Kultura digitalizatua: Europa eta Euskal Herria

Itziar Gonzalez-Dios eta Izaskun Etxeberria
UPV/EHUko HiTZ zentroa, Ixa taldea
@Hitz_zentroa @IxaTaldea

Bikaintasuna Euskal Ikasketetan IX.
Euskara eta euskal kultura eremu digitalean
Donostia, 2019/VII/10

Nor gara?

Universidad del País Vasco Euskal Herriko Unibertsitatea

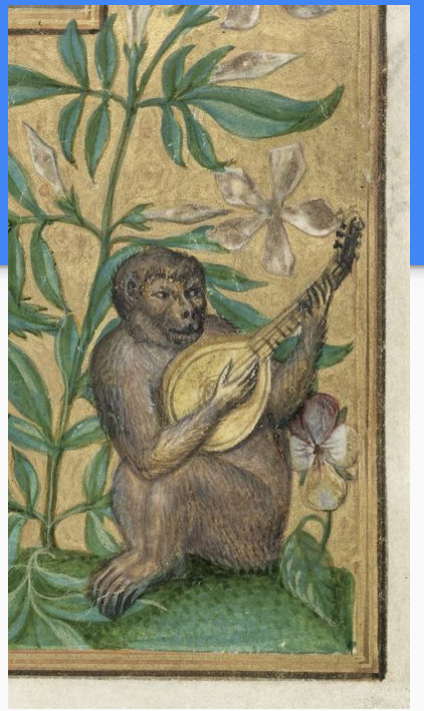
HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

- Ixa ikerketa-taldea: 30 urteko eskarmentua
  - Hizkuntzalaritza eta Informatika
- Izaskun Etxeberria
  - Informatikan doktorea, Informatika Fakultateko irakaslea (Konputagailuen Arkitektura eta Teknologia saila)
- Itziar Gonzalez-Dios
  - Hizkuntzalaritza konputazionalean doktorea, Bilboko Ingeniaritza Eskolako irakaslea (Euskal Hizkuntza eta Komunikazioa saila)

Kultura digitalizatua

Gero eta gehiago digitalizatu…

- Europeana: *metaliburutegia*, 50 milioi item baino gehiago
  - Liburuak, musika, artelanak…
- Project Gutenberg: liburutegi digitala
  - 59.000 *ebook* baino gehiago
- Eusko Ikaskuntzaren Dokumentazio Zentro Digitala
  - Eusko ikaskuntzak sortu dituen eduki zientifiko eta kulturalak
    - Adib. Auñamendi entziklopedia, kantutegia, multimedia- eta dokumentu-fondoak



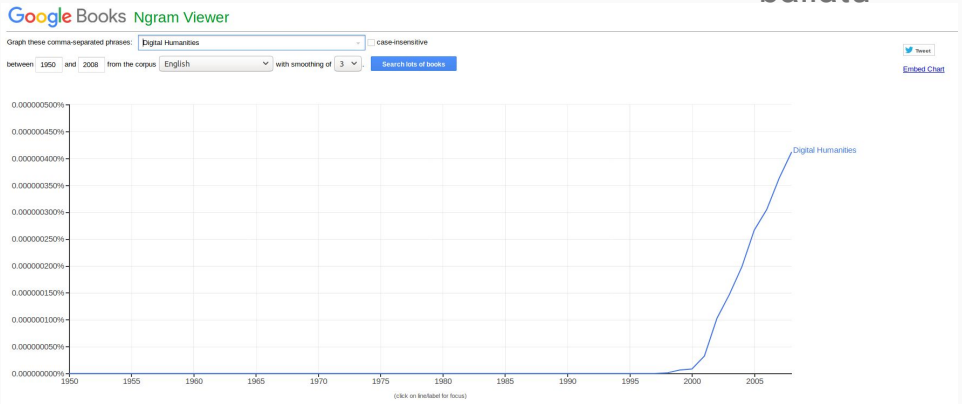
Monkey playing lute
from BL Harley (1582)

*Meatze*ak Giza eta Gizarte Zientzien ikerketan!

Humanitate digitalak (eHumanitateak) eta Gizarte Zientzia digitalak

- Jakintza-alor emergenteak
- Talde, erakunde ugari azken urteetan
- Aitzindaria: *Literary and Linguistic Computing*

