



HiTZ

Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Humanitate Digitalak

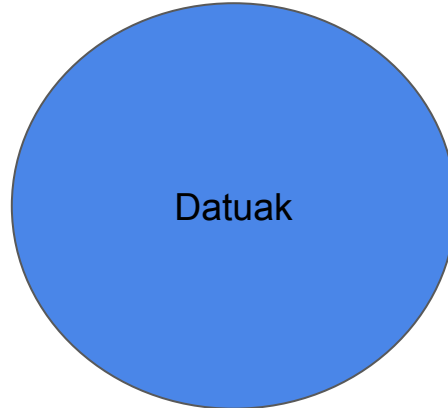
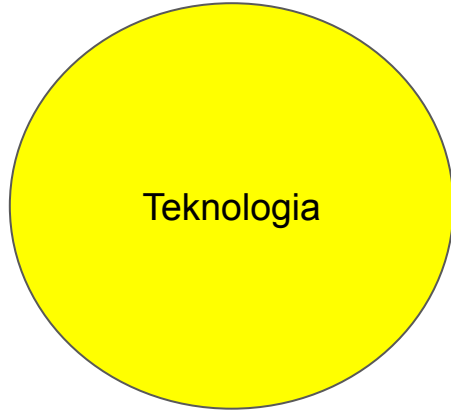
Sarrera, Digitalizazioaren Aukerak eta Erronkak

Rodrigo Agerri

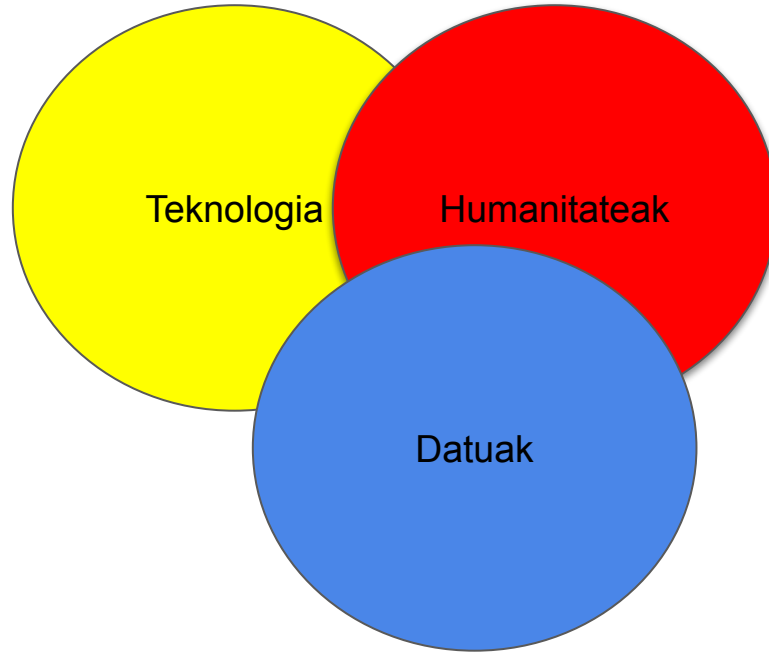
IXA taldea, HiTZ zentroa

<http://www.rodrigoagerri.net/>

Zer dira humanitate digitalak?



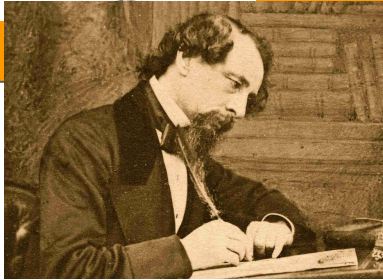
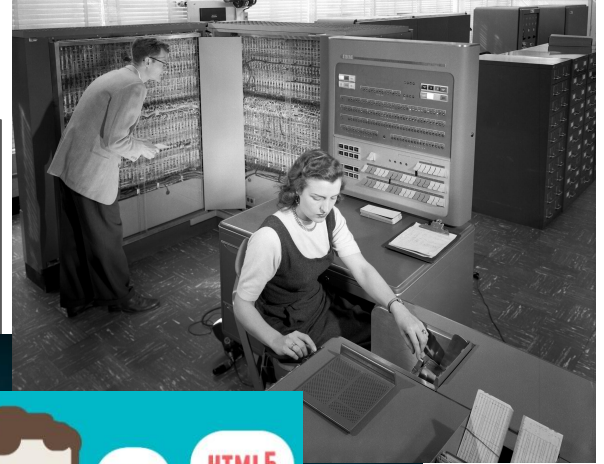
Zer dira humanitate digitalak?



