Universidad del País Vasco Euskal Herriko Unibertsitatea

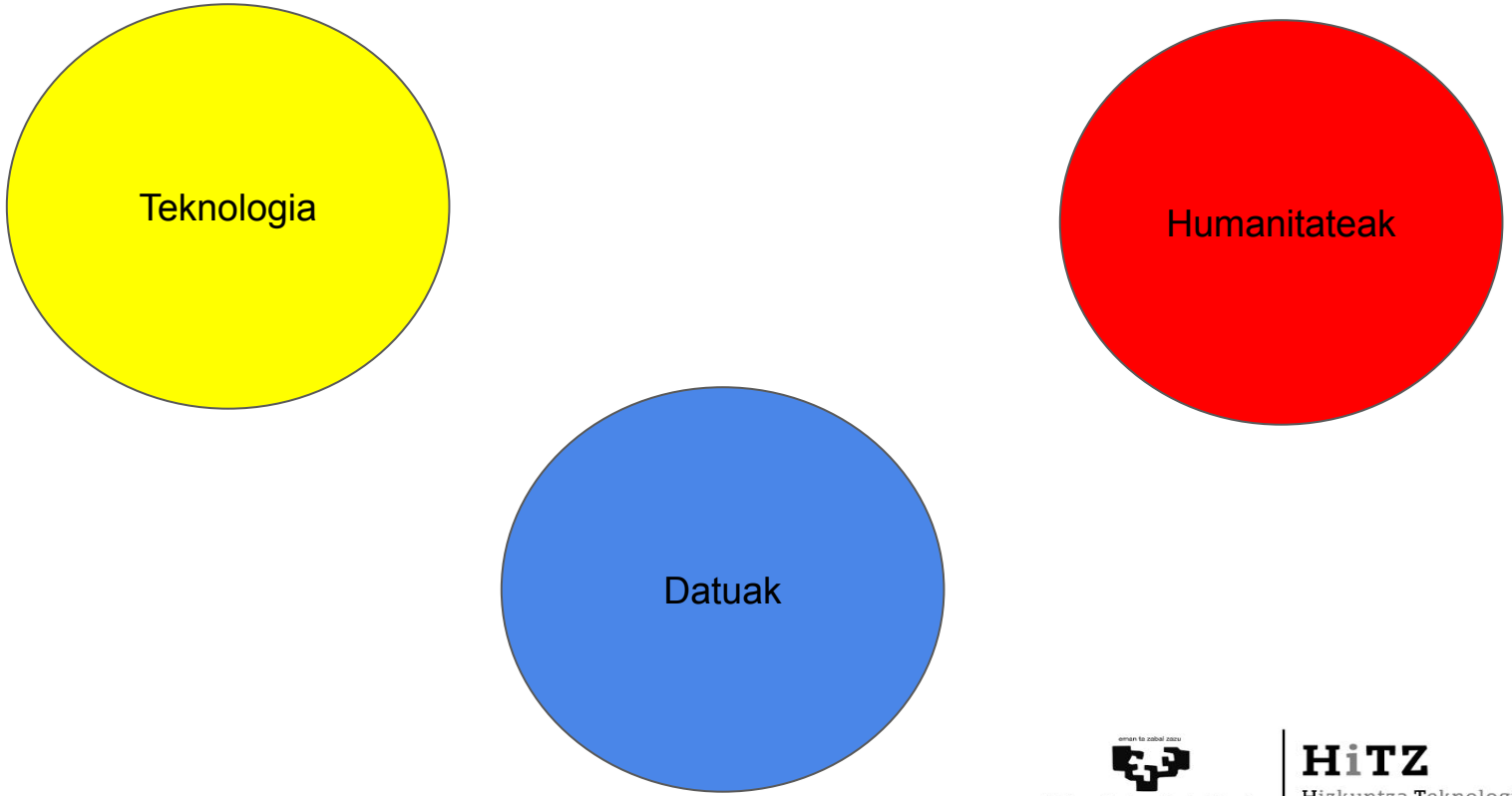
HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Humanitate Digitalak

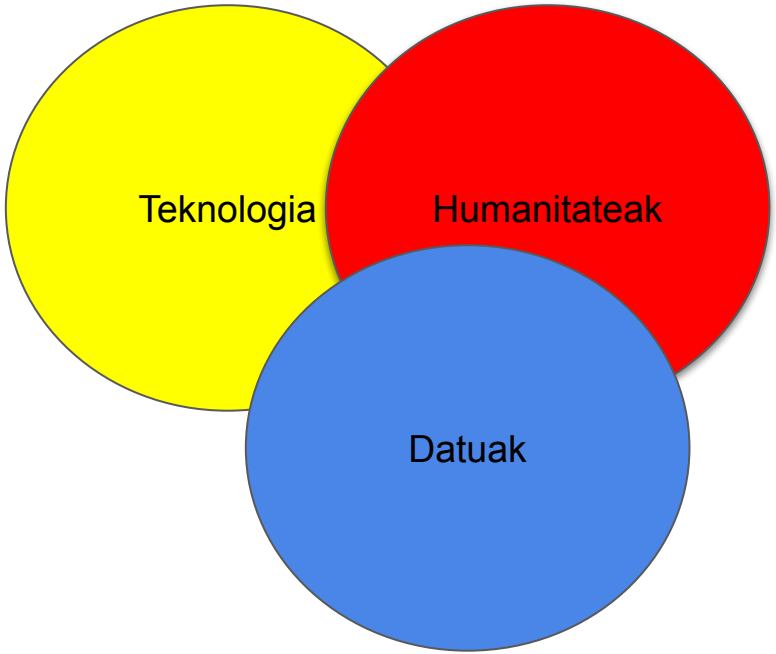
Sarrera, Digitalizazioaren Aukerak eta Erronkak

Rodrigo Agerri
IXA taldea, HiTZ zentroa
http://www.rodrigoagerri.net/

Zer dira humanitate digitalak?