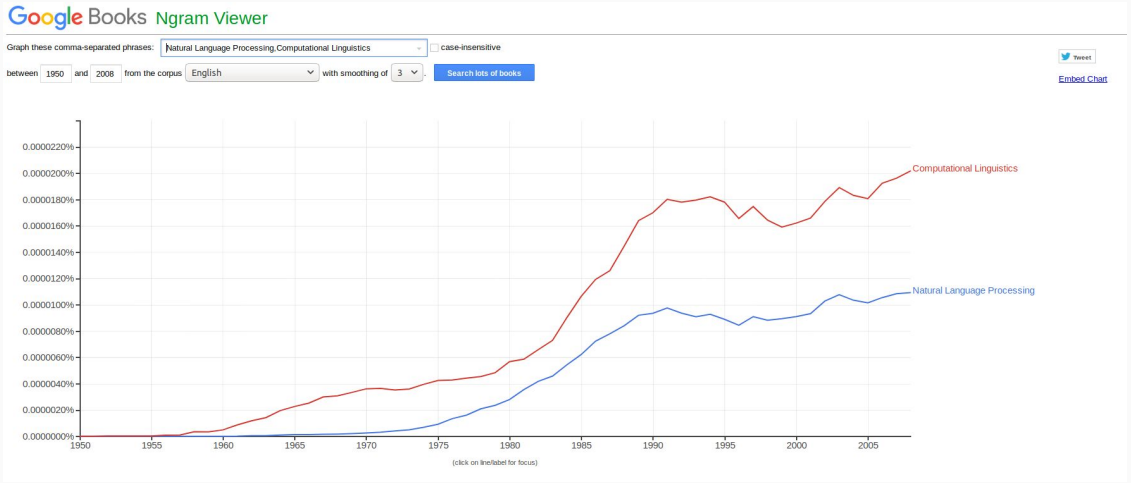
- Humanitateetako eta Gizarte Zientzietako gaiak landu
- IKTak erabili
- Datu (testuak, irudiak...) digitalizatuak baliatu



Nola *ustiatu*?

Hizkuntzaren Prozesamendua (HP)

- Informatika, Adimen artifiziala, Hizkuntzalaritza uztartu
- Zer egin?
  - Hizkuntza baliabideak sortu
  - Informazioa erauzi eta ikasi
  - Sare sozialak monitorizatu
  - ...
- Aplikazioak:
  - Itzulpen automatikoa
  - Elkarrizketa-sistemak
  - Zuzentzaile ortografikoak
  - ...

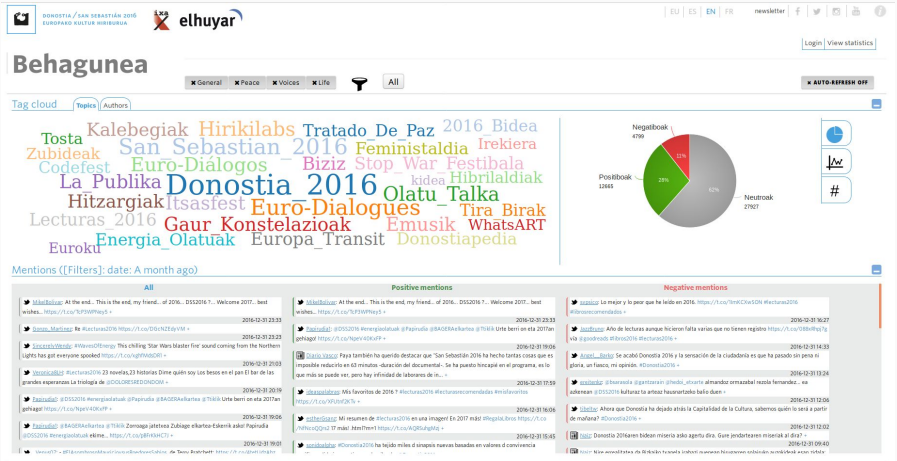


Zer egin daiteke HPko tresnekin?

- *Wordcloud*-ak: maiztasun handiena duten hitzak irudikatu
- Hitzak antonimoekin ordezkatu
  - Few years ago there lived a dear large girl who was beloved by every one who ignore her.
- Beste hizkuntzatako hitzekin edo irudiekin ordezkatu
  - La bedstemor comió el kage y bebió el vinsort.
- Erabilitako teknologia:
  - Analisirako tresnak: analisi morfosintaktikoa + adieren desanbiguazioa
  - Datu-baseak: Open Multilingual Wordnet eta ImageNet
- Informazio gehiago: Agirrezabal et al. (2019)

Baina ez hori bakarrik...

- Zeri buruz ari gara Twitteren?
  Sentimenduen analisia, iritzien erauzketa...



- Zein dira testuaren topikoak (*topic modelling*)?
- Adib., Axularren *Gveron*
  - Miserikodia, bekhatu, esperantza...
  - iuramentu, denbora...
  - erran, hil...
  - haragia, gaitza, etsaia, barkhatu
  - gaizki, konzientzia, ongi
- Baina prozesamendua behar!
  - Maiztasun altuko hitzak ezabatu
  - Aldakiak identifikatu...

Paste your text below!

A Downing Street spokesman said: "We have made clear to the US now unfortunate this leak is. The selective extracts leaked do not reflect the closeness of, and the esteem in which we hold, the relationship."
But he said ambassadors needed to be able to provide honest assessments of the politics in their country, and the prime minister stood by Sir Kim.
"The UK has a special and enduring relationship with the US based on our long history and commitment to shared values and that will continue to be the case," he said.

Go!