Zer dira humanitate digitalak?



ki konput
teak oso



Zer dira humanitate digitalak?

- Dokumentuen digitalizazio masiboarekin, sare sozialen ez-tandarekin eta hizkuntz teknologien hedapenarekin arloen arteko lankidetzaren 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

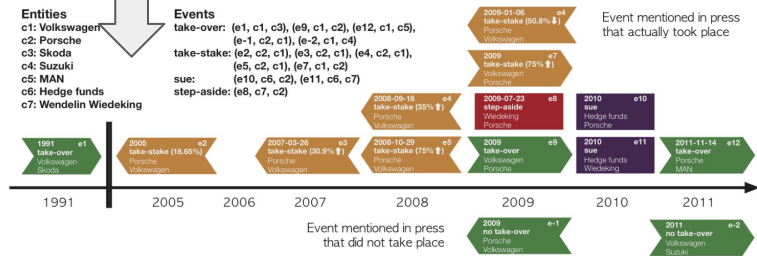
Teknologiak

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

Tehnologiak



```
<NAF xml:lang="eu" version="v1.naf">
  <nafHeader>
    <linguisticProcessors layer="text">
      <lp name="ixa-pipe-tok-eu" beginTimestamp="2016-04-25T11:25:29+0200" endTimestamp="2016-04-25T11:25:29+0200" version="2.0.0-0512ff8407b44bd9a9f9be9b27ca520758b03f" 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="2.0.0-2837752920a35cb0b501f7163aef2e2f7ba5f" 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 kontaktak (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

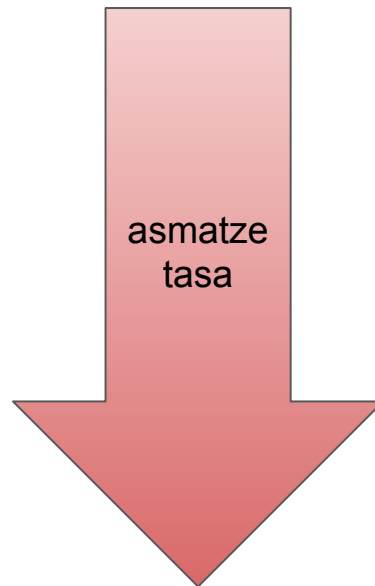
- Lehenetsuna lehiatzea da (pribatuen interesak ikerkuntzan)
