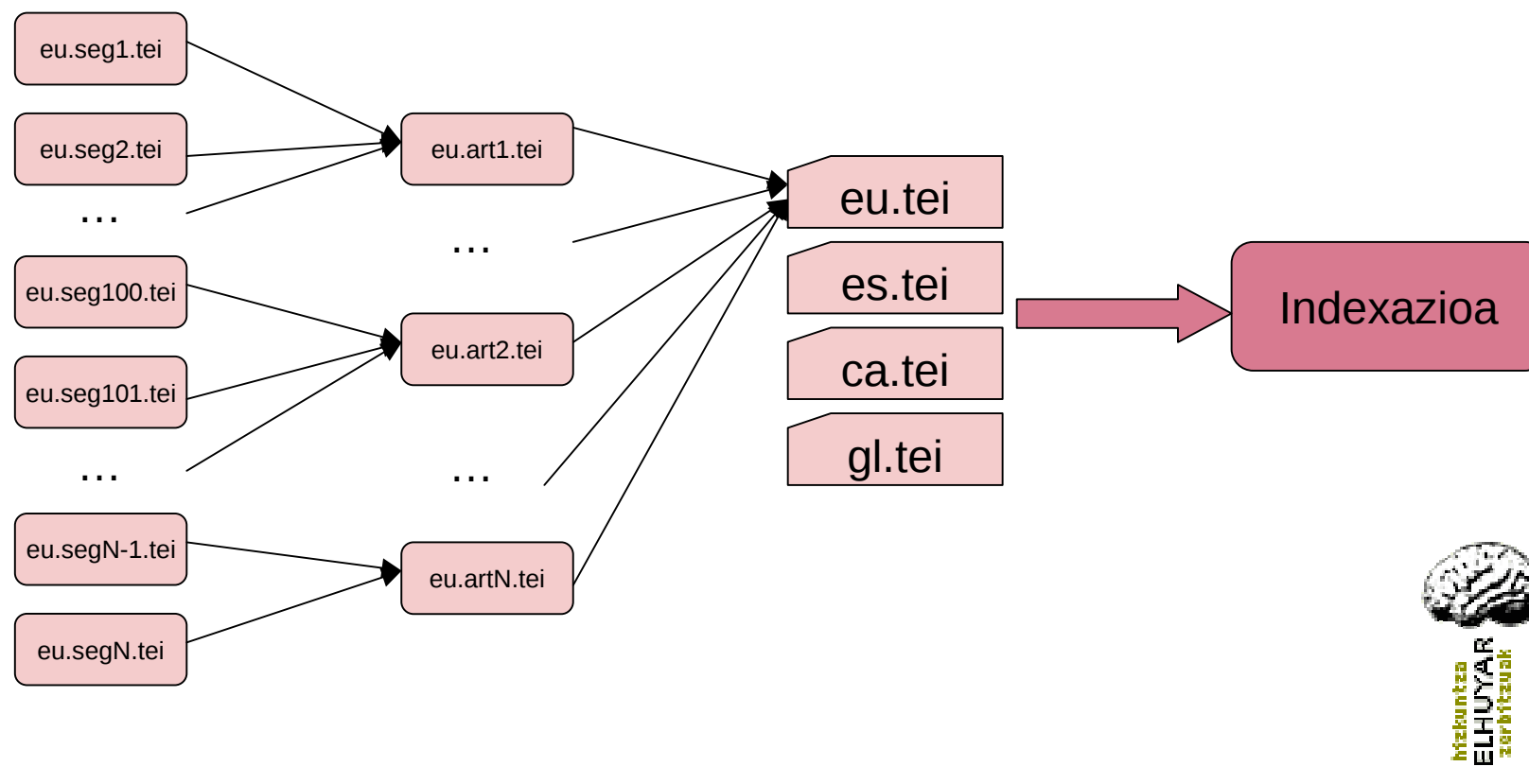
```
<TEI>
<teiHeader xml:lang="eu">
  <fileDesc>
    <titleStmt>
      <title level="a" type="Entrevista">
        &#34;Ongi zahartzeko, osasuna behar da, dirua eta maitasuna&#34;;
      </title>
    </titleStmt>
    <publicationStmt>
      <distributor>Eroski consumer.es archive</distributor>
      <availability>
        <p>Available from: consumer.es archive</p>
        <p>URL: http://revista.consumer.es/web/eu/20091201/entrevista/75315.php</p>
      </availability>
      <date when="2009">2009-12</date>
    </publicationStmt>
    <seriesStmt><title level="j">Consumer.es</title><idno type="alea">20</idno></seriesStmt>
    <sourceDesc>
      <p>Eroski consumer.es archive</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
<text><body><p>
<s xml:id="S2009_12_20_001">
...
<s xml:id="S2009_12_20_078">
...
</p></body></text></TEI>
```