Spiral: Archimedean Rectangular | 5 orientations from −60 ° to 60 ° | Number of words: 250

Scale: log n √n n | One word per line

Font: Impact | Download: SVG

Copyright © Jason Davies | Privacy Policy. The generated word clouds may be used for any purpose. How the Word Cloud Generator Works.

Paste your text below!

* Ezporogin 25 urteko gazte bat hil da. Ibaiak arrastaka eraman duen auto baten barruan gelditu da harrapatuta.
* Tafallako udalak larrialdi bilera egin du goizeko 7:30ean, kalteei aurre nola egin erabakitzeko. Uxue Barkos Nafarroako jarduneko presidentea kaltetutako eremuko alkateekin bilduko da goizean.
* Uholdeek etxebizitza, saltoki eta garaje ugaritan eragin ditu kalteak.
* Zidakos ibaiak 3,5 metro egin zuen gora bost orduko tartean.

Go!

Spiral: Archimedean Rectangular | 5 orientations from −60 ° to 60 ° | Number of words: 250

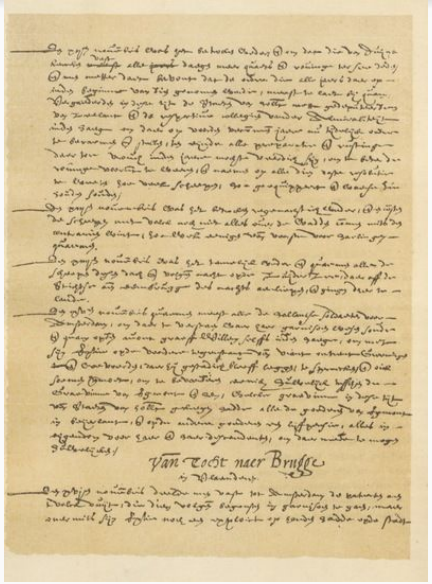
Scale: log n √n n | One word per line

Font: Impact | Download: SVG

Copyright © Jason Davies | Privacy Policy. The generated word clouds may be used for any purpose. How the Word Cloud Generator Works.

HPak dituen humanitate digitaletan erronkak

- **EZ DA KLIK EGITEA!!!**
- Argazki/testu hutsetatik abiatuta,
  - OCR akatsak
  - Hizkera ez-estandarra, batez ere testu zaharretan
    - BIM proiektua
- Lizentziak
  - Libreak vs jabegodunak
- Mugak
  - Noren esku daude datuak?
  - Nola interpretatu emaitzak?



manuscript from "Journaal van A. D., ... 1591-1602


Help

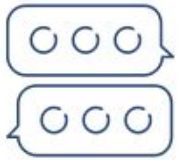
Nola lortu *tresneria*?
Eta zer erabili?


support

Europako azpiegiturak

- Common Language Resources and Technology Infrastructure (CLARIN)
  - Europako ikerketa-azpiegitura (ERIC)
  - **Helburua:** Europako hizkuntza guztien baliabide digitalak webgune bakarrean eskuragarri egitea, Humanitate eta Gizarte Zientzietako ikertzaileei laguntzeko
    - CLARIN K-zentroen sarearen bidez
  - **Kideak:** Austria, Bulgaria, Txekiar Errepublika, Danimarka, *Dutch Language Union*, Estonia, Finlandia, Alemania, Grezia, Hungaria, Italia, Letonia, Lituania, Herbehereak, Norvegia, Polonia, Portugal, Eslovenia eta Suedia
  - **Begiraleak:** Frantzia, Erresuma Batua, Islandia eta Hego Afrika

CLARIN Resource Families (portal)



Computer-mediated communication corpora

Historical corpora

L2 learner corpora

Manually annotated corpora

Newspaper corpora

Parallel corpora

Parliamentary corpora

Spoken corpora

es-CLARIN-K

- Lau talde:
  - Pompeu Fabra Unibertsitateko (UPF) CLARIN IULA
  - UNEDeko Laboratorio de Innovación en Humanidades Digitales
  - EHUko Ixa Taldea
  - Vigoko unibertsitateko TALG taldea
- Zer eskaini?
  - **Aholkularitza birtuala** gai praktikoei buruzko galderak erantzuteko eta aholkuak emateko
  - **Auto-ikasketetarako laguntza** baliabide espezializatuak erabilita
  - **Prestakuntza-programen antolakuntza**
  - Proiektu teknologikoen **kudeaketa- eta plangintza-zerbitzuak**

es-CLARIN-K

Meatzea ustia dezakegu!

Eta Euskal Herrian?