- Datuak enpresa erraldoien esku egotea eta ez denon eskura
- Eraitzen interpretagarritasun falta (teknologia kutxa beltza izan daiteke)
- Moda izatea

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 Identifikazioa
- Denbora-lerroak

Hizkuntza ez-estandarren normalizazioa



Hurbilpenak

- Datuetan oinarrituta
 - Ikasketa automatikoan oinarrituta
 - Emaizta 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>



Laburpenak egiteko tresna:

Compress eu

Compress eu erabili

- Eskolako laburpenak jasotzeko tresna, diskurtso-egiturako informazioan oinarritua:
 - Estrakziozko laburpenak
 - Abstrakziozko laburpenak

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...**

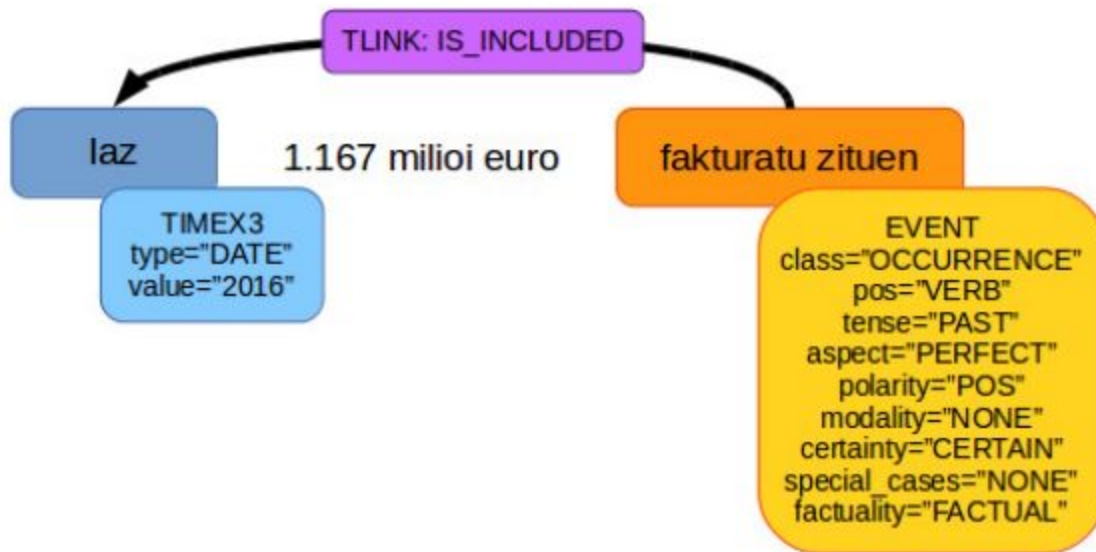
```
<entity id="e7" type="PER">
  <references>
    <!--Varufakisek-->
    <span>
      <target id="t109" />
    </span>
  </references>
</entity>
```

Iritzien Azterketa

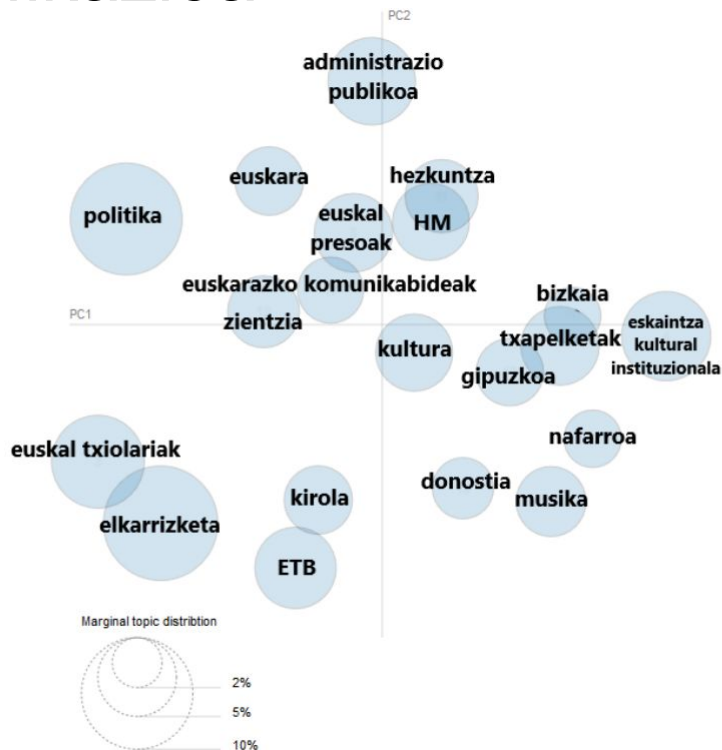
Jabeak ere maitagarriak dira.