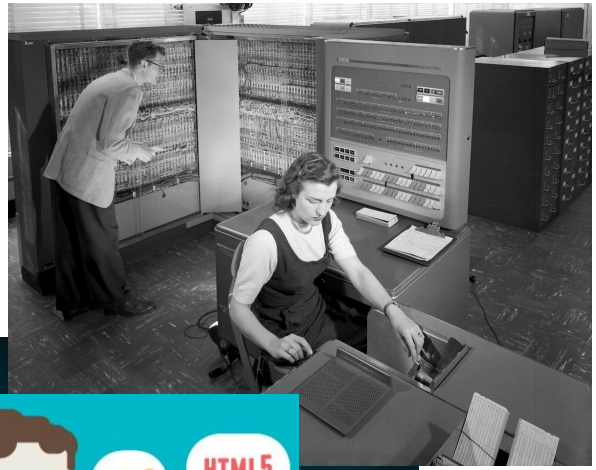


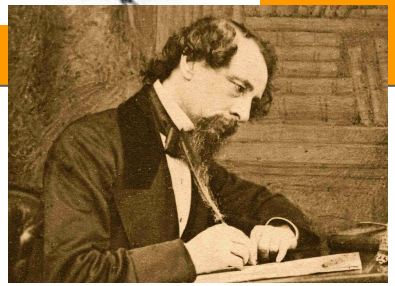
Zer dira humanitate digitalak?



Zer dira humanitate digitalak?

ki konput
teak oso





Zer dira humanitate digitalak?

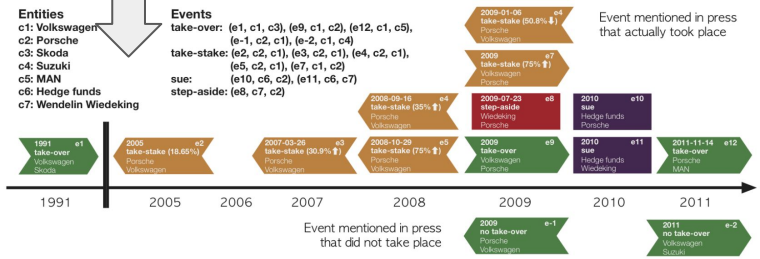
- Dokumentuen digitalizazio masiboarekin, sare sozialen eztandarekin eta hizkuntz teknologien hedapenarekin arloen arteko lankidetza handiagoa
- Humanitate digitalak eta Gizarte-Zientzia digitalak jakintza-arlo emergenteak
- Ikerkuntzan batez ere baina gero eta aplikazio “arrunt” gehiagotan
- Historia, literatura, hezkuntza, hizkuntzalaritza, kazetaritza, soziologia, soziolinguistika, arkeologia eta beste jakintza arlo askotan aldaketa sakonak

Teknologiak

- Digitalizazioa eta OCR
- **Hizkuntza-teknologiak (Hizkuntzaren prozesamendua)**
  - **Testu / Hizketa**
- Dokumentaziorako metadatuak
- Bisualizazio-tresnak
- Big data / hodeia

Teknologiak


```
<NAF xml:lang="eu" version="v1.naf">
  <nafHeader>
    <linguisticProcessors layer="text">
      <lp name="ixa-pipe-tok-eu" beginTimestamp="2016-04-25T11:25:29+0200" en
mestamp="2016-04-25T11:25:29+0200" version="2.0.0-0512ff8407b44bd9a9fbe9b27ca
520758b03f" hostname="mariturri" />
    </linguisticProcessors>
    <linguisticProcessors layer="terms">
      <lp name="ixa-pipe-pos-eu-pos-perceptron" beginTimestamp="2016-04-25T11
:34+0200" endTimestamp="2016-04-25T11:25:34+0200" version="2.0.0-2837752920a3
5cb0b501f7163aef2e2f7ba5f" hostname="mariturri" />
    </linguisticProcessors>
  </nafHeader>
  <text>
```



Analisi kuantitatiboak

- Iturrietan eskuz bilatu behar denean analisi kuantitatiboak egitea (ia) ezinezkoa da
- Aurrerapen handia baina mugatua
- Datu gordinak (askotan kantitate handiegiak ez bada automatikoki sintesi bat egiten)
- Hitz mailako kontaketak (ez lema, are gutxiago kontzeptua)
- Hizkuntza teknologiak muga horiek gainditzen saiatzen dira
- Beti ere errore-tasa batekin

Arazoak

- Digitalizaziorik ez (askotan eskuz daude idatzita)
- OCR egin gabe: Argazkin hutsekin urrunera ez
- OCR eginda baina akatsekin (askotan)
- Hizkuntza ez-estandarra: Testu zaharrak, ez-formalak
    - Normalizazioa eta errore-tasa handiagoa
- Lizentziak: Publikoa eta irekia?