BIM proiektua

Iñaki Alegria, Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Céline Mounole, Manuel Padilla, Ander Soraluze

AURKIBIDEA

1. SARRERA

- *Basque in the making (BIM): A Historical Look at a European Language Isolate*
  - ANR-ek finantzatutako 4 urteko proiektua (2017-2021)
  - Hainbat partaide: **IKER UMR 5478 (CNRS, UBM, UPPA), Ixa (UPV/EHU),** Monumenta Linguae Vasconum (UPV/EHU), HiTT (UPV/EHU), Basdisyn (UPV/EHU), Deustuko Unibertsitatea
  - Ikerketa-burua: Ricardo Etxepare

1.1. Helburu nagusia

- **Euskararen gramatikaren ezaugarri zenbaiten azterketa diakroniko zehatza egitea**
  - Determinatzaileak
  - Aditz-egitura perifrastikoak
  - Numero morfologikoaren jatorria eta bilakaera
  - Kasu morfologikoaren eta funtzio gramatikalen arteko komunztadura
  - Forma pronominal zehaztugabeen sorrera
  - Postposizio-egituren bilakaera
  - Aditz laguntzaile zehaztuen egoera
  - Galdegaiaren eta aditzaren arteko hurrentasun-hertsidura
- **Horretarako, sintaktikoki etiketatutako corpus historikoa sortzea**

Jainkoaren hitz purari jarreikiteko desira dutenék,
sporzu dugu eridenen dutela (suporturekin) zerzaz kontenta.

"<**Jainkoaren**>"
"jainko" IZE ARR BIZ+ GEN NUMS MUGM HAS_MAI @IZLG>

"<**hitz**>"
"hitz" IZE ARR BIZ- ZERO @KM>

"<**purari**>"
EZEZAG "pura" ADJ ARR DAT MG @ZOBJ
EZEZAG "pura" ADJ ARR DAT NUMS MUGM @ZOBJ

"<**jarreikiteko**>"
EZEZAG "jarreikit" IZE ARR GEL NUMS MUGM @IZLG>

"<**desira**>"
"desira" IZE ARR BIZ- ZERO AORG @KM>

"<**dutenék**>"
EZEZAG "dutenk" ADJ ARR ABS MG @OBJ

*Jainkoaren hitz purari **jarreikiteko** desira **dutenék**,*
***sporzu** dugu eridenen dutela (suporturekin) **zerzaz** kontenta.*

"<**sporzu**>"
EZEZAG "sporzu" ADJ ARR ABS MG @OBJ

"<**dugu**>"
"ukan" ADT PNT A1 NOR_NORK NR_HURA NK_GUK @+JADNAG

"<**eridenen**>"
"ediren" "eriden" ADI SIN PART GERO NOTDEK @-JADNAG %ADIKATHAS

"<**dutela**>"
"*edun" ADL KONPL A1 NOR_NORK NR_HURA NK_HAIEK-K @+JADLAG_MP_OBJ %ADIKATBU

"<**suporturekin**>"
EZEZAG "suportu" IZE ARR SOZ MG @ADLG

"<**zerzaz**>"
EZEZAG "zerz" IZE ARR INS NUMS MUGM @ADLG

"<**kontenta**>"
"kontent" ADJ ARR IZAUR- ABS NUMS MUGM @OBJ

2. URRATSAK

- Corpusa bildu eta prestatu (IKER)
- Corpusa etiketatu (Ixa)
  - Corpusa “normalizatu”
    - Eskuz
    - Automatikoki
  - Etiketatze morfosintaktikoa
- Bilaketa-interfazea garatu (Ixa)

2.1. Corpusa bildu eta prestatu

- Corpusaren ezaugarriak
  - Euskara Arkaikotik XVIII. mendera
  - Euskalki guztiak
  - Genero aniztasuna
  - 750.000 hitz inguru
- Lan filologikoa
  - Dauden bertsioetatik abiatuta eta edizioaren kalitatearen arabera
    - Testua orraztu
    - Transkripzio berria egin
  - Grafia *OEH*ren ereduari jarraiki: egungo grafia sistema, baina ezaugarri fonologikoak errespetatuz

2.2. Corpusa etiketatu
2.2.1. Normalizazioa

- Aurrekaria: I. Etxeberriaren tesia (2016): Axularren *Gero* eta Mogelen *Peru Abarka*
- Normalizazioa: zati bat eskuz (zoriz), gainontzekoa “automatikoki”
- Eskuzko normalizazioa:
  - Tamaina egokiko “sekzioak” sortu (30-60 hitz)
  - “Sekzioen” zozketa: obraren % 10
  - Tokenizatzailea
  - Entitate (izen bereziak) ezagutzailea
  - EDBLrekin sortutako ezagutzailea
- Hasierako etiketak ezarri hitz bakoitzari: **ENT-Zuz**, **STD-Zuz**, **OOV (Out Of Vocabulary)**

2.2.1. Normalizazioa: eskuzko zatia

- Hitzez hitzeko etiketatzea
- Beharrezkoa denean, forma “estandarra” esleitu (*Euskaltzaindiaren Hiztegia*)
  - *etzuen > ez zuen*
  - *burhaso > guraso*
  - *hetarik > haietarik*
- Fenomeno morfosintaktiko interesgarriak markatu: **SEK_Ber** etiketa
  - Aoristoa / subjuntibo zaharra / prosekutiboa...
- Brat tresna erabili dugu eskuzko etiketatzea egiteko
- Eskuzko lanaren denbora-estimazioa: 140 hitz inguru orduko
  - Pertsona batek corpus osoa eskuz etiketatzeko: 750 lanegun inguru (3 urte)