```
<opinion oid="00">
  <!-- 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

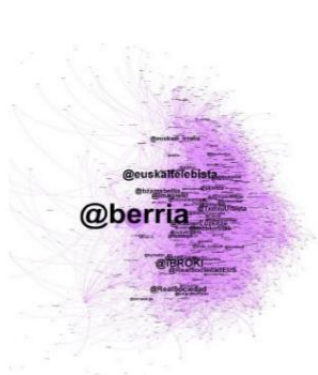


Gaien Identifikazioa



LDavis: [20 topiko](#)

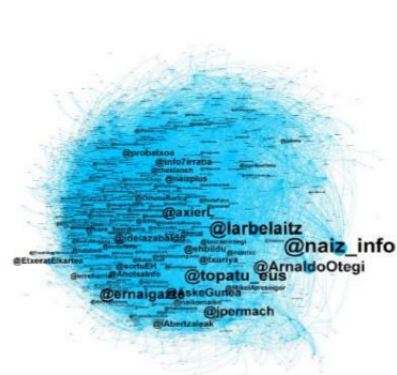
Gaien Identifikazioa



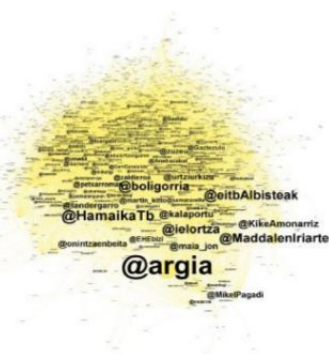
(a) Kirolak (% 21,61)



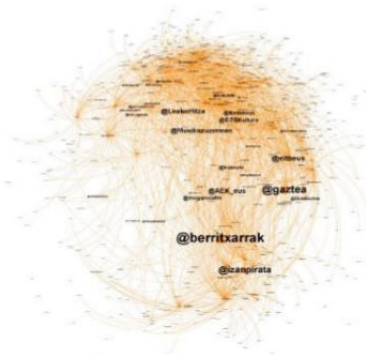
(b) Euskara (% 20,70)



(c) Ezker Abertzalea (% 17,12)



(d) Albisteak (% 14,92)



(e) Musika (% 11,35)

Denbora-lerroak

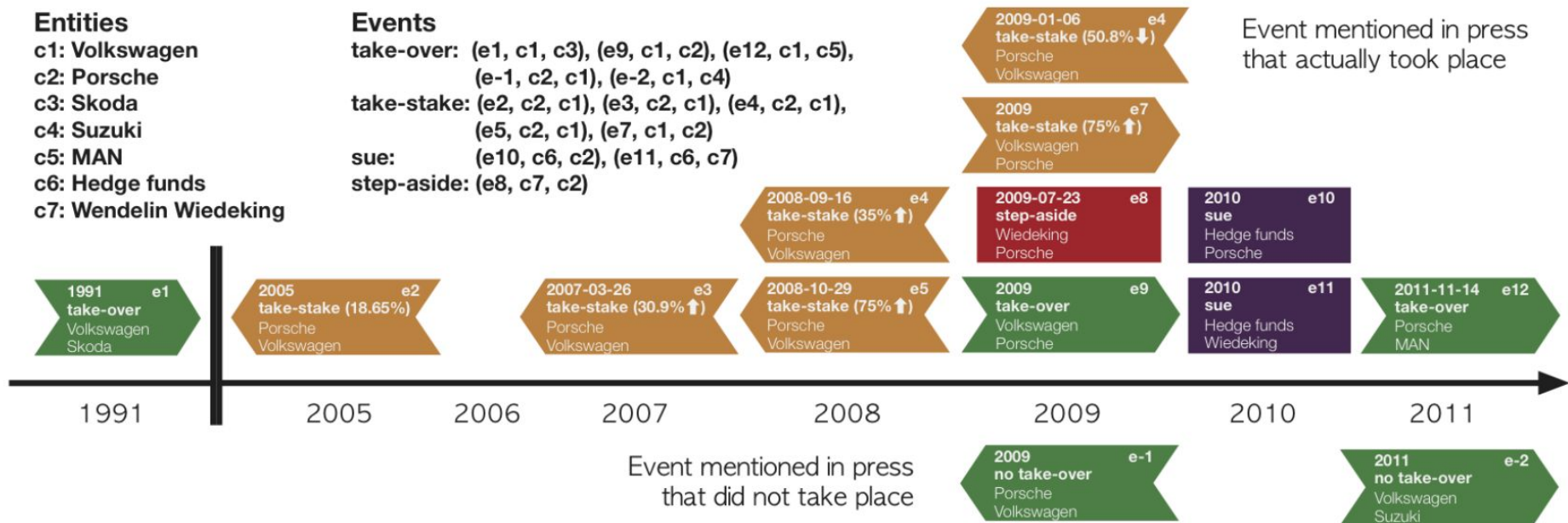
Entities

- c1: Volkswagen
- c2: Porsche
- c3: Skoda
- c4: Suzuki
- c5: MAN
- c6: Hedge funds
- c7: Wendelin Wiedeking

