Erreferentzia-corpusen (edo eskura ditugun corpusen) aplikazioak

Eneko Agirre
LNPrako Ixa taldea
http://ixa.si.ehu.es

Egitura

- LNPan corpusak duen garrantzia
- Ikerkuntzarako ditugun corpusak
- Corpusen markaketa
- Corpusen markaketa eta erabilera adibideak
- Ondorioak

LNPan corpusak duen garrantzia

- Lengoaia Naturalaren Prozesamendua
  - ✓ hizkuntza "ulertzeko" programak
- Aplikazioak
  - ✓ zuzentzaileak (ortografia, gramatika, estiloa)
  - ✓ elkarrekintza
    - ingurunearekin (ahotsa)
  - ✓ informazioaren kudeaketa
    - bilaketa
    - sailkapena
    - ustiapena
    - laburpena
  - ✓ itzulpengintza
  - ✓ ...

LNPan corpusak duen garrantzia

- Ezagutzan oinarritutako sistemak *(arrazionalistak)*:
  linguistek egindako erregelak
  - ✓ ebaluaziorako arazoak
  - ✓ erregela arraro asko
  - ✓ erregelak ez dira bere horretan betetzen: zurruntasuna
  - ✓ anbiguotasunari aurre egiteko arazoak
- Corpusetan oinarritutako sistemak *(enpirikoak)*:
  linguistek markatutako corpusak
  - ✓ datuetatik erregelak induzitu: ikasketa automatikoa
  - ✓ ebaluazioa
  - ✓ erregela asko ikasteko aukera
  - ✓ malgutasuna
  - ✓ erregela usuenak aukeratu
- Konbinazioa hoberena

LNPan corpusak duen garrantzia

- LNParen azpilanak
  - ✓ lematizazioa
  - ✓ kategoria
  - ✓ morfologia
  - ✓ hitz anitzeko unitateak
  - ✓ postposizioak
  - ✓ analisi sintaktikoa
  - ✓ adiera desanbiguazioa
  - ✓ analisi semantikoa
  - ✓ diskurtsoaren egitura
  - ✓ ...
- Denetan laguntzen dute corpusek

Kontzeptuak

- Corpus orekatua ala corpus berezitua
  - ✓ Orekatuan fenomeno linguistikoen maiztasuna "naturala" da
  - ✓ Berezitua domeinu, genero, fenomeno konkretuak aztertzeko
  - ✓ Orekatua ez daukazunean? Ahal duzuna... baina kontuz
- Corpus gordina edo markatutako corpusa
  - ✓ Inongo markarik gabe
  - ✓ Hainbat informazio
- Corpus elebakarrak eta elebidunak
  - ✓ Elebidun konparagarriak (ad. egunkariak)
  - ✓ Elebidun baliokideak – esaldiak lerrokatuta (ad. aldizkari ofizialak, itzulpen zuzenak, ...)

Ikerkuntzarako ditugun corpusak

- Egungo euskara estandarra lantzen dugu
- **EEBS**ren zati bat (UZEI): 1 Mhitz
- **Egunkaria**: 10 Mhitz urteko (3 urte)
- Aldizkari ofizialak (internet): BAO, GAO, NAO, EHAA
- Liburuak:
  - ✓ klasikoak (internet)
  - ✓ literatura (EIZIEko itzultzaile sortari esker)

Ikerkuntzarako ditugun corpusak

- Interneteko testuak
  - ✓ Bilatzaileetan ez dago hizkuntza ezagutzailerik (eta, edo,dira google-en 40.000 dokumentu, kalitatea?)
  - ✓ Euskarazko bilatzailea (aurki, 3.800 dok, eskuz aukeratutakoak)
- Elebidunak
  - ✓ Egunkariak (2000ko egunkariak gaztelera, catalan, ingelesez)
    - 2 Mhitz (3 hilabete)
  - ✓ Aldizkari ofizialak, liburuak (euskara-gaztelera, euskara-ingelesa)
    - 2 Mhitz (12 liburu)
- Etab.

Corpusen markaketa

- Dokumentuaren metadata:
  - ✓ Egilea, data, generoa, euskalkia, etab.
  - ✓ Domeinua: erlijioa, sukaldaritza, kirolak, etab.
- Dokumentuaren egitura:
  - ✓ Titulu eta azpitituluak, paragrafo, esaldi, zerrenda, etab.
- Informazio lexikoa:
  - ✓ Hitzen kategoria, lema eta egitura morfosintaktikoa
    - ad: *etxekoen alde egin du*
  - ✓ Hitz anitzetako unitateak: lokuzioak, zenbakiak, datak, entitateak (lekuak, pertsonak, ...), etab.
    - ad: *irailaren 18an*
    - ad: *Estatu Batuetako presidentea*

Corpusen markaketa

- Informazio sintaktikoa
  - ✓ Esaldien egitura sintaktikoa
- Informazio semantikoa
  - ✓ Hitzen adierak: *Arte*
  - ✓ Analisi semantikoa: *Altzairuz egin ezazu kupela.*
- Informazio pragmatikoa
  - ✓ Korreferentzia: *Neskatoak altxatu zuen lurretik.*
  - ✓ Diskurtsoaren egitura
- Etab.