65 Eta anhitz berze gauzarik erraiten zutén haren kontra desondratzen zutela .

66 Eta argitu zenean bil zitezen populuko Anzianoak , eta Sakrifikadore prinzipalak , eta Skribák , eta eraman zezaten hura bere konseillu barnera ,

---------7 1081,1082---------

5 Baina orain banoa ni igorri nauenaganát , eta zuetarik nehork ez nau interrogatzan : Norat oha ?

6 Baina zeren gauza hauk erran drauzkizuedan , tristiziák bethe du zuen bihotza .

---------8 721,722---------

Halakoetan , gauzá zeren den dakitenetarik batbedera , orhoituren da , othoi , ezen hunelako gauzák , guziz lengoaje oraino usatu gabe batetan , ezin behingoaz halako perfekzionetan jar daitezkela , nola behar bailizateke .

Guziagatik ere minzatzeko manerán anhitz arrastatu gabe , Jainkoaren hitz purari jarreikiteko desira dutenék , sporzu dugu eridenen dutela (suporturekin) zerzaz kontenta .

Eta baldin speranza dugun bezala , oraindrano egin den hunetán heuskaldunak gozorik edo edifikazionerik hartzen badu , hunetan enplegatu izan diradenék bihotz harturen duté , oraindanik gogo ere duten bezala , egin denaren berriz ikhusteko eta korrijitzeko , bai eta , baldin Jaunak hala plazer badu , pasaje difizilenén deklaragarri anotazionén ezarteko .

Edit Annotation

Text

anhitz Link

Search

Google, Wikipedia

Entity type

- OOV
  - OOV-Zuz
  - OOV-Ald
    - OOV-AldADIJ
  - OOV-Zalant
    - OOV-ZalantADIJ
- STD-Zuz
  - STD-Ald
    - STD-AldADIJ
  - STD-Zalant
    - STD-ZalantADIJ
- ENT-Zuz
  - ENT-Ald
  - ENT-Zalant
- LEX-Berri
- ErrTipo
- HEE
- EEE1

Notes

anitz

Add Frag. | Delete | Move | OK | Cancel

2.2.1. Normalizazioa automatikoki egiteko

- Eskuz etiketatutako zatiarekin **ikasi**
  - *Phonetisaurus* tresnan oinarritzen da (Novak, 2012)
  - Grafemetan oinarritzen da
  - “aldaera/estandar” eta “estandar/estandar” bikoteekin ikasi
  - Sarrera berri bat → dagozkion *n* erantzun probableenak → iragazpen prozesua
- Ikasitakoaren ebaluazioa
  - *Leave-one-out-cross-validation* metodoa
  - 10 zati, 9 ikasteko eta 1 testeatzeko (10 esperimentu)

2.2.2. Normalizazioaren emaitzak: Ebaluazioa (Leizarraga eta *RS*)

	Leizarraga
	Asmatze-tasa
Test: guztiak	94,24
Test: aldaerak	87,68

	RS
	Asmatze-tasa
Test: guztiak	80,66
Test: aldaerak	66,83

Leizarraga: **% 68,36 STD** / % 28,42 OOV / % 3,21 ENT

Refranes y Sentencias: **% 55,43 STD** / % 43,85 OOV / % 0,71 ENT

2.2.2. Errore-analisia: sailkapena

- Normalizazio-sistema doitzeko erroreak errepasatu
- Leizarragaren kasuan, errore guztien heren bat hartu eta sailkatu (131 errore):
  - Ondorengo analisian eraginik ez dutenak > **% 22,90 (30)**
  - Ondorengo analisian eragina dutenak > % 72,51 (95)
  - Eskuzko etiketatzearen akatsa > % 3,05 (4)
  - Bestelakoak > % 1,52 (2)

2.2.2. Errore-analisia: ondorioak

- Zenbaitetan errorea BAI, analisian eraginik EZ (% 22,90) > analisiaren araberako bilaketak arazorik gabe
  - *franzesez > frantzesez (frantsesez)*
  - *berthuterekilako > bertuterekilako (bertuterekiko)*
  - *erekharri > ekarri (erakarri)*
- Errore asko adizkiekin (% 25,19)
  - *lerroten > lerran diezaioten (liezaioten erran)*
  - *daidigularik > dagigularik (dagikegularik)*
  - *dagienzat > dagin dadin (dagien)*
  - *lizenzat > zizen (izan ledin)*

2.2.2. Errore-analisia: ondorioak

- Zailtasunak 2>1 eta 1>2 aldaketak normalizatzeko (% 12,21)
  - *trebuka eraziten > treboerazten (trebukarazten)*
  - *goga eraziteko > gogaerazteko (gogarazteko)*
  - 1>2 adibideak: ikasi du banatu egin behar dela eta nola, baina adizkia gaizki:
    - ***etz**akion > **ez z**ekion (ez zitzaion)*
    - *e**zp**alitzaue > e**z b**alitzaue (ez balitzaie)*