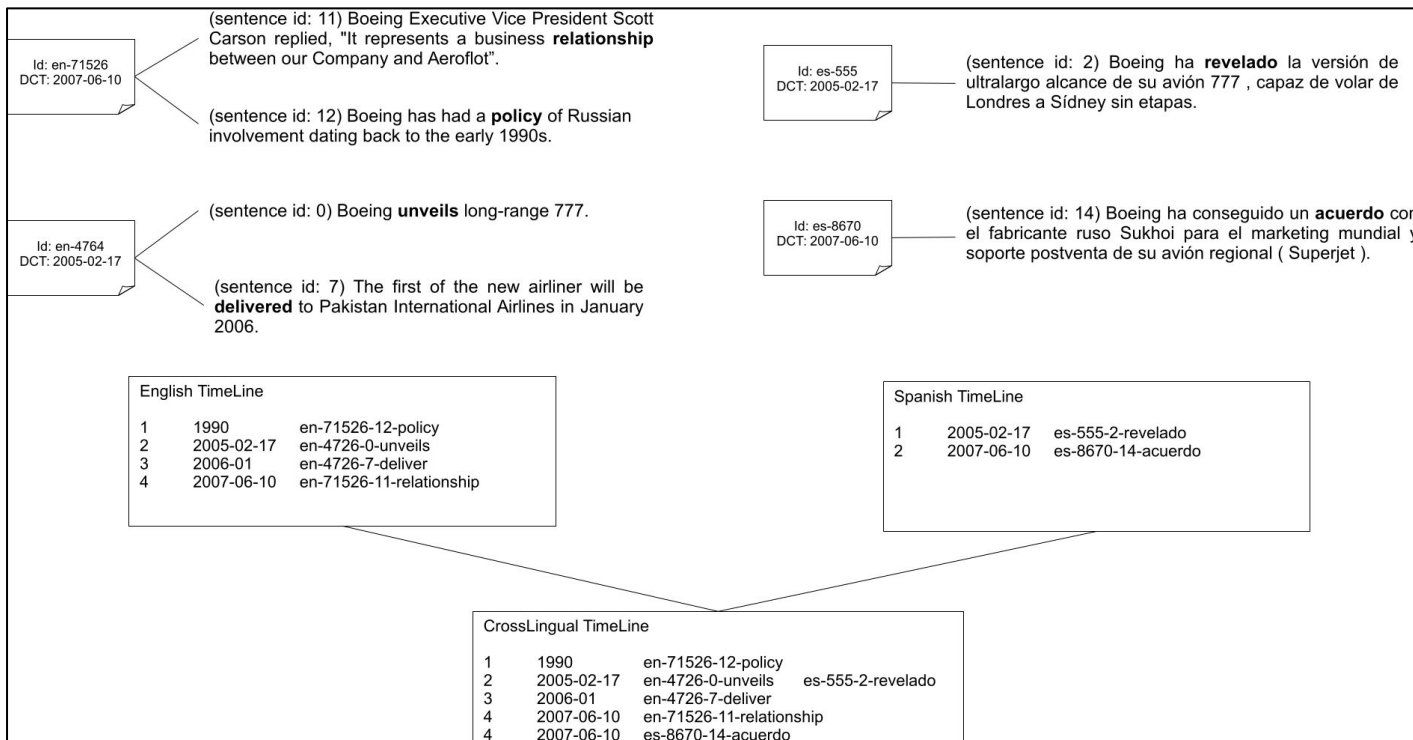
Events

- take-over: (e1, c1, c3), (e9, c1, c2), (e12, c1, c5), (e-1, c2, c1), (e-2, c1, c4)
- take-stake: (e2, c2, c1), (e3, c2, c1), (e4, c2, c1), (e5, c2, c1), (e7, c1, c2)
- sue: (e10, c6, c2), (e11, c6, c7)
- step-aside: (e8, c7, c2)