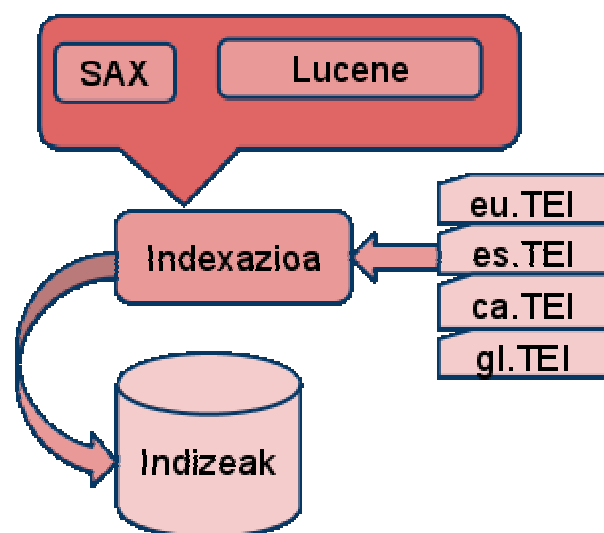
Prozesua: Analisia (TEI-segmentua)



Prozesua: Analisia (TEI-artikulua)



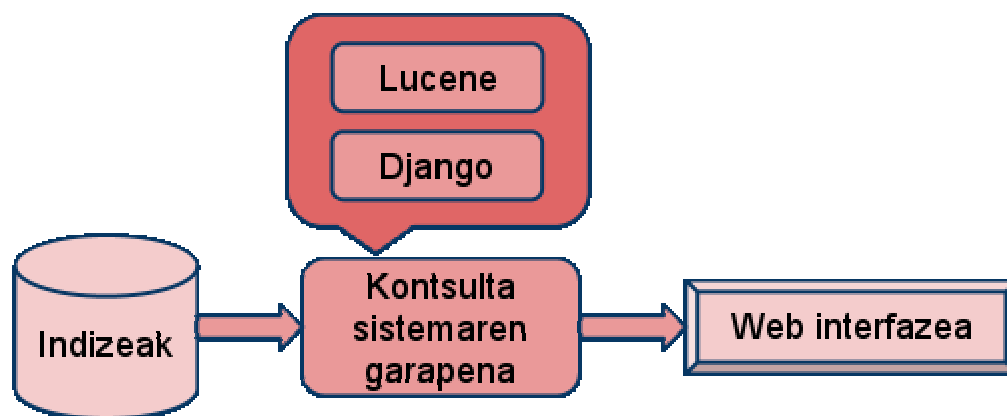
Prozesua: Indexazioa



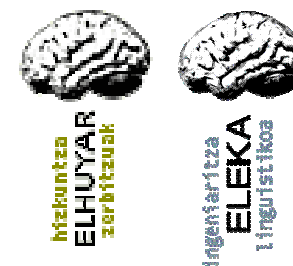
- **SAX** parserra (XML)
- **Lucene:** Testua indexatu eta bilatzeko liburutegia
 - ***Indexazio prozesuan, indizeak sortzeko erabiltzen da.***



Prozesua: Kontsulta sistemaren garapena



- **Django:** kode irekiko garapen frameworka
- **Lucene:** Testua indexatu eta bilatzeko liburutegia
 - ***Kontsulta garaian, indizeen gaineko bilaketak egiteko erabiltzen da.***





Kontsulta-sistema nola erabili

- Hementxe dago: <http://corpus.consumer.es>

Consumer Corpora

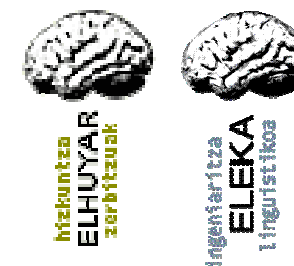
Aurkezpena | Laguntza

Galdera

Hizkuntza	Zer bilatu	Aukerak	Hitza	Kategoria	Emaiza
Gaztelania	Lema	Da	sistema		☒ Gaztelania ☒ Euskara ☒ Galegoa ☒ Katalana
Hizkuntza	Zer bilatu	Aukerak	Hitza	Kategoria	Grafiko aukerak
Gaztelania	Lema	Da			1. hitzaren forma