- Arazoa: STD-Zuz zirenak gaizki egitea (% 5,34)
  - *haur > hau (haur)*
  - *juje > juja (juje)*

2.2.2. Ondo ikasitakoak

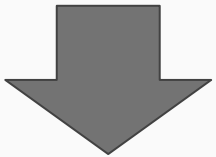
- Kontsonante hasperendunak
  - *aiphatzez > aipatzez / berthute > bertute / bekhatuez > bekatuez / anhitz > anitz*
- *-zione > -zio*
  - *adopzionearen > adopzioaren / afekzione > afekzio / estimazionerik > estimaziorik*
    - *detestazionetan > *detestaziotan (detestazionetan)*
- Azentuak kendu
  - *aiték > aitek / bethé > bete / enseiñák > entseinak / duté > dute*
- Subjuntibo zaharra *-zat*
  - *dadinzát > dadin / dezagunzat > dezagun / ditzagunzat > ditzagun / eztezadanzat > ez dezadan*
    - *behinzat > *behin dadin (behintzat) / dagoenzat > *dagoen (egon dadin)*
- Hasierako *-r / -s > erre- / es-*
  - *regina > erregina / rendatu > errendatu / skriba > eskriba / speranza > esperantza / spiritu > espiritu*

2.2.2. Ondo ikasitakoak

- Adizkiak
  - *dirade > dira / date > dateke / drauku > derauku / dadukan > daukan*
  - *dauen > dagoen /baitzateken > baitzatekeen / nezakela > nezakeela*
- Aditz-izenetan *-ite > -te*
  - *egoite > egote / emaiten > ematen / erraiten > erraten*
- Aldaerak
  - *bertze > beste / guziagatik > guztiagatik / saindu > santu / kristek > kristok / hek > haiek*
  - *xipitako > txikitako / hunek > honek*
    - *hetango > *haietango (haietako)*

2.2.3. Etiketatze morfosintaktikoa

- Behin testua normalizatuta, egungo tresnak gai dira analizatzeko
- *Eustagger* tresnarekin etiketatu
  - Tokenizazioa
  - Lematizazioa
  - Analisi morfosintaktikoa
    - Analisi morfologikoa
    - Funtzio sintaktikoak



Jainkoaren hitz purari jarreikiteko desira dutenék, sporzu dugu eridenen dutela (suporturekin) zerzaz kontenta.
Jainkoaren hitz puruari jarraikitzeko desira dutenek, esportzu dugu edirenen dutela (suporturekin) zertaz kontenta.

"<Jainkoaren>"
"jainko" IZE ARR BIZ+ GEN NUMS MUGM ZERO HAS_MAI @IZLG>
"<hitz>"
"hitz" IZE ARR BIZ- ZERO @KM>
"<puruari>"
"puru" ADJ ARR IZAUR- DAT NUMS MUGM @ZOBJ
"<jarraikitzeko>"
"jarraiki" ADI SIN ADIZE GEL ZERO @-JADNAG_MP_IZLG> %ADIKAT
"<desira>"
"desira" IZE ARR BIZ- ABS NUMS MUGM AORG @OBJ
"<dutenek>"
"ukan" ADT_IZEELI PNT ERG NUMP MUGM A1 NOR_NORK NR_HURA NK_HAIEK-K @SUBJ
"<esportzu>"
"esportzu" IZE ARR BIZ- ABS MG @OBJ
"<dugu>"
"ukan" ADT PNT A1 NOR_NORK NR_HURA NK_GUK @+JADNAG
"<edirenen>"
"ediren" ADI SIN PART GERO NOTDEK @-JADNAG %ADIKATHAS
"<dutela>"
"*edun" ADL KONPL A1 NOR_NORK NR_HURA NK_HAIEK-K @+JADLAG_MP_OBJ %ADIKATBU
"<suporturekin>"
EZEZAG "suportu" IZE ARR SOZ MG @ADLG
"<zertaz>"
"zer" DET NOLGAL MG INS @ADLG
"<kontenta>"

2.2.3. Etiketatze morfosintaktikoa

- Lantzen ari garen fasea
- Sekzio bereziak **eskuz** errepasatu behar ditugu (aoristo, subjuntibo zaharra…)
- Aurreikuspena: fenomeno sintaktiko bereziak analizatzeko, erregela berriak sortu
- Testu historikoak analizatzeko analizatzaile morfosintaktikoa

2.3. Bilaketa-interfazea

- Zeren arabera bilatu
  - Metadatuen arabera: egilea, garaia, euskalkia, generoa
  - Testuaren ezaugarri gramatikalen arabera: hizkuntzalariekin batera lantzen ari gara
    - Forma originala edo lema estandarra
    - Deklinabide kasuak
    - Kategoria gramatikalak
    - Adizkitegia (nnn, indikatibo orainaldia…)
    - Egitura sintaktikoak