Event mentioned in press that actually took place



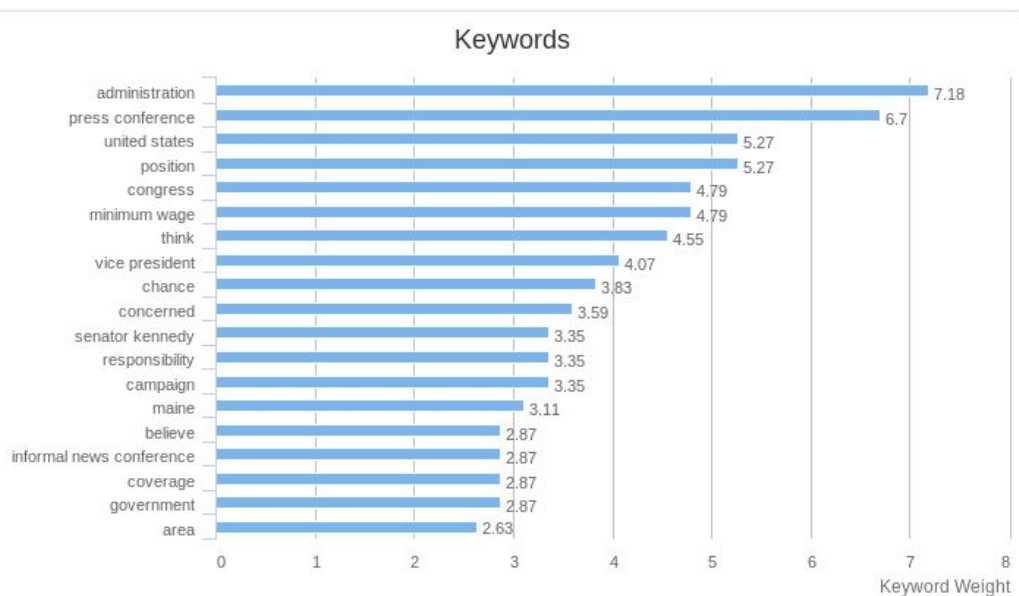
Denbora-lerroak (eleanitzak)



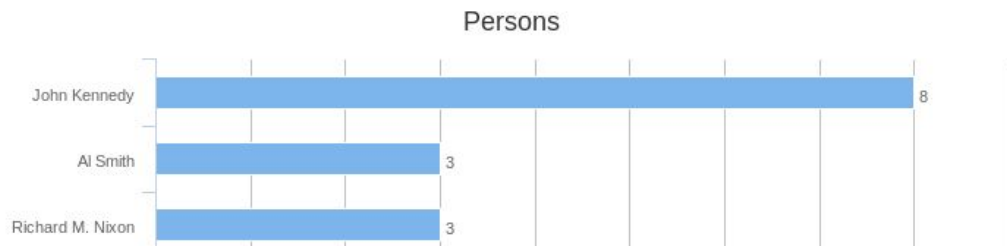
Bisualizazioa: <https://voyant-tools.org/>