Corpusen markaketa eta erabilera

- IXA taldea
- Informazio linguistikoa eskuz gehituta
  - ✓ Domeinuak: 10.000 dokumentu
  - ✓ Morfosintaktikoa: 50.000 hitz EEBS / egunkaria
  - ✓ Izendatutako entitateena: 20.000 entitate (hitz larritan dauden pertsonak, lekuak, erakundeak)
  - ✓ Egitura sintaktikoa: 50.000 hitz EEBS / egunkaria
  - ✓ Adierena: 4.000 hitz
- Automatikoki markatutako corpusen erabilpena
  - ✓ Hitz anitzeko unitate lexikalak lantzeko
  - ✓ Aditzen azpikategorizazio ereduak ikasteko
  - ✓ Itzulpengintza

Adibideak: informazio morfosintaktikoa

- Informazio morfologikoz eskuz etiketatutako corpusa,
  - ✓ EEBS Tamaina: 28.300 token
  - ✓ Egunkaria Tamaina: 14.800 token
- Lema eta kategoriaz gain hitzen egitura morfologikoa
- Erabilera
  - ✓ analizatzaile morfologiko eta lematizatzailea ebaluatzeko
  - ✓ lematizatzaileak desanbiguatzen ikasi dezan
    - eskuzko erregelak
    - erregela estatistikoak

Adibideak: informazio morfosintaktikoa

```
/<Eta>/<HAS_MAI>/
C       ("eta"  LOT JNT EMEN @PJ)
        ("eta"  LOT MEN KAUS @+JADNAG_MP @+JADLAG_MP)
/<,>/<PUNT_KOMA>/
/<azkenik>/
        ("azken"  DET ORD + DEK PAR MG @OBJ @SUBJ)
        ("azken"  IZE ARR + DEK PAR MG @OBJ @SUBJ)
C       ("azkenik"  ADB ADOARR)
/<,>/<PUNT_KOMA>/
/<lurralderik>/
C       ("lurralde"  IZE ARR + DEK PAR MG @OBJ @SUBJ)
/<urrutieneko>/
        ("urruti"  ADJ IZO + DEK GEN NUMP MUGM @IZLG> @<IZLG + DEK NUMS MUGM
             + DEK GEL @IZLG> @<IZLG...)
        ("urruti"  ADJ IZO + DEK GEN NUMP MUGM @IZLG> @<IZLG + DEK NUMS MUGM
             + DEK GEL @IZLG> @<IZLG)
        ("urruti"  ADJ IZO + GRA SUP + DEK NUMS MUGM + DEK GEL @IZLG> @<IZLG
             + DEK ABS MG @OBJ @SUBJ @PRED)
C       ("urruti"  ADJ IZO + GRA SUP + DEK NUMS MUGM + DEK GEL @IZLG> @<IZLG)
/<herriak>/
        ("herri"  IZE ARR + DEK ABS NUMP MUGM @OBJ @SUBJ @PRED)
C       ("herri"  IZE ARR + DEK ERG NUMS MUGM @SUBJ)
```