2.3. Bilaketa-interfazea

- Adibideak
  - Erlatibo postnominala:
    - izena mugagabean + aditz-izena + ERL: *gizon etorri dena*
  - Aoristoa:
    - aditzoina + *edin/*ezan iraganean
  - Ezeztapenaren ordena berezia:
    - partizipioa/aditz-izena/aditzoina + *ez* + adlag.: *enzun eztaizu*
- Ezaugarri garrantzitsua: bilaketaren emaitzetan faksimilea eta bertsio eguneratua lotuta ikusteko aukera (ikus irudiak)

Bilaketa: guzti **Mota:** Lema **Bilatu**

Bilaketa: guzti **Mota:** Lema **Bilatu**

302 emaitza, 2 testutan

Jesus Krist Gure Jaunaren Testamentu Berria [290]

...eta goitiko gauza **guziak** gauza **guzietan** eta lekhu **guzietan**, berak plazer duen bezala. Eta ezta nehor...
...z gidamendurik batre gabe **guziagatik** ere hek haren bide bezala bilhatzera erideitera, ezagutzera...
...z egundano ere izan banitate baizen, berthute **guzitako** Jauna asistitu eta baliatu izan zaion...
...Filiperen emazteagatik, eta berak egin zituen gaixtakeria **guziakgatik**, eratxeki zezan haur ere berze **guzién**...
...lurreko nazione **guzien** artean elejitu eta hautatu ukhan baitzuen eta hek ziraden Israeleko haurrak, haei...

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
56 57 58 » + Ikusi guztiak

Refranes y Sentencias [12]

... . Garagarrilean neskea ezark alborean. . Doguna jan ta txiro izan. . Direanak direanegino. . Andra **guztiok** erzeti...
... . Kibel egiok ekatxari. . Txiroak dirudi insausti, **guztiok** arrika beti. . Katuak daroean okelea geiago da...
... Betiko itoginak arria zulatu ta aldi luzeak **guztia** aztu. . Buruko andia ta jate urria. . Zagokez...
...oro buru balz, andra **duztiok** buru zuri. . Ogiagaz ura, oragaz eroen elikatura. . Putxeak ogi baga...
... gaxtoak bernazakian. . Maiatz iluna ta bagil argia, urte **gustiko** ogia. . Loitzaen ganeko leia, euria...

« 1 2 3 » + Ikusi guztiak

Jesus Krist Gure Jaunaren Testamentu Berria

Testamentu zaharrak eta berriak irakasten draukuten guziaren somarioa.

Testamentu zaharreko liburuek irakasten gaituzte ezen Adamek, Noek, Abrahamek, Isaakek, Jakobek, Dabidek eta berze gure Aiték adoratu ukhan duten Jainkoa, dela egiazko Jainko bat bera, bothere guzitako eta eternal. Bere ontasun handiagatik zerua eta lurra eta hetan diraden gauza guziak bere Hitzaz kreatu dituena, zeinaganik gauza guziak heldu baitirade, eta hura gabe deusek izaterik eztu. Zeinek egiten baitu justizia eta miserikordia, bai eta goitiko gauza guziak gauza guzietan eta lekhu guzietan, berak plazer duen bezala. Eta ezta nehor hain ausartik izateko, non hari derron, zergatik egiten duena, hunela edo hala egiten duen. Gehiago, liburu hek deklaratzen draukute, nola Jainko guzizko gorák eta bothere guzitakoak, berze gauzák kreatu zituen ondoan, kreatu ukhan zuen Adam lehen gizona, bere irudi eta imajina spiritualera, hura egiten zuelarik gauza guzién jabe lurrean. Baina deabruaren inbidiaz eta enganioz desobedienziatan eror zedin Adam bere Kreazalearen manamenduaren iragaitez eta haustez. Eta bere bekhatuaz halako khotsua eta pozoina mundura eman ukhan du, non haren arrazatik heldu garen guziok gure naturaz Jainkoaren hiraren eta punizionearen digne baikara: herioaren eta damnazionearen suiet, eta deabruaren tiraniaren eta botherearen meneko egin izanik. Halaber, liburu noble eta exzelent hetarik ezagutzen dugu, nola Jainkoak lehenago prometatu ukhan zerauen Adami, Abrahami, Isaaki, Jakobi, Dabidi, eta anhitz berze lehenagokori, ezen igorriren luela dohain onezko hazia, Jesus Krist bere Seme gure Salbadorea;

TESTAMENTV ÇAHARRAC
eta berriac iracaſten draucuten guciaren
ſommarioa.