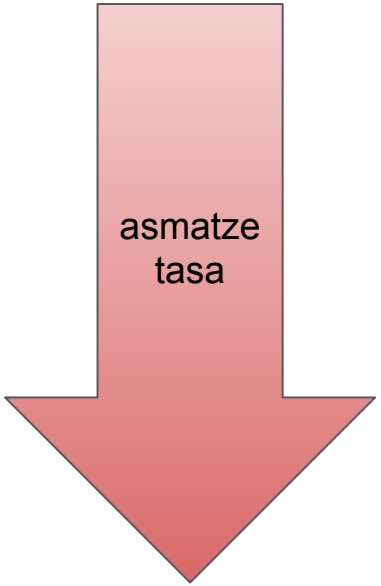
Kritikak eta Mugak

- Lehentasuna lehiatzea da (pribatuen interesak ikerkuntzan)
- Datuak enpresa erraldoien esku egotea eta ez denon eskura
- Emaitzen interpretagarritasun falta (teknologia kutxa beltza izan daiteke)
- Moda izatea

  https://en.wikipedia.org/wiki/Digital_humanities#Criticism

Hizkuntzaren Prozesamendua

- POS tagging (analisi morfosintaktikoa)
- Entitateen Identifikazioa eta Desanbiguazioa
- Entitateen arteko erlazioak
- Iritzien Azterketa
- Rol semantikoak eta analisi sintaktikoa
- Gaien Idenfikazioa
- Denbora-lerroak



Hizkuntza ez-estandarraren normalizazioa

Hurbilpenak

- Datuetan oinarrituta
  - Ikasketa automatikoan oinarrituta
  - Emaitza onak datu nahikoak eta zuzenak daudenean
  - Askotan lan handia datu-prestakuntzan
- Ezagutzan oinarritutako sistemak
  - Lehen hurbilpenerako egokia baina sistema errealak egiteko zailtasunak
  - Egokia ikasteko datuak ez daudenean

Datuetan oinarrituta

```
Lurralde          lurralde          IZE_ARR
historikoek       historiko         ADJ_ARR_ERG
babes   babes     IZE_ARR
zibilaren         zibil     ADJ_ARR_GEN
zabalkundean      zabalkunde        IZE_ARR_INE
parte   parte     IZE_ARR
hartuko hartu     ADI_SIN
```

IXAk garatutako teknologia

- https://ixa.eus/produktuak?language=eu
- http://ixa2.si.ehu.es/clarink/index.php?lang=eu

Tokenizazioa

```
<wf id="w69" sent="4" para="4" offset="354"
        length="9">announced</wf>
```

- Normalizazioa: Corpusen arabera.
- Paragrafoak, esaldiak, hitzak eta puntuazio-markak...
- Erregeletan oinarrituta.

POS tagging

Jeroen Dijsselbloem Eurotaldeko presidentearekin egin du bilera Varufakisek...

```
<!--Varufakisek-->
<term id="t64" type="close" lemma="varufakis" pos="R" morphofeat="PROPN">
  <span>
    <target id="w64" />
  </span>
</term>
```

Entitateak

Jeroen Dijsselbloem Eurotaldeko presidentearekin egin du bilera Varufakisek...

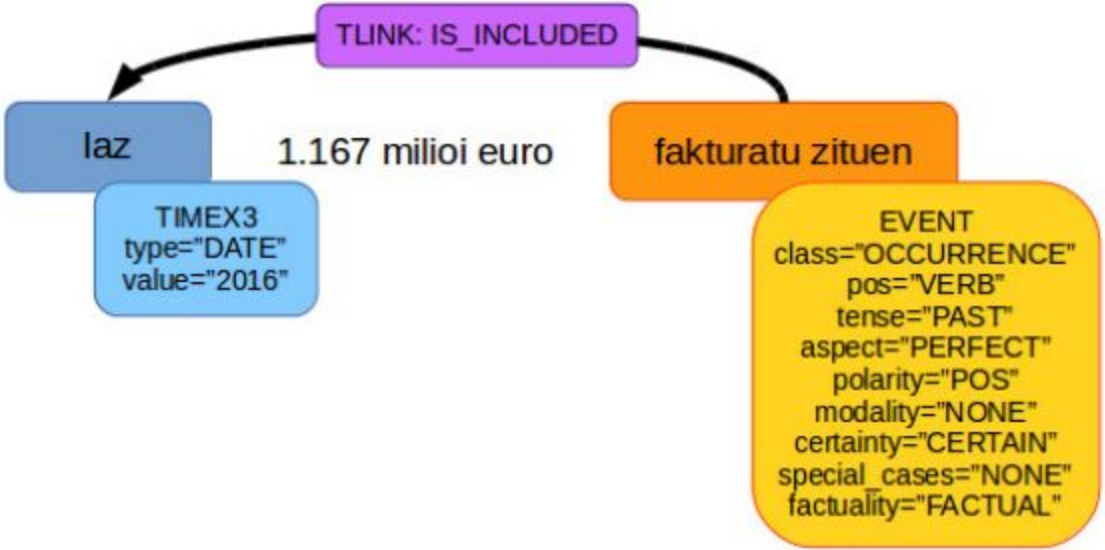
```
<entity id="e7" type="PER">
  <references>
    <!--Varufakisek-->
    <span>
      <target id="t109" />
    </span>
  </references>
</entity>
```

