Laburpena

Lan honen helburua, euskararako sintaktikoki etiketatutako corpusa edo *treebank*-a eraikitzeko oinarriak ezartzea izan da. Aukera desberdinak aztertu ondoren, dependentzia erlazioetan oinarritutako formalismoa jarraitu dugu, 1988an Carrol *et al.*-ek aurkeztutako etiketatze-eskema, hain zuzen ere. Etiketa horiek hitzen arteko interdependentziak erakusten dituzte, esplizituki erakutsi ere, dependentzia-egiturak osatuz. Ondoren, eskema horren aplikazioa egin da euskarazko corpus batean. Aztertutako corpus hau, eskuz etiketaturiko hasierako testu-corpusaren emaitza da.
Mota honetako corpusak baliagarriak dira hizkuntzari buruzko ikerketa linguistikoak egiteko zein tresna informatikoak garatzeko eta ebaluatzeko.

Abstract

The goal of this research is to settle the basis for building the treebank for Basque, i.e., a syntactically tagged corpus. After considering and analyzing the various formalism models, we have selected the Dependency-based model, which follows the scheme presented in (Carrol *et al.*, 1998). It is based on the idea of adding to each sentence in the corpus a series of grammatical relations specifyng the dependencies between modifiers and their nucleus. The chosen tag-set has been applied to a Basque sample corpus.
To conclude, we would like to stress the urging necessity of a syntactically tagged corpus, which would serve to do linguistic research on the language and to develop computational tools and evaluate existing ones.

AURKIBIDEA:

1. AURKEZPEN OROKORRA 3
1.1 Motibazioa 3
1.2 Helburua 3
1.3 Ikerlanaren eskema 4
2. SARRERA 4
3. CORPUS ETA *TREEBANK*-EN INGURUAN 8
3.1 Hizkuntza-corpusak 8
3.2 Hizkuntza arrotzetako treebank-ak 10
3.3 Formalismo nagusiak 11
4. OINARRI METODOLOGIKOAK 13
5. EUSKARARAKO *TREEBANK*-A 14
5.1 Mendekotasun-egituran oinarritutako markaketa sintaktikoa 14
5.2 Euskarazko corpusa 16
5.3 Etiketek adierazten dituzten erlazio gramatikalen deskribapena 17
5.4 Eskuz sintaktikoki aztertutako adibideak 22
5.5 Etiketatzerakoan sortutako arazoak eta irtenbideak 26
6. ESKUZKO MARKAKETAN LAGUNTZEKO TRESNAK 30
7. ONDORIOAK 30
GLOSATEGIA 32
ERREFERENTZIAK 34

1. AURKEZPEN OROKORRA

1.1 Motibazioa

Ikerlan honen zeregin nagusia euskararen sintaxi konputazionalaren azterketan aurrera egitea da, hau da, oraindaino egin dena hobetzea eta sakontzea. Horretarako, euskararen sintaxia sakonago lantzeko, alegia, nahitaezkoa dugu sintaktikoki etiketaturiko corpusa *(treebank)* eraikitzea.

Corpus horretan zehaztuko den informazio sintaktikoa oso lagungarria izango da hizkuntzari buruzko ikerketa linguistikoak egiteko zein tresna informatikoak garatzeko eta ebaluatzeko.

Gertaera linguistiko horien deskribapen zabala egin ahal izateko corpusak, hau da, analisiaren abiapuntua den eta aztergaia osatuko duen enuntziatu multzoak, librea eta orokorra izateaz gain, tamaina handikoa izan behar du. Hain zuzen ere, corpusak ugaltzen diren neurrian, berauetan oinarritutako ikerketek sakontasunean eta zehaztasunean irabazten dutelako.

Jakina, ezaugarri hauetako corpusa eraikitzea ez da nolanahiko lana. Nahiz eta ataza neketsua eta garestia izan, beharrezkoa da; sintaxi-tresnak sortzeko eta aplikazio berrietarako ezinbestekotzat jotzen baita Lengoaia Naturalaren Prozesamenduaren (LNP) arloan. Aipatu aplikazio hauek corpusetan oinarritutako gramatika konputazionalak lortzea izango dute helburu (Charniak E., 1996), (Kübler S. & Hinrichs E.W., 2001), eta gehiago landu beharko diren prozesuen abiapuntu izango dira. Honek guztiak Itzulpen Automatikorako sistemak, Informazio Erauzketa, Informazioaren Berreskurapena, Laburpen Automatikoa eta Galdera-Erantzun motako sistemak hobetzen lagunduko du.

Hizkuntzalaritza mailan, *treebank*-a ezinbesteko datu-basea da hizkuntzaren azterketan, hizkuntza errealari dagozkion aztertutako/etiketatutako adibideak eskaintzen dituelako. Azterketa linguistiko honek eragin zuzena du aipaturiko errekurtso horien hobekuntzan, sendotasun handiagoa ematen dielako.

1.2 Helburua

Lantzen ari garen ikerlan honen helburua, IXA[1] taldean aurretik aztertutako azaleko sintaxia abiapuntu gisa hartuta, *treebank*-a eraikitzeko oinarriak ezartzea da.

[1]Euskal Herriko Unibertsitateko Informatika Fakultateko IXA taldeak Lengoaia Naturalaren Prozesamenduan eginiko ikerketa-lana du helburu nagusi. Arlo zabal horren barruan euskararen gaineko ikerketa aplikatua da bere xede nagusia.
Azken helburua euskararako hain beharrezko eta onuragarri izan daitezkeen sistema (erdi)automatikoak egitea edo egiten laguntzea, teknologia berrien eragina hizkuntzan ahalik eta onena izan dadin.
Lan-talde honi buruzko informazioa ondoko web-orrian aurki daiteke: http://ixa.si.ehu.es

Helburu horretara iristeko ondorengo puntuetan zerrendatzen direnak dira eginiko lanak:

- *Treebank*-a osatzeko etiketatze-eskema orokorra definitu; zeregin horretan oinarrizko erabaki batzuk hartu dira.
- Etiketatze hori garatzeko erabili behar den formalismoa hautatu: erabakiak formalizatzeko eredu desberdinak aztertu ondoren, mendekotasun-egituran oinarritzen dena aukeratu dugu.
- Eskemaren aplikazioa euskarazko corpus batean: lagin hau, eskuz etiketaturiko hasierako testu-corpusaren emaitza da. Etiketa sintaktiko horiek hitzen arteko interdependentziak erakutsiko dituzte, esplizituki erakutsi ere, dependentzia-egiturak osatuz.

Honetatik guztitik lortzen den informazioa gure taldean landutako analizatzaile sintaktikoak garatzeko eta sendotzeko erabili nahi dugu; hartara, tresna horiek ahalmen handiagoa izango dutelako analisi sakonagoak egiteko. Hala ere, hasieran aipatu bezala, ikerketa teorikoetarako zein gainerako aplikazioetarako ere baliagarria izan daiteke. Beraz, lan hau diziplinartekoa da, informatikarien eta hizkuntzalarien jakintza-arloak elkarren osagarri baitira, lan egiteko ikuspegiak zabalduz eta ezagutza aberastuz.