Adibideak: informazio morfosintaktikoa

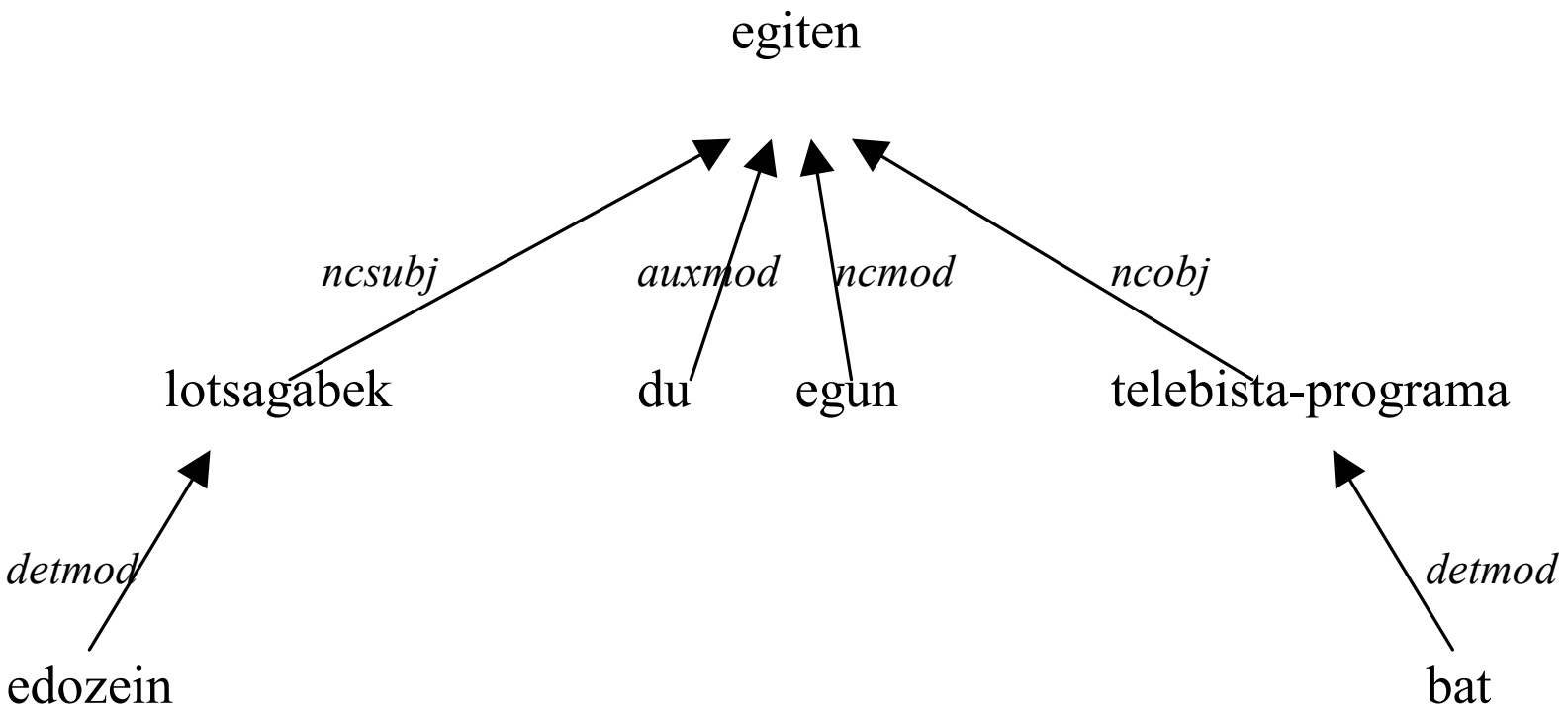
"<Gero>" D:395
  "gero" ADB ADO HAS_MAI @ADLG
"<,>"
  PUNT_KOMA
"<hegoak>" D:223
  "hego" IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ
"<moztu>" D:16
  "motz" ADI SIN ASP PART ZERO NOTDEK @-JADNAG
"<eta>" D:392
  "eta" LOT JNT @PJ @SJ AORG
"<poxpolu>"
  "poxpolu" IZE ARR ZERO @KM
"<kaxa>" D:30
  "kaxa" IZE ARR ZERO AORG @KM
"<batean>" D:164
  "bat" DET DZH DEK NUMS MUGM DEK INE @ADLG
"<gartzelaratuko>" D:187
  "gartzelara" ADI SIN ASP PART ASP ETOR NOTDEK AORG @-JADNAG
"<zizkizun>" D:208
   "*edun" ADL B1 NR_HK NI_ZU NK_HU @+JADLAG
"<$.>" PUNT_PUNT

Adibideak: informazio sintaktikoa

- Informazio sintaktikoz eskuz etiketatutako corpusa
- Jatorria: morfosintaxiaz etiketatutako corpus bera
- Bi eredu:
  - ✓ zuhaitz egitura
  - ✓ dependentzien zuhaitza
- Erabilera:
  - ✓ analizatzaile sintaktikoen ebaluazioa
  - ✓ analizatzaile sintaktikoak ikasteko (txikia)
  - ✓ azpikategorizazioa aztertzeko (txikia)

Adibideak: informazio sintaktikoa

- *Edozein lotsagabek egiten du egun telebista-programa bat.*



Adibideak: informazio sintaktikoa

- *Ikasleek oporrak hartu dituzte, baita irakasleak ere.*

  ncsubj (erg., hartu, ikasleek)

  ncobj (abs., hartu, oporrak)

  auxmod (- , hartu, dituzte)

  lot (baita, e, irakasleak, ere)

  ncsubj (erg., hartu, irakasleak)