Iritzien Azterketa

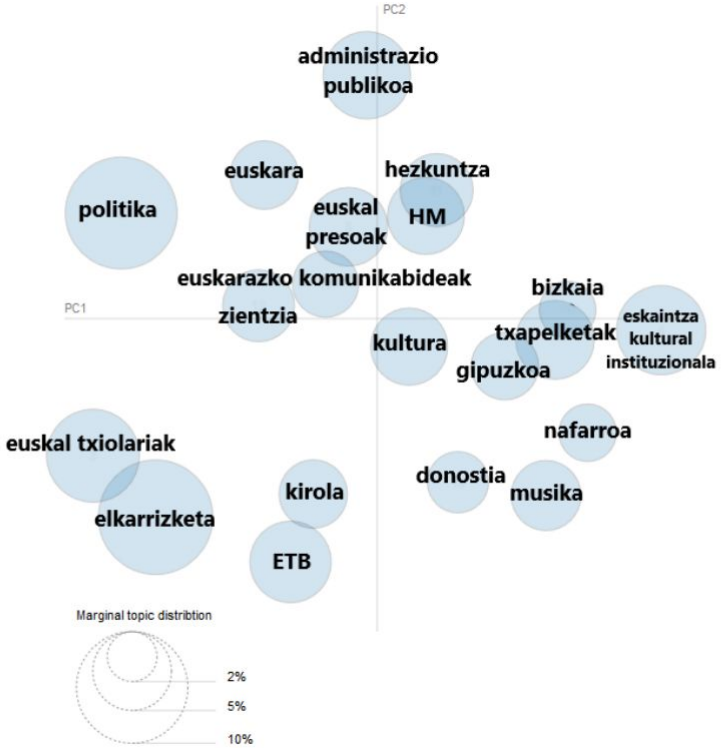
Jabeak ere maitagarriak dira.

```
<opinion oid="o0">
  <!-- Tag opinion scope id=8 -->
  <opinion_holder/>
  <opinion_target>
    <!-- Jabeak -->
    <span>
      <target id="t10"/>
    </span>
  </opinion_target>
  <opinion_expression polarity="Positive">
    <!-- maitagarriak -->
    <span>
      <target id="t12"/>
    </span>
  </opinion_expression>
</opinion>
```

Rol semantikoak



Gaien Identifikazioa



LDAvis: [20 topiko]

Gaien Identifikazioa

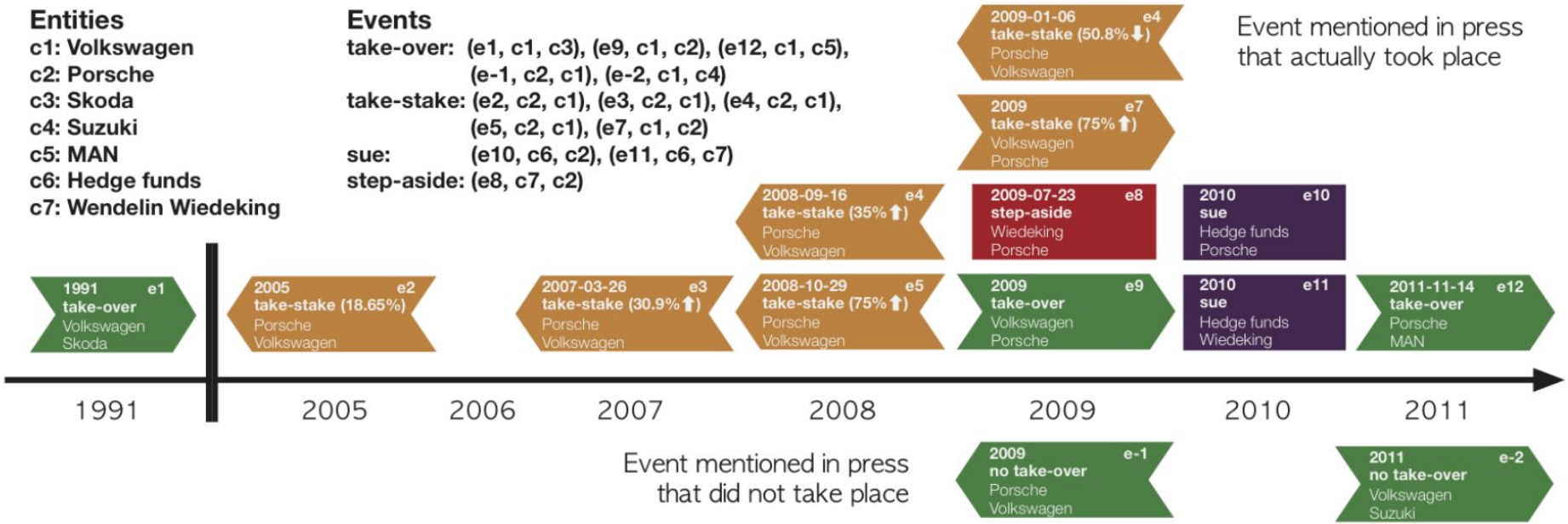
(a) Kirolak (% 21,61) (b) Euskara (% 20,70) (c) Ezker Abertzalea (% 17,12)

(d) Albisteak (% 14,92) (e) Musika (% 11,35)