1.3 Ikerlanaren eskema

Motibazioak eta helburu nagusiak azalduta, ikerlanean jarraitu dugun eskema egin dugu atal honetan. Horrela, 2. atalean eta sarrera modura, esku artean dugun lana LNPren barruan kokatu eta LNPrako *treebank*-a sortzeko premia aztertu dugu. 3.ean, corpusek hizkuntza-ikerketetan duten garrantzia azaldu dugu batetik, eta bestetik beste hizkuntzetarako garatu diren *treebank* nagusiak aipatu ditugu, tradizionalki erabili diren bi formalismo desberdinen azterketa zehatzagoa eginez. 4.ean euskararako *treebank*-a eraikitzeko oinarri metodologikoak ezarri ditugu. 5. atalean euskararako *treebank*-a deskribatu dugu, euskarazko corpus errealaren gainean hartu behar izan ditugun erabakiak arrazoituz eta erabaki horiek formalismo batez adieraziz. 6.ean, hizkuntzalariari lana erraztearren eta lanaren beraren fidagarritasuna bermatzearren paraleloan egiten ari garen lana aipatu besterik ez dugu egin. Eta bukatzeko ikerlan honetatik ateratako ondorioak aipatu ditugu.

2. SARRERA

Sintaxia funtsezkoa dugu hizkuntzaren tratamenduaren arloko edozein lani ekiteko, hizkuntza ezagutzea nahiz sortzea delarik sintaxiaren edo hizkuntzaren tratamenduaren arloko edozein lanaren helburua. Hizkuntzaren gramatika formalizatu eta konputazionalki tratatzeko moduan adierazi behar da, morfologiaz harantzago joan nahi duen edozein aplikazio edo tresnatan erabiliko bada.

Morfologiatik sintaxira doan bidea, ordea, oso luzea gertatzen da LNPren munduan; morfologia-sistema osoak eraikitzea posible den bitartean, ez dago, oraindik, sistema sintaktiko automatiko oso bat garatuta; are gutxiago ingelesa bezala ikertu ez den hizkuntza baterako. Hori dela-eta tarteko bideak hartu dira eta analizatzaile sintaktiko orokorrak baino sinpleagoak diren tresnak bultzatu dira azken urteetan. Arrakastatsuenak etiketatzaileak izan dira eta, haiekin batera lematizatzaileak. Etiketatzaileek testuko hitz bakoitzak dituen analisi desberdinen artean zuzena dena aukeratu behar dute; lematizatzaileek, aldiz, lema posibleen artean dagokiona.

Ingeleserako sortu da ahalmen zabaleko sistemarik, baina sistema horiek ere oztopo itzel batekin aurkitzen dira beren emaitzak aplikatu nahi dituztenean: emaitza posible bakarra ez, baizik eta dozenaka-edo analisi posible lortzen baitituzte testu libreetako perpaus arruntak analizatzerakoan.

Analisi sintaktiko konputazionalaren inguruan kontuan hartu beharrekoa da laborategiko esaldi-multzo bat prozesatuko duen analizatzailea eraikitzea erraza bada ere, oro har, hizkuntzaren teoria hutsetan oinarritu direnak ez direla erabilgarriak izan testu libreetan aplikatzeko. Testu errealak behar dira, egindako programa horiek gero benetako testuei aplikatzerakoan porrot egingo ez badute. LNPko sistema gehienek testu errealetan aplikatzeko helburua dute. Kasu guztietan, aurre egin beharreko arazoa anbiguotasunarena dugu, esan bezala, arazoa areagotu egiten baita aplikazio errealekin lan egin behar denean.

Analisi zein sorkuntza sintaktikoak ezinbesteko tresnak ditugu aplikazio gehienetan. Horrela bada, eta euskarari dagokionean beste hizkuntza batzuetarako baliagarri suertatu diren formalismoak eta analisi-teknikak erabiltzen ari gara gu ere. Horien artean Murrizpen Gramatika (Karlsson *et al.*, 1995) azpimarratu nahi genuke, analisi morfologikotik ateratzen den anbiguotasun maila jaisteko eta esaldien analisi sintaktiko azalekoa egiteko erabiltzen ari garena. Horretaz gain, PATR-II izeneko baterakuntza-formalismoaz (Shieber M., 1986) euskarazko izen-sintagma eta perpaus bakunen egitura deskribatzen duen gramatika konputazionala ere garatu dugu.

Beraz, euskararen analizatzaile sintaktikoa eraikitzeko taldean garatu diren tresna hauetan oinarrituko gara:

a) analizatzaile morfologikoan (Alegria *et al.*, 1996): analizatzaile (eta sintetizatzaile) morfologikoaren zeregina hitz-forma osatzen duten morfemak ezagutzea (eta konposatzea) da, eta morfema bakoitzari dagokion informazio morfologiko-lexikala ematea. I.1 irudian *"Gero hegoak moztu eta poxpolu kaxa batean gartzelaratuko zizkizun"* esaldiaren analisia daukagu. Ikus daitekeenez, hitz bakoitzeko analisi posible bat baino gehiago ematen da.

```
        "<$.>"
          PUNT_PUNT
        "<Gero>"
          "gero"  ADB ADOARR
          "gero"  IZE ARR + DEK ABS MG @OBJ @SUBJ @PRED
          "gero"  IZE ARR
        "<,>"
          PUNT_KOMA
        "<hegoak>"
          "hego"  IZE ARR + DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
          "hego"  IZE ARR + DEK ERG NUMS MUGM @SUBJ
        "<moztu>"
          "moztu"  ADI SIN + AMM PART + ASP BURU
          "moztu"  ADI SIN + AMM PART + DEK ABS MG @OBJ @SUBJ @PRED
          "moztu" ADI SIN + AMM PART
        "<eta>"
          "eta"  LOT JNT EMEN @PJ
          "eta"  LOT MEN KAUS @+JADNAG_MP @+JADLAG_MP
        "<poxpolu>"
          "pospolo"  /poxpolu/ IZE ARR + DEK ABS MG @OBJ @SUBJ @PRED
          "pospolo"  /poxpolu/IZE ARR
        "<kaxa>"
          "kaxa"  IZE ARR + DEK ABS MG @OBJ @SUBJ @PRED
          "kaxa"  IZE ARR + DEK ABS NUMS MUGM @OBJ @SUBJ @PRED
          "kaxa"  IZE ARR
        "<batean>"
          "bat"  DET DZH NUMS + DEK NUMS MUGM + DEK INE @ADLG
          "bat"  IZE ARR RARE+ + DEK NUMS MUGM + DEK INE @ADLG
          "bate"  IZE ARR RARE+ + DEK NUMS MUGM + DEK INE @ADLG
          "batean" ADB ALGARR
        "<gartzelaratuko>"
          "kartzelaratu"  /gartzelaratu/ ADI SIN + AMM PART + ASP GERO
          "kartzelaratu"  /gartzelaratu/ ADI SIN + AMM PART + DEK NUMS
        MUGM + DEK GEL @IZLG> @<IZLG + DEK ABS MG @OBJ @SUBJ @PRED
          "kartzelaratu"  /gartzelaratu/ ADI SIN + AMM PART + DEK NUMS
        MUGM + DEK GEL @IZLG> @<IZLG
        "<zizkizun>"
          "*edun"  ADL B1 NR_HK NI_ZU NK_HU + ERL MEN ERLT
        @+JADNAG_IZLG> @+JADLAG_IZLG<
          "*edun"  ADL B1 NR_HK NI_ZU NK_HU + ERL MEN MOS @+JADNAG_MP
        @+JADLAG_MP
          "*edun"  ADL B1 NR_HK NI_ZU NK_HU + ERL MEN ZHG
        @+JADNAG_MP_OBJ @+JADLAG_MP_OBJ
          "*edun"  ADL B1 NR_HK NI_ZU NK_HU
        "<$.>"
          PUNT_PUNT
```

I.1 irudia. "*Gero, hegoak moztu eta poxpolu kaxa batean gartzelaratuko zizkizun"* esaldiaren analisia.