Adibideak: hitzen adierak

- Adierak eskuz aukeratu Euskal Hiztegiaren arabera
- 40 hitz (izen, adjektibo, aditz)
- > 100 agerpen bakoitzeko
- Jatorria: egunkaria, EEBS (nahiko agerpen ez)
- Erabilera:
  - ✓ adieren zerrenda fintzeko / luzatzeko
    - Euskarazko hitzen "ontologia"-ren hezurdura: EusWordNet
  - ✓ adieren maiztasunak jakiteko
  - ✓ hitzen adiera topatzen duen sistemak ikas dezan
  - ✓ ebaluazioa (*Senseval-2*)

Adibideak: hitzen adierak

```
<entry>
  <form><orth>koroa</orth></form>
  <GramGrp><pos>iz.</pos></GramGrp>
  <usg type=time>1571</usg>
  <sense n='A1'>
    <def>Eraztun formako apaingarria, buruan ezartzen dena, abarrez, lorez... egina
         edota metalezkoa, <hi rend=italic>berezk.</hi> agintaritzaren ezaugarri dena.</def>
    <xr type = syn><lbl>Ik.</lbl><ref>burestun; buruntza</ref></xr>
    <eg><q>Alkatearen zumezko koroa. Urre eta diamantezko koroa. Elorrizko,
         arantzazko koroa. Erregeren koroa. Koroa irabazi nahi duenak.</q></eg>
  <sense n='A1.N2'>
      <def>Erregetza.</def>
      <eg><q>Espainiako Koroa. Ingeles koroaren mendean.</q></eg>
    </sense>
  </sense>
  <sense n='A2'>
    <def>Zirkulu formako gauzakia.</def>
    <eg><q>Zerraldo gaineko lorezko koroa.</q></eg>
  </sense>
</entry>
```

Adibideak: hitzen adierak

```
<instance id="koroa.IZE.50" docsrc="2000-09-
   23.kirola3.txt" topic="kirola" sentsrc="4" positsrc="2">
<answer instance="koroa.IZE.50" senseid="koroa.A1"/>
<context>
Final gutxi baina izar asko izan ziren atzo olinpiar
   estadioan.
Jokoetako errege-erreginen <head>koroak</head> janztera
   etorri diren atletak - Marion Jones, Maurice Green,
   Cathy Freeman eta Michael Johnson - atzo estreinatu
   ziren Sydneyko Jokoetan, ondo estreinatu ere.
Guztiek erraz egin zuten aurrera euren kanporaketetan, 100
   metroetakoek bi alditan, eta 400ekoak behin.
</context>
</instance>
```

Adibideak: hitzen adierak

- *Koroa*:
  - ✓ Adiera nagusia egunkarian %39 A1.N2 (erregetza)
  - ✓ Adiera berriak: moneta
- *Tentsio*:
  - ✓ Bi adiera:
    - gatazkei lotutakoa
    - elektrizitateari lotutakoa
  - ✓ Adiera nagusia egunkarian, lehenbizikoa %98
  - ✓ Adiera nagusia EEBSn, bigarrena %72
    - 58 agerpen (gure zatian)

Erabilpena: azpikategorizazioa

- Aditzen azpikategorizazio ereduak corpusetatik zuzenean ikasi
- Corpusa ez dago eskuz markatuta
  - ✓ Jatorria: egunkaria, 1.4 Mhitz
  - ✓ Automatikoki lematizatua
  - ✓ Azaleko sintaxia automatikoki egina: IS eta aditz kateak
- Metodoa:
  - ✓ Kasu/aditz konbinazioak kontatu
  - ✓ Maiz gertatzen diren kasuak: argumentuak
  - ✓ Elkarrekin maiz agertzen diren argumentuak: ereduak
- Aditz batzuek egunkarian esanahi berezia

Erabilpena: azpikategorizazioa

```
[ IS [ hori ]
          [ KASUA  Absolutiboa (ABS) ]
          [ NUMEROA Singularra ]
          [ FUNTZIO_SINTAKTIKOA    SUBJ ]
          [ GUNEA hori ]
]

[ ADIKAT [ isladatzen  da ]
          [ ASPEKTUA Ez burutua (EZBU) ]
          [ MODUA-DENBORA  Indikatibozko orainaldia (A1) ]
          [ FUNTZIO_SINTAKTIKOAK
               [ ADITZ NAGUSIAN    JADNAG (Ez jokatua) ]
               [ ADITZ LAGUNTZAILEAN    JADLAG (Jokatua) ] ]
          [ PERTSONAK
             [ NOR  Hura ]]
          [ GUNEA islatu ]
]

[ IS [ Mediterranioko  zonaldean ]
           [ KASUA  Inesiboa (INE) ]
           [ NUMEROA Singularra ]
           [ MUGATASUNA Mugatua ]
           [ FUNTZIO_SINTAKTIKOA    ADLG ]
           [ GUNEA zonalde ]
 ]
```