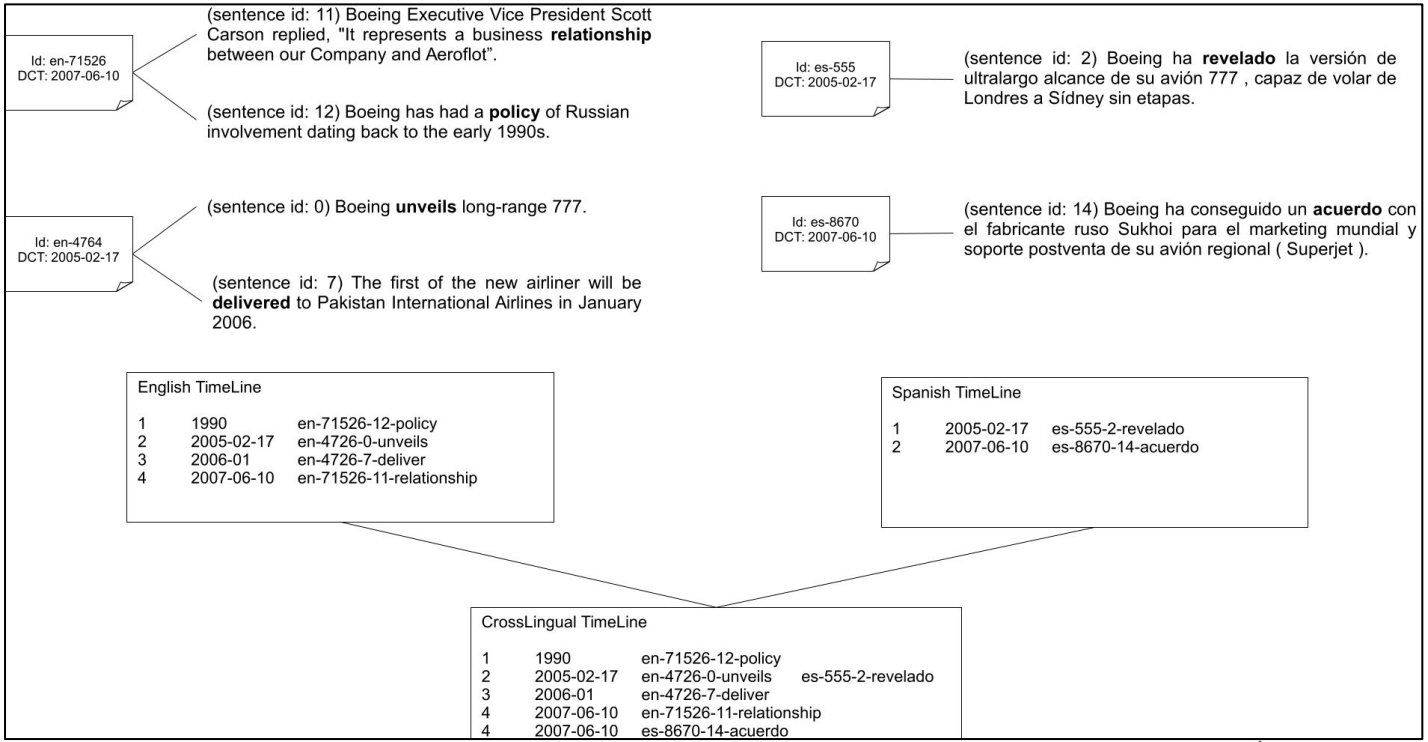
ixa

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