b) desanbiguatzaile morfologikoan (Aduriz *et al.*, 1996): analisi morfologikotik irtendako emaitza anbiguoa tratatzen da. Murrizpen Gramatikako (MG) erregela batzuk aplikatu eta gero (395, 223, 16, 392, 30, 164, 208), hitz bakoitzeko analisi bakarra uzten saiatzen gara. (Ikus I.2 irudia)

```
"<$.>"
  PUNT_PUNT
"<Gero>" D:395
  "gero"  ADB ADOARR  @ADLG
"<,>"
  PUNT_KOMA
"<hegoak>" D:223
  "hego"  IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"<moztu>" D:16
  "moztu"  ADI SIN AMM PART  @-JADNAG
"<eta>" D:392
  "eta"  LOT JNT EMEN @PJ
"<poxpolu>"
  "pospolo"  /poxpolu/ IZE ARR DEK ABS MG @OBJ @SUBJ @PRED
  "pospolo"  /poxpolu/ IZE ARR  @KM>
"<kaxa>" D:30
  "kaxa"  IZE ARR  @KM>
"<batean>" D:164
  "bat"  DET DZH NUMS DEK NUMS MUGM DEK INE @ADLG
"<gartzelaratuko>" D:187
  "kartzelaratu"  /gartzelaratu/ ADI SIN AMM PART ASP GERO
@-JADNAG
"<zizkizun>" D:208
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU ERL MEN MOS @+JADNAG_MP
@+JADLAG_MP
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU ERL MEN ZHG @+JADNAG_MP_OBJ
@+JADLAG_MP_OBJ
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU @+JADLAG
"<$.>"
  PUNT_PUNT
```

I.2 irudia: I.1 irudiko esaldiaren analisi desanbiguatua

c) azaleko analizatzaile sintaktikoan (Arriola *et al.*, 1999): perpausaren azaleko analisia[2] lortzen da, hau da, zein diren osagai posibleak eta beren arteko loturak. Orain arte xedea ez da esaldi oso-osoa analizatzea izan.

[2] azaleko analisia: testuan ageri diren elementuak bakarrik hartzen dituena kontuan.

```
"<$.>"
  PUNT_PUNT
"<Gero>" %SINT
  "gero"  ADB ADOARR  @ADLG
"<,>"
  PUNT_KOMA
"<hegoak>" %SINT
  "hego"  IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @PRED
"<moztu>" %ADIKAT
  "moztu"  ADI SIN AMM PART  @-JADNAG
"<eta>"
  "eta"  LOT JNT EMEN @PJ
"<poxpolu>" %SIH
  "pospolo"  /poxpolu/ IZE ARR DEK ABS MG @OBJ @SUBJ @PRED
  "pospolo"  /poxpolu/ IZE ARR  @KM>
"<kaxa>"
  "kaxa"  IZE ARR  @KM>
"<batean>" %SIB
  "bat"  DET DZH NUMS DEK NUMS MUGM DEK INE @ADLG
"<gartzelaratuko>" %ADIKATHAS
  "kartzelaratu"  /gartzelaratu/ ADI SIN AMM PART ASP GERO
@-JADNAG
"<zizkizun>" %ADIKATBU
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU @+JADLAG
"<$.>"
  PUNT_PUNT
```

I.3 irudia. Zatiak ezagutu ahal izateko etiketak (aurretik % ikurra dutenak)[3] dituen adibidea.

Osatzen jarraituko dugun analizatzaile sintaktiko edo *parser* hau aurrez etiketaturiko corpus horrekin ebaluatua izango da eta bere emaitzak testu errealen azaleko zein sakoneko sintaxia deskribatzea eta konputazionalki tratatzea izango dira.

Azken finean helburua hau da: inplizitu dauden harreman sintaktikoak esplizitu egitea. Arriolaren (2000), Gojenolaren (2000) eta Adurizen (2000) tesietan ildo honetatik egindako lana agertzen da eta aurrerantzean ere bide beretik jarraitu beharko da lanean.

3. CORPUS ETA *TREEBANK*-EN INGURUAN

3.1 Hizkuntza-corpusak

Lanean buru-belarri sartu baino lehen corpusa zer den eta hauek hizkuntzaren ikerketan izan dezaketen papera aztertuko dugu.

[3] Azaleko analisietan azaltzen diren zatiak ezagutzeko etiketen esanahia: %ADIKATHAS: osagai bat baino gehiagoko aditz-kate bateko lehenengo elementuari esleitzen diogun etiketa; %ADIKATBU: osagai bat baino gehiagoko aditz-kate bateko azken elementuari esleitzen diogun etiketa; %ADIKAT: elementu batez osaturiko aditz-katea; %ADIKATETEN: aditz-kate ezjarrai baten bigarren osagaia; %ADIKATETENBU: aditz-kate ezjarrai baten azken elementua; %SIH: sintagma-hasiera; %SIB: sintagma-bukaera eta %SINT: hitz bakarreko sintagma.

Testu-corpusak testu-masa handiak dira, informazio linguistikoaren iturri nagusietako bat eta gorago aipatutako aplikazio eta oinarrietarako probaleku ezinbestekoak.

Hizkuntzaren azterketan corpusek duten garrantzia datu enpirikoek duten berbera da. Kontzeptu hau hizkuntzalaritza enpirikoari lotuta azaldu zaigu XX. mendearen hasieran. Berez, hizkuntza da hizkuntzalariak deskribatu nahi duena, baina hori egiteko enuntziatu linguistikoetara jo behar du, hau da, hizkuntzaren erabilera gauzatzen duten ekintzetara, bai ahozko hizketara, baita idatzizko testuetara ere. Bestela esanda, hizkuntzalariak ere haren teoriak babesten dituen eta hizkuntzaren joera nagusiak erakusten dizkion erreferentzia-elementuak (datu enpirikoak) beharko ditu, eta corpusek erreferentzia hori sistematizatuko duten testu edo hizketa-multzoa osatzen dute. Azken batean datu enpiriko hauen bitartez, hizkuntzalariek adierazpen objektiboak egin ditzakete, subjektiboetara mugatu gabe.

Hasierako hizkuntzalaritza sortzaileak (Chomsky, 1957), ordea, hizkuntza batek izan ditzakeen enuntziatuak ezin zenbatuzkoak direla dio, eta hauen ustez, ez dago hizkuntzaren mekanismoak azalduko dituen datu egokien testu multzo (corpus) finiturik. Deskribatu behar den objektuaren adibidea bere hizkuntza hitz egiteko konpetentzia duen hiztun ideal batengan bilatu beharko dela diote. Korronte honen ondorioz, hizkuntzalaritza ikuspegi enpirista batetik ikuspegi arrazionalista batera igarotzen da. Orientazio berri honek berekin dakarren kritika zera da, corpusek ez dutela balio hizkuntza deskribatzeko.

Aurrerago, eta batez ere hizkuntzalaritza aplikatua egiten hasi zenetik, corpusen helburua ez da hizkuntzaren ikuspegi osoa ematea, corpusen helburu berria hizkuntzalaritza inguruko ikerkuntzaren oinarri izango den lagin eredugarri bat izatea da, bertan baitaude, hain zuzen ere, datu objektiboak. Corpusa ezin da hizkuntzarekin parekatu; besterik gabe, ezaugarri egokiak edo ez hain egokiak izango dituen datu-multzoa izango da. Helburu berri honen harira hizkuntzalaritza enpiriko eta, berarekin batera, corpusen gaineko interesa handitu egin da berriz.

Hizkuntzalaritza arrazionalista eta enpirikoaren arteko eztabaidak bere horretan dirauen arren, ikerlan honetan ez dugu horretan sakondu. Kontuan hartu behar duguna da corpusen erabilgarritasuna ez dela egun zalantzan jartzen, areago LNPren arloan.