Erabilpena: HAUL erauzketa

- Corpus orekatu handia behar da
- Ez da beharrezkoa markatuta egotea
- Metodoa:
  - ✓ hitzak elkarrekin ausaz espero zitekeena baino gehiago azaltzen badira: HAULa izan daiteke
  - ✓ hitzen agerpenak kontatu
  - ✓ hitz bikoteen agerpenak kontatu
    - martxan jarri BAI
    - etxe txikia EZ
- Terminoak bereizteko, corpus berezitu bateko agerpen kopuruarekin konparatu

Erabilpena: Itzulpengintza

- Corpus elebidun baliokideak, markatu gabe
- Esaldiak automatikoki lerrokatuta
- Erabilera:
  - ✓ Ordenadoreak lagundutako itzulpena
  - ✓ Itzulpen automatikoa
  - ✓ Itzulpen memoriak eraikitzeko
  - ✓ Hizkuntza bateko ezagutza linguistikoa bestera pasatzeko
- Corpus elebidun baliokide orekaturik ez dago munduan
- Corpus elebidun konparagarria

Erabilpena: Itzulpengintza

```
<p id=.p1>
<s id=.s1 corresp=.s1> Euskal Herriko
Agintaritzaren Aldizkaria - EHAA 1999189
Aldizkaria 1999/10/01 Orrialdea: 16367
Bestelako Xedapenak Hezkuntza, Unibertsitate
eta Ikerketa Saila AGINDUA,
1999ko irailaren 13koa, Hezkuntza,
Unibertsitate eta Ikerketa sailburuarena,
1999ko Euskadi Ikerketa Sariko Epaimahaiaren
erabakia jakinarazteko dena. </s> </p>
<p id=.p1>
<s id=.s2 corresp=.s2> Maiatzaren  7ko
93/1996  Dekretuak  arautzen  du  Euskadi
Ikerketa Saria. </s> </p>
...
<s id=.s5 corresp='.s5 .s6'>  Epaimahaia
1999ko  irailaren  11n  bildu  zelarik,
honakoa EBATZI DUT:
</s> </p>
```

```
<p id=.p1>
<s id=.s1 corresp=.s1> Boletín Oficial del
País Vasco - BOPV Boletín N. 1999189
01/10/1999 Pág: 16367 Otras Disposiciones
Educación, Universidades e Investigación ORDEN
de 13 de septiembre de 1999, del Consejero de
Educación, Universidades e Investigación, por
la que se hace público el fallo del Jurado del
Premio Euskadi de Investigación para 1999.
</s> </p>
<p id=.p1>
<s id=.s2 corresp=.s2> El Decreto 93/1996  de
7  de  mayo, regula  el  Premio  Euskadi  de
Investigación. </s> </p>
...
<s id=.s5 corresp=.s5> Habiéndose  reunido
dicho Jurado el día 11 de septiembre de 1999.
</s> </p>
<p id=.p4>
<s id=.s6 corresp=.s5> RESUELVO: </s> </p>
```

Ondorioak

- Tamaina garrantzitsua da:
  - ✓ Neurri estatistikoak fidagarriak izateko
  - ✓ Fenomeno linguistiko ez hain usuak (gehiengoak!)
    - adierak, azpikategorizazio ereduak
- Orekatua izatea bai da garrantzitsua:
  - ✓ Adieren maiztasunak errealak izan daitezen
  - ✓ Azpikategorizazio ereduak errealak izan daitezen
- Corpusa markatuta egon behar da:
  - ✓ Egitura, generoa, domeinua, kategoria, lema
- Fenomeno linguistiko jakinez markatutako corpus orokorrak:
  - ✓ Maila lexiko, sintaktiko, semantiko eta pragmatikoan
  - ✓ Lan handia: ad. 40 hitzen 100 agerpen, 120 ordu.
- Publikoa

Etorkizunean nahi duguna 2010

Beste hizkuntzetarako dagoena:

• Markaketa minimoa:	5 Mh (EEBS)	100 Mh
• Hitzen egitura morfosintaktikoa	50 Kh	2 Mh
• Sintaxia	300 h	1 Mh
• Semantika		
✓ adierak	4000 h	1 Mh
✓ analisi semantikoa	0 h	1 Mh

- Corpus elebidunak
- Ahozko corpusak