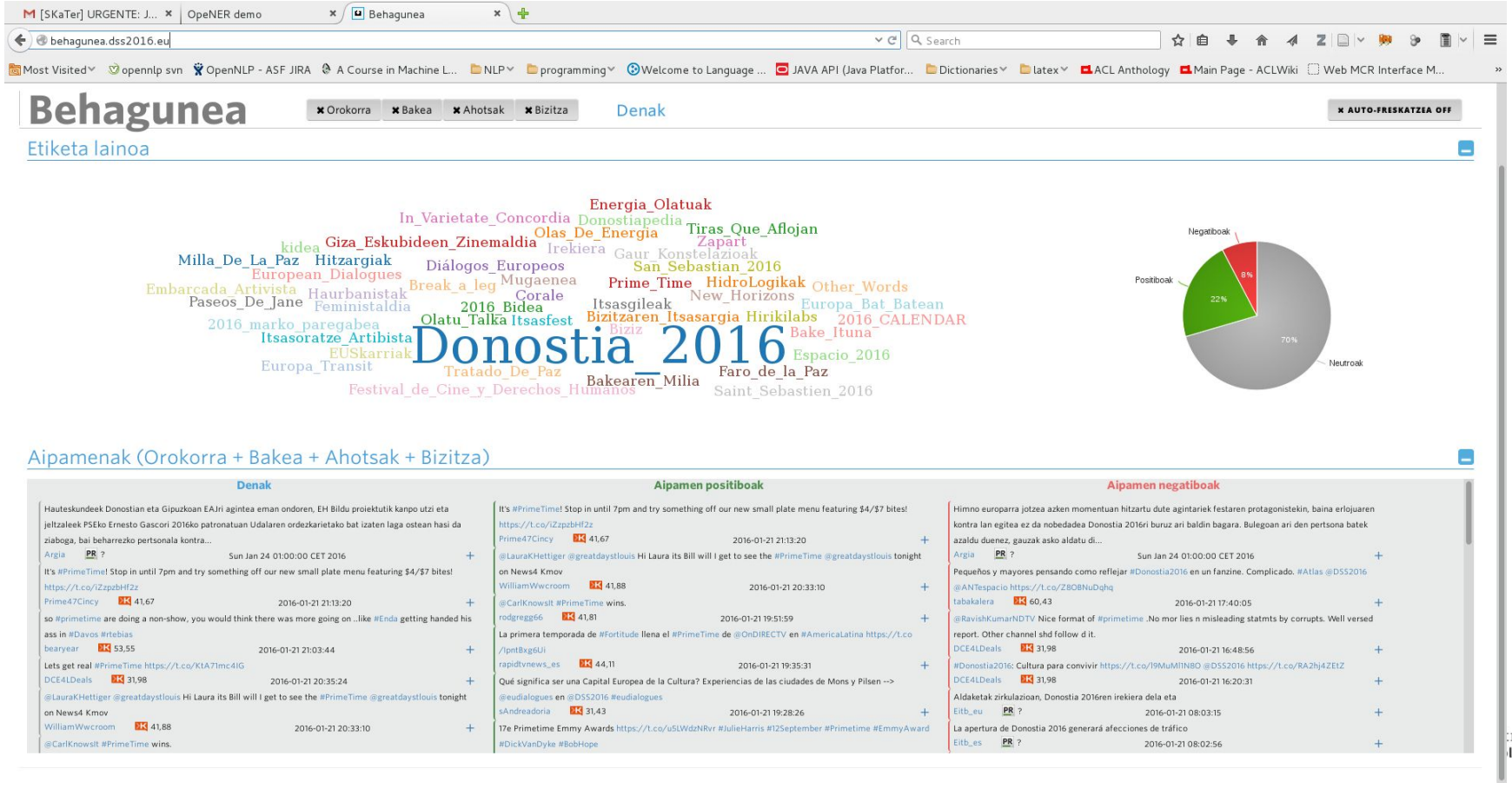
Denbora-lerroak



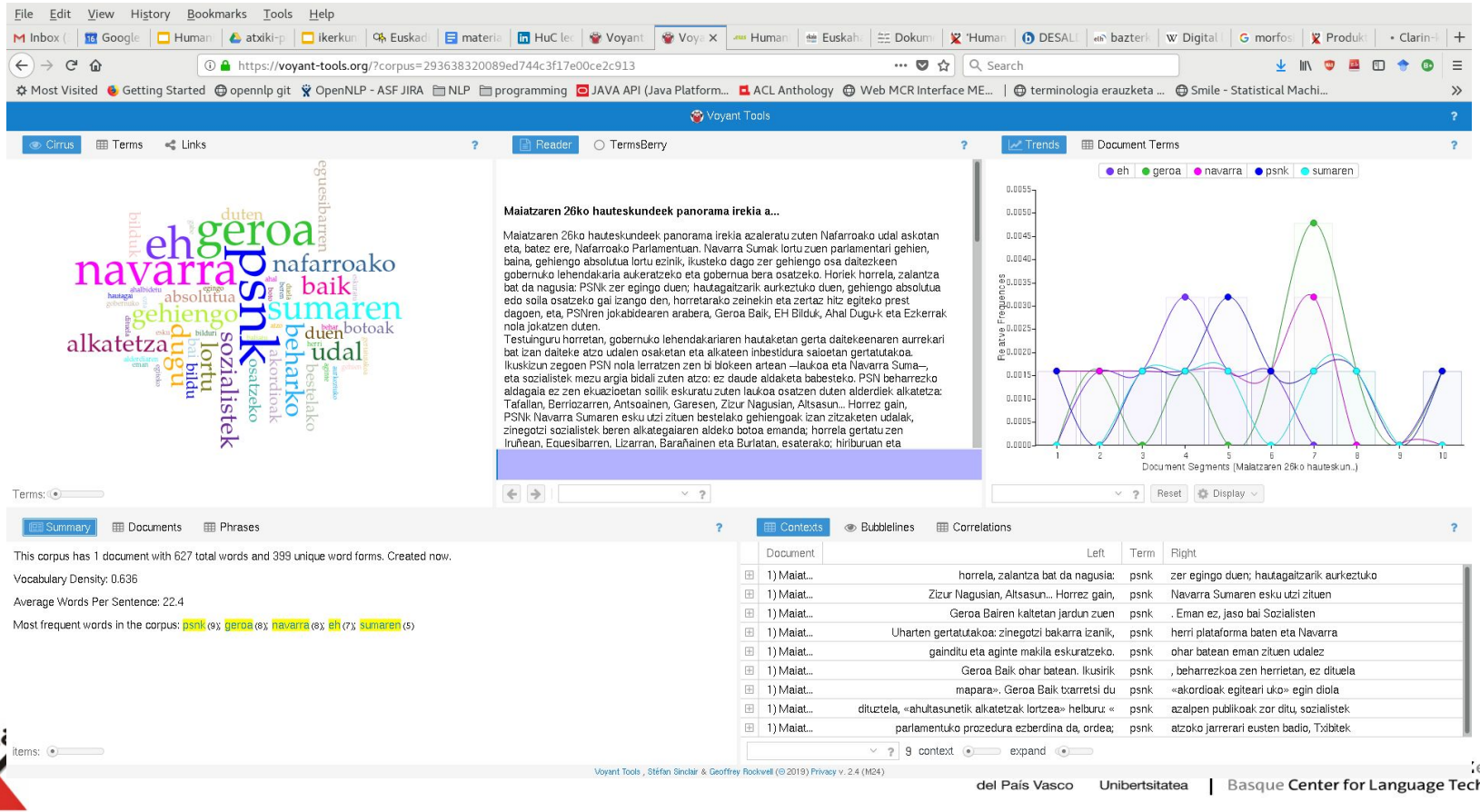
Denbora-lerroak (eleanitzak)