Hizkuntza bat konputazionalki tratatu ahal izateko, argi dago testu etiketatu erraldoi baten beharra dagoela. Etiketatze hau dena den, maila desberdinetan egin daiteke. Etiketaturiko corpusak informazio asko eta desberdinarekin aberastu izan ohi dira: morfosintaktikoa, sintaktikoa edo semantikoa. Etiketatze-motak:

- Morfologikoa, hitz bakoitzak bere deskribapen morfologikoa du.
- Sintaktikoa, hitzak sintagma eta esaldietan biltzen dituena.
- Semantikoa, hitzei ezaugarri semantikoa eransten diena, hau da, bakoitzak duen esanahia.

Euskararen kasuan, IXA taldean aspaldi ekin genion lan honi eta morfologia mailako markaketa nahikoa aurreratu bada ere, maila sintaktikoan lan sakonagoaren beharra dugu zabalduenak diren hizkuntzen (ingelesa, alemana, ...) ildotik jarraitu nahi izanez gero. Hauetan indar handiak jarri eta jartzen dira horrelako errekurtsoak lortzeko. Ondorioz, tresna eta errekurtso horiek edukitzeak hizkuntzaren tratamendu konputazionalean garapen handiagoa izaten laguntzen die. Hortik gure interesa sintaktikoki etiketatutako corpusa eraikitzeko.

3.2 Hizkuntza arrotzetako *treebank*-ak

Treebank-ak (zuhaitz-banku izenez ere ezagutzen dira) hizkuntza baliabideak dira eta hizkuntza naturaleko egiturak etiketatzea da beren eginkizuna. Esan dugun bezala, etiketatze hau maila desberdinetan egiten da: hitzaren mailan, esaldiarenean, perpausarenean eta zenbaitetan baita funtzio sintaktikoen mailan ere.

Corpus horiek ingeleserako landu dira gehien bat, *Penn Treebank* (Marcus *et al.*,1993), *British National Corpus* (Burnage *et al.*, 1993), *Susanne* (Sampson, 1995), esaterako. Azken urteetan, ordea, lengoaiari aplikatutako teknologia berrien gorakada dela-eta, zuhaitz-banku sintaktiko eta semantikoen datu-basea osatzea helburu duten proiektu ugari garatzen hasi dira hizkuntza desberdinetan, adibidez, frantseserako (Abeillé *et al.*, argitaratzear), alemanerako (Brants *et al.*, argitaratzear), italierarako (Bosco *et al.*, 2000) eta (Montemagni *et al.*, argitaratzear), turkierarako (Oflazer *et al.*, 1999b), polonierarako (Marciniak *et al.*, argitaratzear)[4], gaztelaniarako (Moreno *et al.*, 2000) edo txekierarako (Böhmova *et al.*, 1999).
Hona hemen aipaturiko proiektu horiek, bertan landutako hizkuntza eta tamaina zehazten delarik:

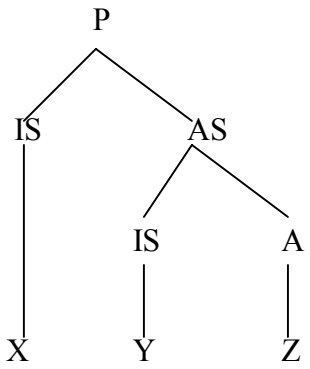
1. NEGRA/TIGER (alemana; 350.000 token[5])
2. PDT: Prague Dependency Treebank (txekiera; 450.000 token)
3. French Treebank (frantsesa; 1.000.000 token)
4. TUT: Turin University Treebank (italiera; 1.000 esaldi)
5. Spanish Treebank (UAM) (gaztelania; 1.500 esaldi)
6. ISST: Italian Syntactic-Semantic Treebank (italiera; 300.000 token)
7. Penn Treebank (ingelesa; 7 milioi/2milioi hitz)
8. Susanne Corpus (ingelesa; 120.000 token)

Euskarari dagokionean, Madrileko Zientzi eta Teknologia Ministeritzak babesturik estatu mailan garatzen ari den *" IXA taldea, Euskararen tratamendu automatikorako tresnak: arbola sintaktiko-semantikoz osatutako datu-base baten sorkuntza"* proiektuan[6] parte hartu du IXA taldeak.

[4] argitaratzear dauden corpus hauei buruzko informazioa web-orri honetan dago ikusgai: http://treebank.linguist.jussieu.fr/toc.html
[5] *token*: testu-unitatea.
[6]PROFIT proiektua. Erreferentzia: FIT-150500-2002-244. Bertan parte hartu duten unibertsitateak, Universitat d´Alacant (UA), Universitat Politècnica de València (UPV), Universitat de Barcelona (UB), Universidad Politécnica de Catalunya (UPC) eta Euskal Herriko Unibertsitatea (UPV/EHU) dira.

3.3 Formalismo nagusiak

Esaldi bat sintaktikoki analizatzen denean, egituraren bat esleitzen zaio. Demagun esaldi batean osagai gerta daitezkeen elementuen segida hau dugula: *" X Y Z"*; hau da, aditza: *Z,* eta beste osagaiak: *X Y*



Egitura horrek esaldiko osagai linguistikoak errepresentatzen ditu eta beraien arteko harreman gramatikalak azalarazten ditu.

Ordea, esaldi-mota bakoitzeko zer-nolako analisia emango den zehaztea ez da lan makala. Guk hizkuntza desberdinetarako erabili diren etiketatze sintaktikoak aztertu ditugu eta *" X Y Z "* motako esaldia, esaterako, era desberdinean etiketa daitekeela ikusi dugu.
Horrela, erabili diren etiketatze sintaktikoen artean bi dira nagusitzen diren formalismoak, bata osagai-egituran (constituency-based) oinarritzen dena eta etiketatze sintaktiko hau jarraituz, aurreko esaldiaren analisia da:

(P (IS X) (AS (IS Y) (A Z)))

Eta bestea, mendekotasun-egituran (dependency-based) oinarritzen dena:

ncsubj[7] (- , Z, X)
ncobj[8] (- , Z, Y)

Jarraian bi formalismo hauek banan-banan aztertuko ditugu.

3.3.1 Osagai-egituran (constituency-based) oinarritzen dena.

Osagai-egituraren bidezko analisian, osagai sintaktikoa osatzen duten osagai bakoitza dago etiketatuta, baita bere kategoria sintaktikoa ere; hau da, emaitza, lortutako osagaiak eta beren kategoriak definituz ematen da (izen-sintagmak, esaldiak, ...)

[7] ncsubj: "*non-clausal subject*"
[8] ncobj: "*non-clausal object*"

Ingeleseko corpus zabalena eta gehien erabilia den *Penn Treebank*-ak (Marcus *et al.*, 1993) etiketatze modu hau jarraitzen du.

Adibidez, ondoko esaldiak honako errepresentazioa du:

(1). *John tried to open the window*[9] (Jon leihoa irekitzen saiatu zen)

```
(S  (NP  (N1  (N  John_NP1)))
  (VP  (V  tried_VVD)
    (VP  (V  to_TO)
      (VP  (V  open_VVO)
        (NP  (DT  the_AT)
          (N1  (N window_NN1) ) ) ) ) ) )
```

Metodo honek hiru ezaugarri nagusi ditu:

1. hurrenkera linealean oinarritzen da; hau da, osagai sintaktikoak esaldian ageri diren ordena berean adierazten dira.
2. Informazio hierarkikoa esplizitu geratzen da.
3. Inplizitu dagoen funtzio informazioak ez du garrantzirik.

3.3.2 Mendekotasun-egituran (dependency-based) oinarritzen dena.

Mendekotasun egituraren kasuan (Järvinen & Tapanainen, 1997), berriz, osagaien arteko erlazioak deskribatzen dira.
Etiketatze eskema hau alemaneko (*NEGRA)* eta txekierako (*PDT)* corpusetan erabili izan da, besteak beste.