The screenshot displays the Voyant Tools web interface. On the left, a word cloud features prominent terms like 'navarra', 'psnk', 'geroa', 'sumaren', 'duen', 'botoak', 'alkatetzatza', 'sozialistek', 'duen', 'lorru', 'absolutua', 'gehiengo', 'nabarroako', 'baik', 'duen', 'botoak', 'udal', 'bestelako', 'akordioak', 'osatzeko', 'bestelako', 'duen', 'botoak', 'udal', 'bestelako', 'akordioak', 'osatzeko'. The central text area shows a snippet titled 'Maiatzaren 26ko hauteskundeek panorama irekia a...' with the following text: 'Maiatzaren 26ko hauteskundeek panorama irekia azalatu zuten Nafarroako udal askotan eta, batez ere, Nafarroako Parlamentuan. Navarra Sumak lortu zuen parlamentari gehien, baina, gehiago absolutua lortu ezinik, ikuskeriko dago zer gehiago osa daitezkeen gobernuko lehendakaria aukeratzeko eta gobernu bera osatzeko. Horiek horrela, zalanba bat da nagusia: PSNK zer egingo duen; hautagaitzarik aurkeztuko duen, gehiago absolutua edo soil osatzeko gai izango den, horretarako zinekin eta zertaz hitz egiteko prest dagoen, eta, PSNren jokabidearen arabera, Geroa Baik, EH Bilduk, Ahal Duguk eta Ezkerrek nola jokatu zuten. Testuinguru horretan, gobernu lehendakariaren hautaketan gerta daitezkeenaren aurrekari bat izan daiteke atzo udalen osaketan eta alkatteen inbestidura sailotan gertatutakoa. Ikuskituz zegoen PSN nola lerratzen zen bi blokean artean —laukoa eta Navarra Suma—, eta sozialistek mezu argia bidali zuten atzo: ez daude aldatuta batesteko. PSN beharrezko aldagaiak ez zen ekiazioetan soilik eskuratu zuten laukoa osatzen duten alderdiak alkatetzatza: Tafallan, Berriozarren, Antsoainen, Garresen, Zizur Nagusian, Altsasun... Horrez gain, PSNK Navarra Sumaren esku utzi zituen bestelako gehiengoak izan zitzaizkiren udalak, zinegotzi sozialistek beren alkatetza beren aldeko botoa emanda; horrela gertatu zen Iruhean, Edeusbarren, Lizarran, Barañainen eta Burlatan, esaterako; hiriburuan eta...'. On the right, a line graph titled 'Relative Frequencies' shows the frequency of terms across 10 document segments. The legend includes 'eh' (green), 'geroa' (blue), 'navarra' (purple), 'psnk' (red), and 'sumaren' (cyan). The graph shows peaks for 'psnk' at segments 4 and 5, 'geroa' at segment 7, and 'sumaren' at segment 10. Below the graph is a table with columns 'Document', 'Left', 'Term', and 'Right'. The table contains 10 rows of text segments with corresponding terms highlighted in the original image. The bottom of the interface shows a 'Summary' tab with statistics: 'This corpus has 1 document with 627 total words and 399 unique word forms. Created now. Vocabulary Density: 0.636. Average Words Per Sentence: 22.4. Most frequent words in the corpus: psnk (9), geroa (8), navarra (8), eh (7), sumaren (5)'. The bottom left corner features the 'ix' logo.