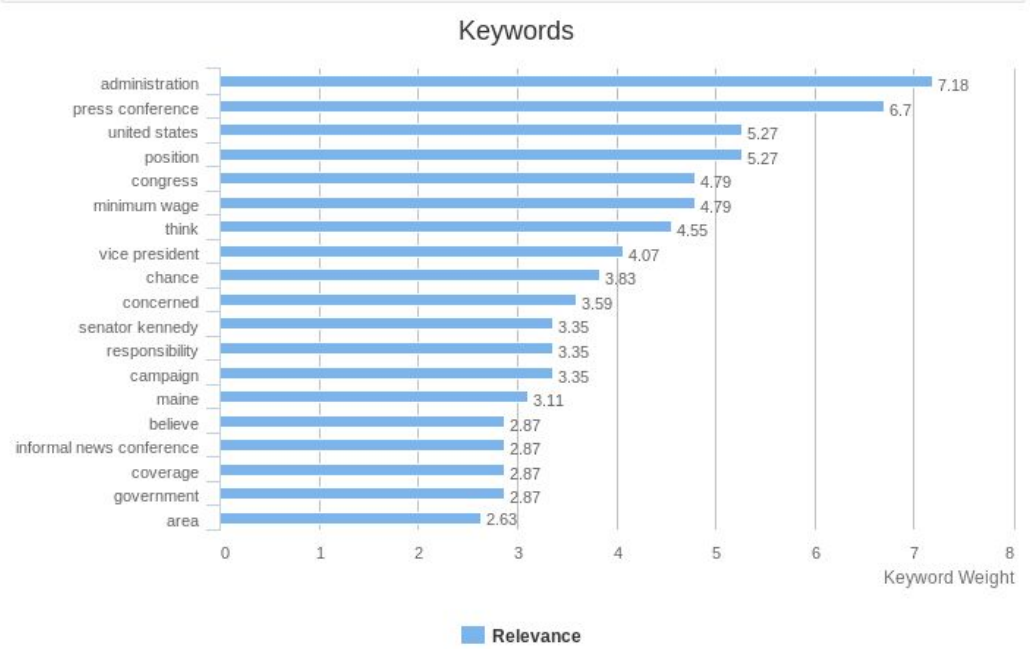
Bisualizazioa: http://behagune.elhuyar.eus/