Adibidez, (1) esaldiaren errepresentazioa, hau da:

Metodo honek dituen ezaugarriak:

1. hurrenkera linealak garrantzi txikiagoa du.
2. Hierarkian oso oinarrituta dagoen metodoa da.
3. Funtzio informazioak oso garrantzi handia du.

Bi mutur hauen artean tarteko bideak ere erabiltzen dira, adibidez, (Basili *et al.*, 2000) artikuluan mendekotasun-egitura erabiltzen da esaldiaren oinarrizko osagaiak konbinatzeko (izen-sintagmak, preposizio-sintagmak eta aditza), baina horien barruan osagai-egitura lortzen da, mendekotasuna hitzetaraino eraman gabe.

[9] Carroll, J., Briscoe, T., & Sanfilippo, A. (1998). "Parser Evaluation: a survey and a new proposal" artikulutik hartutako adibidea.

Deskribatu diren bi formalismo horiek egokiak badira ere, bakoitzaren arrakasta eta aplikazioaren eraginkortasuna hizkuntzaren araberakoa da. Horrela bada, eta euskararen beraren ezaugarriak kontuan hartuta, mendekotasun-egituran oinarritzen den formalismoa jarraitu dugu aurrerago ikusiko dugun bezala.

4. OINARRI METODOLOGIKOAK

Hizkuntza desberdinetako corpusak etiketatzeko egin diren azterketa edo lan horiek kontuan izanda, parametro batzuk definitu ditugu, oinarri teoriko eta metodologiko nagusienak ezarri nahian *treebank*-a eraikitzeko.

Oinarrizko erabakiak honako hauek dira:

1. Zein elementu etiketatuko ditugu?

Etiketatze lan honetan, IXA taldean aztertu den azaleko analisi hori abiapuntu izanik analisi sakonago bat egin nahi da. Azalekoa da, testuan ageri diren elementuak bakarrik hartzen dituelako kontuan, sakoneko egituretan sartu gabe. Beraz, orain arte jorratutako sintaxian ez dago zuhaitzik edo antzeko egitura hierarkikorik; esaldi baten analisi sintaktikoa, funtzio sintaktikoa adierazten duten etiketa sintaktikoek osatzen dute. I.2 irudian ikus daitekeenez funtzio sintaktiko horiek elementuen arteko erlazioak erakusten dizkigute. Guk esaldia izan dugu aztergai; hau da, puntutik puntura doan testu-zatia[10]. Esaldiko elementu esplizituez gain, elidi tutako zenbait osagai etiketatu ditugu, hau da *pro*[11] gisa ezagun diren elementuak aintzat hartu ditugu. Anaforak eta antzeko zenbait kohesio elementu mementoz, alde batera utzi dira.

2. Osagai-egitura (constituency-based) *versus* mendekotasun-egitura (dependency-based)

Treebank-a osatzeko garaian erabili behar den etiketatze-eskemari buruzko eztabaidak irekita jarraitzen du.
Guk, zenbait saio egin ondoren eta (Skut *et al.*, 1997; Tapanainen, 1998; Oflazer 1999b) autoreen lanak kontuan harturik, euskara bezalako ordena libreko hizkuntzaren sintaxia lantzeko mendekotasun edo dependentzia egituren bidetik jarraitzea erabaki dugu.

Erabaki hori hartzeko garaian eragin zuzena izan dute puntu hauek:

a) Erlazio semantikoak adierazteko bidea ematen du.

b) Taldean garatu ditugun tresna informatikoen bidez mendekotasun erlazioak lortzea posible ikusten dugu. Horrela ez balitz ere, dependentzi-

[10] Perpaus diogunean, berriz, aditzetako bakoitzari dagokion zatiaz (perpaus nagusia, mendeko perpausa, ...) ari gara.
[11] *pro*: aditz laguntzaileekin komunztadura egiten duten osagai sintaktikoen elipsia; *pro-drop* fenomeno ezagunaren ondorioz agertzen dena.

zuhaitz batetik beste errepresentazio modu batera igarotzeko aukera dago, oso informazio aberatsa kudeatzeko aukera ematen duelako.

c) Ebaluaziorako.
Guk uste dugu, garbiagoa dela esaldia osatzen duten elementuen arteko lotura ebaluatzea, parentesiak non hasten eta bukatzen diren ikustea baino.
Dena den, ebaluazio erabilgarriena zein den erabakitzen lagunduko duten zenbait ikerlan aztertu beharra badugu ere, kontuan hartzekoa da (Carroll *et al.*, 1999)-k proposatzen duten ebaluazio-sistema; sistema hau dependentzia-egituren antzeko erlazio gramatikaletan oinarritzen dena baita.

3. Teoriarik jarraitzen al da?

Teoria bat jarraitzeak duen alde ona da arazo askori irtenbidea eman dakiokeela. Kontrakoa, berriz, teoria horiek ez dituztela corpusean ageri diren gauza asko aurreikusten.

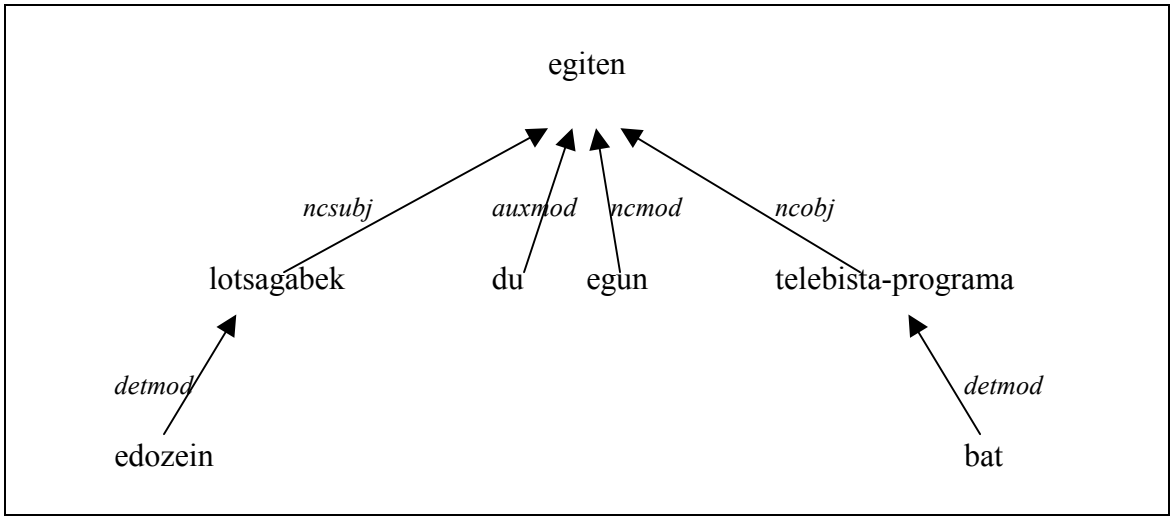
Guk, eta teoria bat izateak duen babesa kontuan izanda, zorrozki bete ez badugu ere, sortzailearen ildotik jarraitu dugu zenbaitetan. Esaterako, kategoria isilak (*pro*-ak besteak beste) aztertzerakoan.

5. EUSKARARAKO *TREEBANK*-A

Atal honetan oinarri linguistiko eta metodologiko sendoa izango duen etiketatze-sintaktikoaren eskema diseinatu eta definitu da. Era berean, hartutako lehen erabakiak eta corpusaren zati bat aztertzean aurkitu ditugun arazoak eta emandako irtenbideak papereratu ditugu.