Emaitzak: artikulu, 2023 agerpen

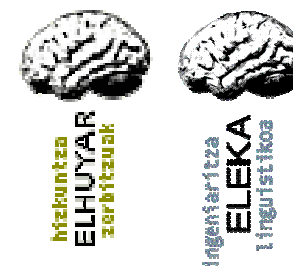
Gaztelania	Euskara	Galegoa	Katalana
2009-12: CONTENIDO ESTÁNDAR	2009-12: EDUKI ESTANDARRA	2009-12: CONTENIDO ESTÁNDAR	2009-12: CONTINGUT ESTÀNDAR
La videocámara menguante	Txikitzen, txikitzen...	As novas cámaras de video dixitais dispoñen de mellores condicións técnicas en dimensións cada vez menores	Les noves càmeres de video digitals disposen de millors condicions tècniques amb unes dimensions cada vegada més petites
Sin embargo, no es un sistema propicio para visualizar la grabación en un ordenador, ya que precisa de adaptadores especiales.	Sistema hori, ordea, ez da egokiena grabazioa ordenagailuan ikusteko, egokigailu bereziak behar izaten baitira.	Así e todo, non é un sistema propicio para visualizar a gravación nun ordenador, xa que precisa de adaptadores especiais.	Tot i això, no és un sistema propici per a visualitzar la gravació en un ordinador, ja que requereix uns adaptadors especials.
La CCD, sin embargo, contiene un sistema llamado 3 CCD que capta la imagen con mayor sensibilidad.	CCD motakoak, ordea, sistema berezi bat du, 3 CCD izenekoa, eta sentiberatasun gehiagorekin hartzen ditu irudiak.	A CCD, non obstante, contén un sistema chamado 3 CCD que capta a imaxe con maior sensibilidade.	La CCD, però, conté un sistema anomenat 3 CCD que capta la imatge amb més sensibilitat.
Se trata del sistema encargado de captar la imagen y transformarla en digital.	Bi sistema daude egun, eta alde handia dago bien artean, kalitatean bezala prezioan ere: zoom optikoa da bat eta zoom digitala bestea.	Trátase do sistema encargado de captar a imaxe e de transformala en dixital.	Es tracta del sistema encarregat de captar la imatge i de transformar-la en digital.
Así se podrá ponderar si merece la pena o no pagar un mayor precio.	Irudi grabatuaren kalitatean egon ohi da alderik handiena kameratik kamerara (zenbat eta garestiagoa, orduan eta kalitate hobea), eta irudiak gordetzeko sistemak ere badu zeresana, jakina.	Así se poderá ponderar se paga a pena ou non pagar un maior prezo.	Així es podrá ponderar si val la pena o no pagar un preu més alt.
2009-12: ENTREVISTA	2009-12: ELKARRIZKETA	2009-12: ENTREVISTA	2009-12: ENTREVISTA
"Para envejecer bien, se necesita salud, dinero y amor"	"Ongi zahartzeko, osasuna behar da, dirua eta maitasuna"	"Para avellentar ben, còmpre saúde, diñeiro e amor"	"Per a envellir bé, fan falta salut, diners i amor"





Etorkizuneko aukera interesgarriak

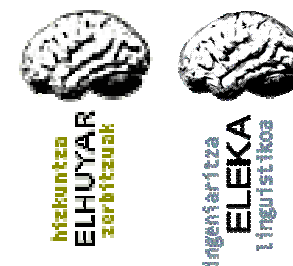
- Eguneratzen joatea aldizkariaren ale berriekin
- Hitzen mailako lerrokatze automatikoa neurri estatistikoen bidez
- Dataren arabera bilaketak mugatzea
- ...





Estekak

- CLUVI: <http://sli.uvigo.es/CLUVI/>
- Consumer aldizkaria: <http://consumer.es>
- Django: <http://www.djangoproject.com/>
- Lucene: <http://lucene.apache.org>
- pyLucene: <http://lucene.apache.org/pylucene/>
- Tagaligner: <http://tag-aligner.sourceforge.net/>
- TEI: <http://www.tei-c.org>
- ZT Corpora: <http://www.ztcorpusa.net>



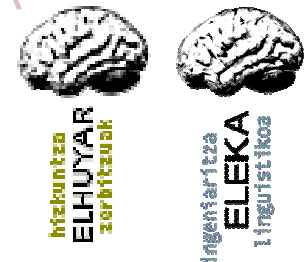


Eroski Consumer Corpusaren aurkezpena

Corpusgintza gaur egun

MINTEGIA

Igor Leturia - Eurne Martinez



2010eko urtarrilaren 21a