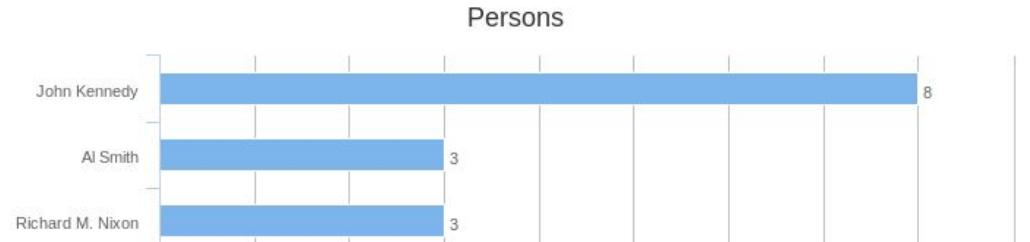


Bisualizazioa: https://voyant-tools.org/



Bisualizazioa: http://celct.fbk.eu:8080/Alcide_Demo/login



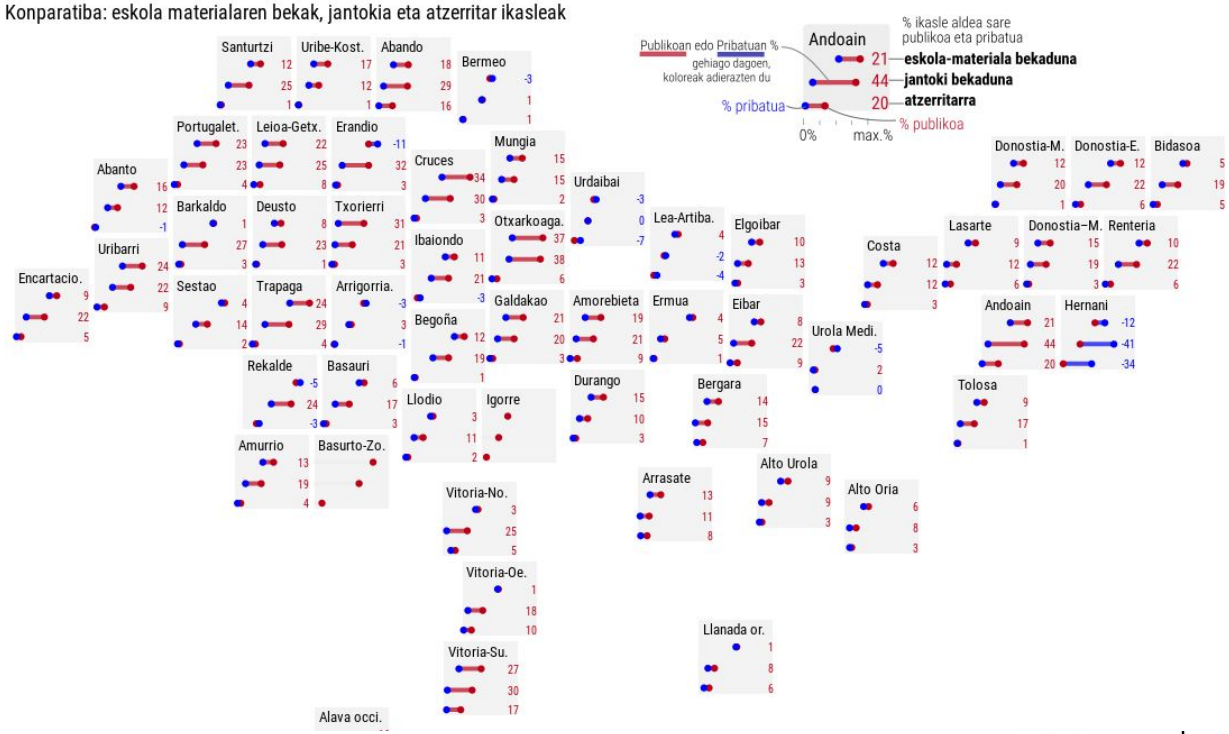


ixa

Jniversidad País Vasco · Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Bisualizazioa: https://lab.montera34.com/segregacionescolar/eus.html



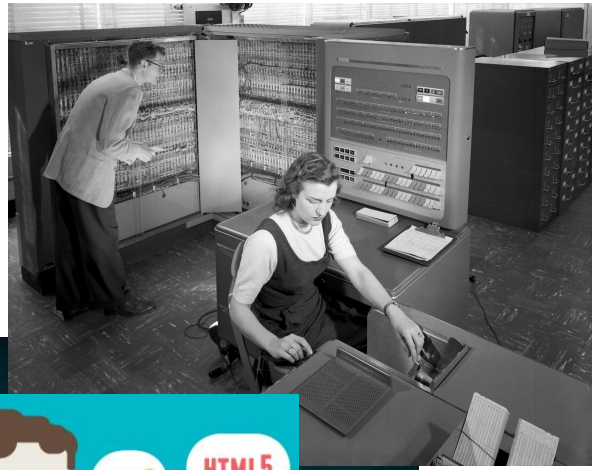
Oso sinple izan daiteke

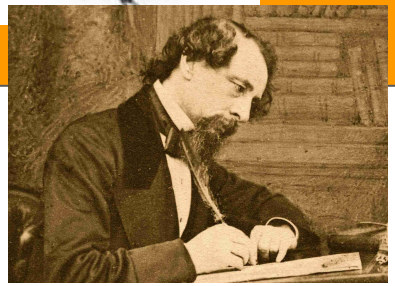
```
ragerri@mariturri:~/javacode/examples$ cat nafarroa-hauteskundeak-2019.txt | java -jar ~/javacode/ix
a-pipe-tok/target/ixa-pipe-tok-2.0.0-exec.jar tok -l eu -o conll |  sed '/^[[:punct:]]/d' | sort | u
niq -c | sort -nr | less
```

```
9 PSNk
9 ez
8 Navarra
8 Geroa
8 bat
7 zen
7 EH
7 du
6 ere
5 zuen
5 Sumaren
5 izan
5 ditu
5 da
4 zer
4 udal
4 sozialistek
4 Nafarroako
4 lortu
4 gehiengo
4 beharko
4 Baik
4 alkatetza
3 zuten
3 osatzeko
```

Erronkak

ki konput
teak oso

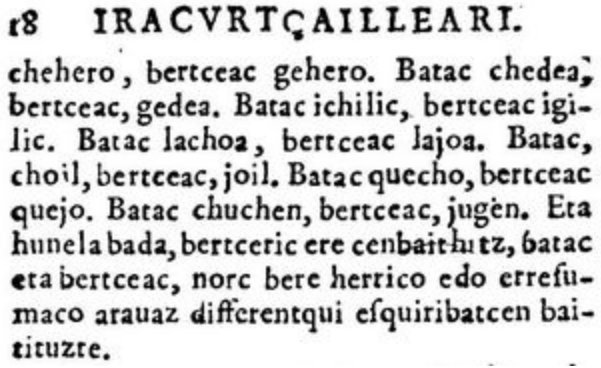




Eskerrik asko!

http://ixa2.si.ehu.es/testuhistorikoak/s/lemma/hitz

IRAKURTZAILLEARI chehero, bertzeak gehero. Batak chedea, bertzeak gedea. Batak ichilik, bertzeak igilik. Batak lachoa, bertzeak lajoa. Batak choil, bertzeak joil. Batak kecho, bertzeak kejo. Batak chuchen, bertzeak jugen. Eta hunela bada bertzerik ere zenbait hitz, batak eta bertzeak, nork bere herriko edo erresumako arauaz diferentki eskiribatzen baitituzte. Ordea zeren ezpaitira hamar bat hitz edo baizen, hunela diferentki, eta bi aldetara eskiribatzen direnak: halatan nik ere zenbait aldiz eskiribatukoitut alde batera liburuan barrena, eta bertze aldera liburuaren bazterrean, in margine : bat bederak zerbait


18 IRACVRTÇAILLEARI.

chehero, bertceac gehero. Batac chedea, bertceac, gedea. Batac ichilic, bertceac igilic. Batac lachoa, bertceac lajoa. Batac, choil, berteeac, joil. Batac quecho, bertceac quejo. Batac chuchen, bertceac, jugèn. Eta hunela bada, bertceric ere cenbait hitz, batac eta bertceac, norc bere herrico edo erresumaco arauaz differentqui esquiribatcen baituzte.