5.1 Mendekotasun-egituran oinarritutako markaketa sintaktikoa

Mendekotasun egituren bidezko analisian osagaien arteko erlazioak deskribatzen dira; hala, corpuseko esaldi bakoitzari burua eta bere modifikatzaileen arteko mendekotasun edo dependentzia sintaktikoa zehazten zaio (ikus II. irudia), erlazio hori esplizitu eginez.



II. irudia. *"Edozein lotsagabek egiten du egun telebista-programa bat"* esaldiaren mendekotasun-egitura

Aukeratu dugun metodoak alde on hauek ditu osagai-egituran oinarritzen denarekin erkatuz gero:

a) Euskara hurrenkera libreko[12] hizkuntza izanik, ordena zurruna duten hizkuntzetan gertatzen ez diren zenbait arazori aurre egin beharrean aurkitzen da. Hau da, ingelesez subjektu eta objektuaren bereizketa aditzarekiko posizioaren arabera markatuta dago, baina euskararen ordena librearen eraginez zailagoa da bereizketa hori egitea. Arazo hori are nabarmenago egiten da osagai-egituraren bidezko analisian, bertan hurrenkera linealean oinarritzen delako eta ez duelako hierarkiarik adierazten.
Adibideekin argiago ikusiko dugu:

(2a). *Azeriak oiloa akabatu du*
(2b). *Oiloa azeriak akabatu du*

Bi esaldi hauen arteko desberdintasun bakarra hitz-hurrenkera da. Elementuen ordena diferenteak informazio gehigarria ematen digu, galdegaiarena, alegia.
Osagaietan oinarritutako analisia, ordea, desberdina litzateke:

[12] Ordena librea dugu euskaraz, elementuen eginkizun edo funtzioak baizik ez baditugu kontuan hartzen; esaterako hurrengo adibideak hogeita lau hurrenkera onartzen dituela esan izan da.

a. Aitak haurrari sagarra eman dio.
b. Aitak haurrari eman dio sagarra.
c. Aitak sagarra haurrari eman dio.
d. Aitak eman dio sagarra haurrari.
...

Ez da librea, ordea, perpausaren bigarren egitura ere, mintzagaiarena alegia (mintzaira-egitura deitu izan dena), gogoan atxekitzen badugu. (Euskaltzaindia, 1987. *Euskal Gramatika: Lehen Urratsak-I (Eranskina)*)

(2a). (P (IS azeriak)
(AS (IS oiloa)
(A akabatu du)))

(2b). (P (IS oiloa)
(AS (IS azeriak)
(A akabatu du)))

Parentesiez baliatzen den metodo hau oso nahasgarria izateaz gain, (2a) eta (2b) adibideetan ikus daitekeenez AS, bi analisietan bat ez datozen osagaiez osaturik agertzen zaigu; eta horietako bat, (2b), gaizki dago gainera.

Mendekotasun-egituran oinarritutako markaketan, aldiz, hurrenkera-askearen arazorik ez dago, hierarkietan oinarritzen delako. (2a) eta (2b) adibideak, esaterako, honela errepresentatuko lirateke:

ncsubj[13] (erg., akabatu, azeriak, azeriak, subj.)
ncobj (abs, akabatu, oiloa, oiloa, obj.)
auxmod (- , akabatu, du)

Markaketa honetan "azeriak" subjektua dela eta "oiloa" objektua dela besterik ez da zehazten. Osagaietan oinarritutako analisian ez bezala, hemen etiketa berak bi perpausentzat balio du.

b) Osagaien-egituran elementu edo osagai ez jarraiak modu erraz batean behintzat, tratatu ezin diren bezala, hemen badute irtenbidea.

(3) *Ikasleek oporrak hartu dituzte,* ***baita*** *irakasleak* ***ere***.

ncsubj (erg., hartu, ikasleek, ikasleek, subj)
ncobj (abs., hartu, oporrak, oporrak, obj.)
auxmod (- , hartu, dituzte)
lot (baita, e, irakasleak, ere)
ncsubj (erg., hartu, irakasleak, irakasleak, subj)

c) Horretaz gain eta hurbilpen honen mesedetan, metodo erraza eta intuitiboa dela esan genezake.

5.2 Euskarazko corpusa

Gure ikerlanean Lengoaia Naturalaren Prozesamenduko lanetarako erabiltzen diren honako bi corpus hauetaz baliatu gara:

a) *XX. mendeko euskararen corpus estatistikoa*[14], oraindaino EEBS (Egungo Euskararen Bilketa-lan Sistematikoa) izenarekin ezagutu izan dena.
XX. mendeko euskara jasotzen duen corpus estatistikoa da, eta testu-motei begira, orekatua[15].

[13] Etiketa hauen guztien deskribapena 5.3. puntuan ageri da.
[14] Ikus http://www.euskaracorpusa.net
[15] "Corpusen artean ondoko sailkapen sinplea egin daiteke: orekatua/ez-orekatua. Orekatuetan testu-moten artean halako oreka bat bilatzen da, testu-mota berezituei dagozkien ezaugarri partikularretatik

b) Kazetaritza-corpusa: euskarri elektronikoan ditugun *Euskaldunon Egunkaria*ren 1999ko urtarriletik 2000ko maiatza bitarteko ale guztiak biltzen dituena.
Euskara batua edo estandarra da *EGUNKARIA*ren hizkuntza eredua, hau da, Euskaltzaindiak 1968az geroztik batasunerako eman dituen irizpideak betetzen dituena.

Corpus horien ezaugarririk garrantzitsuena testu errealez osatuta egotea da; euskararentzat nahi dugun analizatzaile sintaktikoak, testu errealetan oinarritzen diren beste aplikazio batzuetarako izan behar baitu.

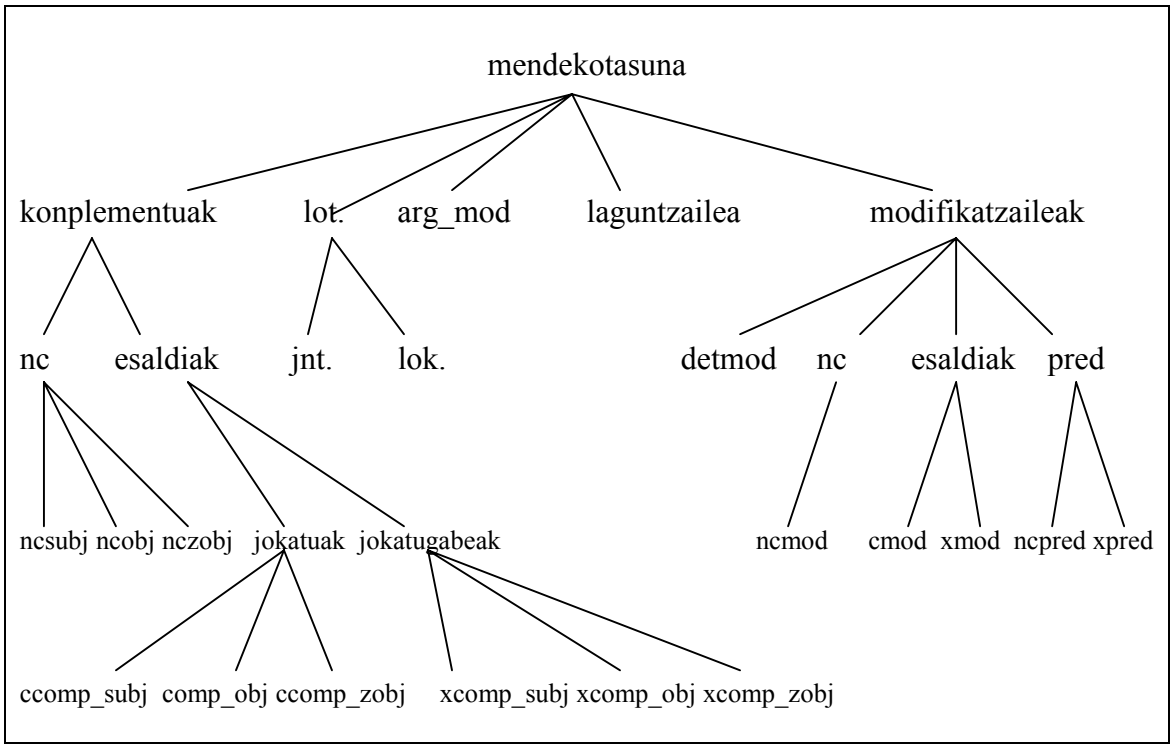
Corpus hauetatik morfosintaktikoki etiketatuta dauden 50.000 hitz hartuko ditugu; helburu orokorra horiek guztiak sintaktikoki etiketatzea bada ere, lan honetan horien guztien adierazgarri izango den lagin bat eman dugu ezagutzera.