Eſtamẽtu çaharreco liburuẽc iracaſten gaituzte ecen Adamec, Noec, Abrahamec, Iſaac-ec, Iacob-ec, Dauid-ec eta berce gure Aitéc adoratu vkan duten Iaincoa, dela eguiazco Iaincobat bera, bothere gucitaco eta eternal: bere ontaſſun handiagatic cerua eta lurra eta hetan diraden gauça guciac bere Hitzaz creatu dituena, ceinaganic gauça guciac heldu baitirade, eta hura gabe deuſec içateric eztu. Ceinec eguiten baitu iuſtitia eta miſericordia, bayeta goitico gauça guciac gauça gucietan eta leku gucietan, berac placer duen beçala. Eta ezta nehor hain auſartic içateco, non hari derrón, cergatic eguiten duena, hunela edo hala eguiten duen.

Guehiago, liburu hec declaratzẽ draucute, nola Iainco gucizco gorác eta bothere gucitacoac, berce gauçác creatu cituen ondoan, creatu vkan çuen Adam lehen guiçona, bere irudi eta imagina ſpiritualera, hura eguiten çuelaric gauça gucién iabe lurrean. Baina deabruaren inuidiaz eta enganioz deſobedientiatan eror cedin Adam bere Creaçalearen manamenduaren iragaitez eta hauſtez: eta bere bekatuaz halaco kotſua eta poçoina mundura eman vkan du, non haren arraçatic heldu garen gucioc gure naturaz Iaincoaren hiraren eta punitionearen digne baicara: herioaren eta damnationearen ſuiet, eta deabruaren tyranniaren eta botherearen meneco eguin içanic.

Halaber, liburu noble eta excellent hetaric eçagutzen dugu, nola Iaincoac lehenago promettatu vkan cerauẽ Adami, Abrahami, Iſaac-i, Iacob-i, Dauid-i, eta anhitz berce lehenagocori, ecen igorriren luela dohain onezco hacia, Ieſus Chriſt bere Seme gure Saluadorea:

3. ONDORIOAK

- Horrelako corpus bat osatzeko ezinbestekoa da lan automatikoa
- Normalizazio-emaitzek metodologia balidatzen dute
- Estandarrarekiko distantzia handiagoa > ataza zailagoa
- Obra “bereziekin” merezi du esfortzua egiteak eskuzko etiketatzean

4. HURRENGO LANAK

- Eskuzko normalizazioa aurrera: Etxepare, Lazarraga
- XV. eta XVI. mendeak bukatu
  - Betolaza
  - Bildumak: *TAV, Contr* eta *ETZ* > eskuz erabat, seguru asko
- Etiketatze morfosintaktikoa doitu
- Bilaketa-interfazea garatu

Trena martxan dago!

Oharrak

Irudi guztiak baimenarekin erabili dira edo gureak dira.

Lan hau Creative Commons Aitortu-EzKomertziala 4.0 Nazioartekoa lizentzia baten mende dago.

Webgune interesgarriak

- Europeana https://www.europeana.eu/portal/en
- Project Gutenberg https://www.gutenberg.org/
- Eusko Ikaskuntzaren Dokumentazio Zentro digitala http://www.eusko-ikaskuntza.eus/eu/dokumentu-fondoa/
- Behagunea http://behagune.elhuyar.eus/
- Clarin https://www.clarin.eu/
- es-Clarin-K center http://clarin-es.org/
- Ixa-Clarin-K center http://ixa2.si.ehu.es/clarink/index.php?lang=eu

Erreferentziak

Itziar Aduriz, Maria Jesús Aranzabe, Jose Maria Arriola, Arantza Diaz de Ilarraza, Koldo Gojenola, Maite Oronoz, Larraitz Uria (2004). A Cascaded Syntactic Analyser for Basque. In Alexander Gelbukh (arg.), Lecture Notes in Computer Science (LNCS), Computational Linguistics and Intelligent Text Processing, 2945, pg. 124-135. ISBN: 3-540-21006-7 ISSN: 0302-9743, LNCS Series. Springer Verlag. Berlin. 2004.

Manex Agirrezabal, Begoña Altuna, Lara Gil-Vallejo, Josu Goikoetxea, Itziar Gonzalez-Dios (2019). Creating vocabulary exercises through NLP. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, CEUR-WS, vol. 2364, pp. 18-32. ISSN:1613-0073.

Izaskun Aldezabal, Olatz Ansa, Bertol Arrieta, Xabier Artola, Aitzol Ezeiza, Gregorio Hernández, Mikel Lersundi (2001). EDBL: a General Lexical Basis for the Automatic Processing of Basque. IRCS Workshop on linguistic databases. Philadelphia (USA).

Izaskun Etxeberria (2016). Aldaera linguistikoen normalizazioa inferentzia fonologikoa eta morfologikoa erabiliz. Ph.D. thesis, Universidad del Pais Vasco / Euskal Herriko Unibertsitatea.

Josef Novak, Nobuaki Minematsu, Keikichi Hirose (2012). WFST-Based Grapheme-to-Phoneme Conversion: Open Source tools for Alignment, Model-Building and Decoding. In Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, Donostia–San Sebastian. Association for Computational Linguistics.

ESKERRIK ASKO!

GALDERARIK?