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

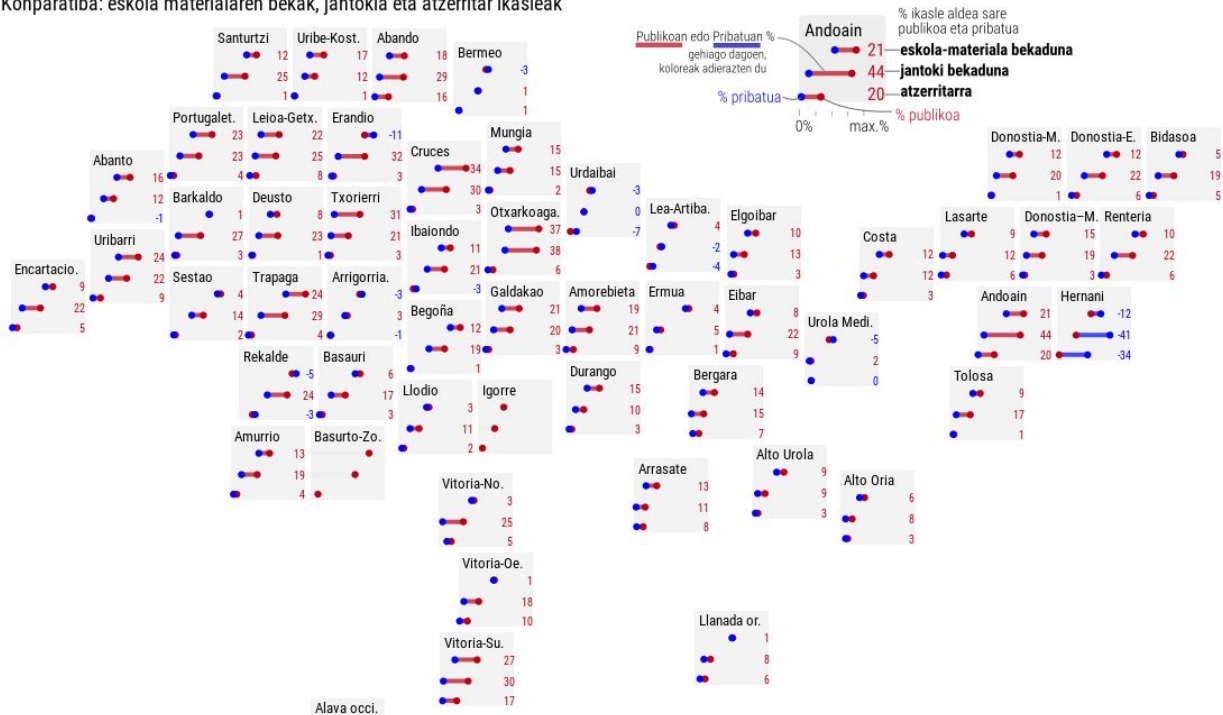


■ Relevance



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

Konparatiba: eskola materialaren bekak, jantokia eta atzerritar ikasleak



Oso simple izan daiteke

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

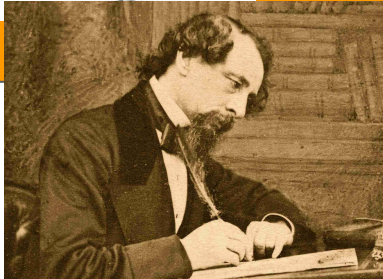
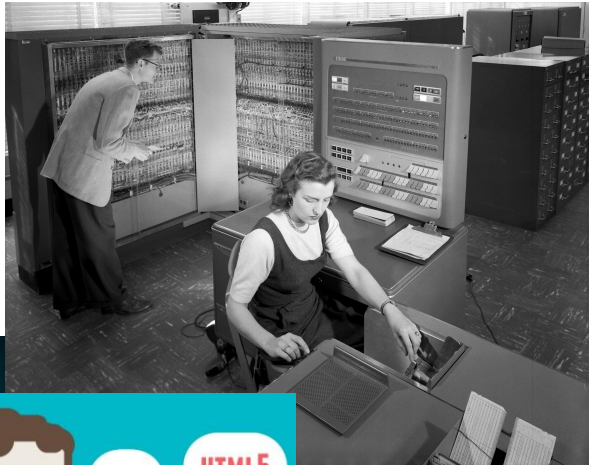
```
9 PSNk
9 ez
8 Navarra
8 Geroa
8 bat
7 zen
7 EH
7 du
6 ere
5 zuen
5 Sumaren
5 izan
5 ditu
5 da
4 zer
4 udal
4 sozialistek
4 Nafarroako
4 lortu
4 gehiengo
4 beharko
4 Baik
4 alkatetza
3 zuten
3 osatzeko
```



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 eskiribatukoituz alde batera liburuan barrena, eta bertze aldera liburuaren bazterrean, in margine : bat bederak zerbaiz

18 IRACVRTÇAILLEARI.

chehero, bertceac gehero. Batac chedea, bertceac, gedea. Batac ichilic, bertceac igilic. Batac lachoa, bertceac lajoa. Batac, choil, bertceac, joil. Batac quecho, bertceac quejo. Batac chuchen, bertceac, jugen. Eta hunela bada, bertceric ere cenbait **hitz**, batac eta bertceac, nork bere herrico edo erresumaco arauaz differentqui esquiribatzen baitituzte.