5.3 Etiketek adierazten dituzten erlazio gramatikalen deskribapena

Esan bezala, mendekotasun harremanak oinarri dituen etiketatze-eskemaren aldeko apustua egin dugu, hurrenkera libreko hizkuntzentzat egokienak direla kontsideratu izan delako.
Etiketatze-eskema definitzeko (Carroll *et al.*, 1998b, 1999) lanetan oinarritu gara. Etiketatze modu honek ez du zerikusirik aplikazioarekin eta orain arte ingelesez, italieraz, frantsesez eta alemanez gertatzen diren fenomeno linguistikoak hartzen ditu kontuan.

Horrela, lan-ildo hau jarraituz, erlazio gramatikalen (errealizazio lexikala duten elementuena zein *pro* direnena) hierarkia batean oinarritutako kodeketa-eskema egin dugu (ikus III. irudia)

aldenduz. Horretarako, iturburu desberdinetatik testu-zati txiki samar anitz, esanguratsuak eta aberasgarriak biltzen dira, teknika estatistikoak erabiliz" (Alegria, 1995)



III.irudia. Erlazio gramatikalen hierarkia

Ikus daitekeen bezala hierarkia horretan maila orokor batzuk agertzen dira, hauek aldi berean beste maila batzuetan zehazten dira. Horrela, esaterako, maila orokorrean konplementuak edo osagarriak, lotura elementuak, paper tematikoak (arg_mod), laguntzailea eta modifikatzaileak dauzkagu. Hauek, berriz ere izen-sintagma (nc)[16] eta perpausetan (ccomp)[17] banatzen dira. Mailakatze honekin jarraituz, maila bakoitzaren zehaztasun handiagoa lortzen da, euren funtzio gramatikalak kontuan hartuz (adib.: ncsubj., ncobj. eta nczobj.)

Ondoren hierarkia honetan adierazitako erlazio gramatikal bakoitzaren deskribapena dator. Deskribapen honek garrantzi handia du, bertan erlazio bakoitzak beharrezkoa duen etiketa kopurua eta mota zehazten delako (eremu kopurua, bakoitzaren ezaugarriak, etab.). Lan hau oso baliagarria izango da etorkizuneko tratamenduetarako, informazio hau guztia SGML[18] formatuan lortzeko adibidez.

5.3.1 Konplementuak

5.3.1.1 Konplementuak nc (IS) direnean (5 eremu) :

ncsubj (kasua, burua, ISren burua, ISren barruan kasu marka daraman elementua, subj)
ncobj (kasua, burua, ISren burua, ISren barruan kasu marka daraman elementua, obj)
nczobj (kasua, burua, ISren burua, ISren barruan kasu marka daraman

[16] nc: "*non-clausal complement*"; izen eta postposizio sintagmak, alegia.
[17] ccomp: "*clausal complement*"; perpausa, alegia.
[18] SGML: testuak markatzeko lengoaia estandar eta orokorra, hots, *Standard Generalized Markup Language* marka-multzo bat baino marka-multzoak espezifikatzeko metalengoaia bat da.

elementua, zobj)

Adib.: (4) *Aitak haurrari sagarra eman dio*.

ncsubj (erg, eman, aitak, aitak, subj)
nczobj (dat, eman, haurrari, haurrari, zobj)
ncobj (abs, eman, sagarra, sagarra, obj)

5.3.1.2 Konplementuak perpausak direnean (4 eremu):

JOKATUAK:

ccomp_subj
ccomp_obj (perpaus mota, burua, menpekoaren burua,
ccomp_zobj erlazio atzizkia daraman laguntzailea)

Adib.: (5) *Nolabait uste izan ninan zerbait egin nezakeela bera hala egon ez zedin.*

ccomp_obj (konp, uste izan, egin, nezakeela)

JOKATUGABEAK:

xcomp_subj
xcomp_obj (perpaus mota, burua, menpekoaren burua,
xcomp_zobj erlazio atzizkia daraman hitza)

Adib.: (6) *Baina haren kontra ez zegonan ezer egiterik.*

xcomp_subj (konp, zegonan, egiterik, egiterik)

5.3.2 Lotura elementuak

5.3.2.1 Koordinazioa.

lot (partikula, maila bereko sintagmak edo perpausak elkartzen dituen juntagailu mota, elkartzen den lehen sintagma edo perpauseko burua)
lot (partikula, maila bereko sintagmak edo perpausak elkartzen dituen juntagailu mota, elkartzen den bigarren sintagma edo perpauseko burua)

5.3.2.1.1 Koordinazioa IS mailan :

Adib.: (7) *Ni eta zu etorri gara.*

lot (eta, izen_EMEN, ni)
lot (eta, izen_EMEN, zu)

5.3.2.1.2 Koordinazioa AS mailan :

Adib.: (8) *Nik egin nuen eta zuk ere bai.*

lot (eta, aditz_EMEN, egin)
lot (eta, aditz_EMEN, e)

5.3.2.1.3 Koordinazio konplexua IS mailan:

Adib.: (9) *Bera, zu eta ni etorri gara.*

lot (puntuazio-marka, izen_EMEN, bera)
lot (eta, izen_EMEN, zu)
lot (eta, izen_EMEN, ni)

5.3.3 arg_mod[19](4 eremu). IS direnek bakarrik daramate etiketa hau. Hasteko, erlazio tematikoak markatzeko balioko du.

arg_mod (- , burua, ISren burua, paper tematikoa[20])

Adib.: (10) *Aitak zulo handi bat egin zinan Timorentzat palaz.*

arg_mod (- , egin, aitak, subj)

5.3.4 Laguntzailea (3 eremu). Aditz nagusiarekin agertzen den aditz laguntzailea.

aux_mod (- , burua, aditz laguntzailea)

Adib.: (11) *Begietatik igarri nionan ez zela bizi.*

auxmod (- , igarri, nionan)

5.3.5 Modifikatzaileak:

5.3.5.1 determinatzaileak (3 eremu)

detmod (- , ISren burua, determinatzailea)

Adib.: (12) *Zenbait ikaslek greba egin dute.*

detmod (- , ikaslek, zenbait)

[19] arg_mod, hau da, etiketa semantikoa.
[20] Aztertutako esaldietan eremu honetan adierazten dena ez da paper tematikoa, funtzio sintaktikoa baizik. Aurrerago zehaztuko da eremu honi dagokion benetako paper tematikoa zein den.

5.3.5.2 Modifikatzaileak nc (IS edo postposizio-sintagma) direnean (4 eremu):

ncmod (kasua, burua, IS edo postposizio-sintagmaren burua, kasu marka daraman elementua)

Adib.: (13) *Begietatik igarri nionan ez zela bizi.*

ncmod (abl, igarri, begietatik, begietatik)

ncmod (- , burua, IS edo postposizio-sintagmaren burua, kasu marka daraman elementua)

Adib.: (14) *Ia gauza guztiengatik egin liteke zerbait.*

ncmod (- , egin, ia, ia)
ncmod (mot, egin, gauza, guztiengatik)

5.3.5.3 Modifikatzaileak perpausak direnean (4 eremu):

JOKATUAK:

cmod (perpaus mota, burua, menpekoaren burua, erlazio atzizkia daraman laguntzailea)

Adib.: (15) *Nolabait uste izan ninan zerbait egin nezakeela bera hala egon ez zedin.*

cmod (helb, egin, egon, zedin)

JOKATUGABEAK:

xmod (perpaus mota, burua, menpekoaren burua, erlazio atzizkia daraman hitza)

Adib.: (16) *Juttak aitonarentzako egindako gurutzearen antz-antzekoa da.*

xmod (erlt, gurutzearen, egindako, egindako)

5.3.5.4 Modifikatzaileak predikatu direnean (4 eremu):

ncpred (- , burua, ISren burua, ISren barruan kasu marka daraman elementua)

Adib.: (17) *Nolabait uste izan ninan zerbait egin nezakeela bera hala egon ez zedin.*

ncpred (- , egon, hala, hala)

xpred (- , burua, menpekoaren burua, erlazio atzizkia daraman hitza)

Adib.: (18) *Zernahi lor dezakegulako ustea hezur muinetaraino sartua dugu.*

xpred (- , ustea, sartua, sartua)

Etiketatze-eskema hau Lin-ek (1995) azaleko egituran erabiltzen duenaren antzekoa da. Dena den, desberdintasunen bat edo beste badago: lexikoki gauzatzen ez diren argumentuak, erlazio gramatikaletan ager daitezke (esaterako, *pro-drop* hizkuntzetan subjektuaren dependentzia **Pro** bezala zehazten da).

5.4 Eskuz sintaktikoki aztertutako adibideak

Aurreko atalean deskribatutako erlazio gramatikal horiek erabiliz, ondorengo adibideetan agertzen diren moduan kodetzen dira esaldiak. Jarraitu dugun formalismoan aditzetik etiketatzen hasten badira ere, guk ezkerretik eskubira doan segida jarraitu dugu. Edozein modutan, ordena ez da garrantzizkoa aplikazioei begira, azken finean mendekotasun-egitura horiek erlazio gramatikal berak adierazten baitituzte. Ordena kontu hauek, batez ere eskuzko etiketatzeaileari begira definitu behar izan dira. Adibide hauek proiektuaren zailtasuna batetik, eta bestetik ezaugarri honetako baliabidea lortzeak LNPn ikerketarako duen garrantzia adierazten dute.

(19) *Aitak zulo handi bat egin zinan Timorentzat palaz.*

ncsubj (erg, egin, aitak, aitak, subj)

arg_mod (- , egin, aitak, subj)

ncobj (abs, egin, zulo, bat, obj)

ncmod (- , zulo, handi, bat)

detmod (- , zulo, bat)

arg_mod (- , egin, zulo, obj)

auxmod (- , egin, zinan)

ncmod (des, egin, Timorentzat, Timorentzat)

arg_mod (- , egin, Timorentzat, mod)

ncmod (ins, egin, palaz, palaz)

arg_mod (- , egin, palaz, mod)

(20) *Zurezko gurutze bat aurrean daukan sastraka baten ondoan daude.*

ncmod (gel, gurutze, zurezko, zurezko)

ncobj (abs, daukan, gurutze, bat, obj)

detmod (- , gurutze, bat)

arg_mod (- , daukan, gurutze, obj)

ncmod (ine, daukan, aurrean, aurrean)

arg_mod (- , daukan, aurrean, mod)

ncsubj (- , daukan, PRO[21], PRO, subj)

arg_mod (- , daukan, sastraka, subj)

cmod (erlt, sastraka, daukan, daukan)

detmod (- , sastraka,baten)

ncmod (-en ondoan , daude, sastraka, baten ondoan)

ncsubj (abs, daude, pro, pro, subj)

arg_mod (- , daude, haiek, subj)

[21] *PRO*: azaleko sintaxian fonetikoki gauzatu ez daitekeen izenordaina. Ingeleseko *"John tried to win"* esaldian, adibidez, PRO "win"-en subjektua da; hau da, *"John tried (PRO to win)"*.
pro, aldiz, ageriko subjektu bat azaleratzeko aukera dagoenean gertatzen da. Bere ezaugarriak "+izenordainkia" eta " - anaforikoa" dira.

(21) *Juttak aitonarentzako egindako gurutzearen antz-antzekoa da.*

ncsubj (erg, egindako, Juttak, Juttak, subj)

arg_mod (- , egindako, Juttak, subj)

ncmod (gel, egindako, aitonarentzako, aitonarentzako)

arg_mod (-, egindako, aitonarentzako, mod)

ncobj (- , egindako, PRO, PRO, obj)

arg_mod (- , egindako, gurutzea, obj)

xmod (erlt, gurutzearen, egindako, egindako)

ncmod (gen, antz-antzekoa, gurutzearen, gurutzearen)

ncpred (- , da, antz-antzekoa, antz-antzekoa)

ncsubj (abs, da, pro, pro, subj)

arg_mod (- , da, hura, subj)

(22) *Gero Dirkek kontatzen dio, nola duela urtebete izan zen hori.*

ncmod (- , kontatzen, gero, gero)

ncsubj (erg, kontatzen, Dirkek, Dirkek, subj)

arg_mod (- , kontatzen, Dirkek, subj)

nczobj (dat, kontatzen, pro, pro, zobj)

arg_mod (- , kontatzen, hari, zobj)

ccomp_obj (zhg, kontatzen, zen, nola)

auxmod (- , kontatzen, dio)

cmod (denb, izan, duela, duela)

ncpred (- , duela, urtebete, urtebete)

auxmod (- , izan, zen)

ncsubj (abs, izan, hori, hori, subj)

arg_mod (- , izan, hori, subj)

(23) *Afganistanen lurrikarak hondatutako ingurura hasi da laguntza iristen.*

ncmod (ine, hondatutako, Afganistanen, Afganistanen)

arg_mod (- , hondatutako, Afganistanen, mod)

ncsubj (erg, hondatutako, lurrikarak, lurrikarak, subj)

arg_mod (- , hondatutako, lurrikarak, subj)

ncobj (abs, hondatutako, pro, pro, obj)

arg_mod (- , hondatutako, hura, obj)

xmod (erlt, ingurura, hondatutako, hondatutako)

ncmod (ala, iristen, ingurura, ingurura)

arg_mod (- , iristen, ingurura, mod)

auxmod (- , hasi, da)

ncsubj (abs, iristen, laguntza, laguntza, subj)

arg_mod (- , iristen, laguntza, subj)

xcomp_obj (konp, hasi, iristen, iristen)

(24) *Gosetea eta gaixotasunak zabal daitezke.*

ncsubj (abs, zabal, gosetea, gosetea, subj)

arg_mod (- , zabal, gosetea, subj)

lot (eta, izen_EMEN, gosetea)

lot (eta, izen_EMEN,gaixotasunak)

ncsubj (abs, zabal, gaixotasunak, gaixotasunak, subj)

arg_mod (- , zabal, gaixotasunak, subj)

auxmod (- , zabal, daitezke)

5.5 Etiketatzerakoan sortutako arazoak eta irtenbideak

Atal honetan euskararen morfosintaxiaren deskribapena egin ez bada ere, esku artean dugun lanean eragin berezia izan duten hizkuntzaren zenbait ezaugarri nola tratatu ditugun eman nahi da jakitera.

5.5.1 Izen-sintagma barruko zein osagai izango da burua?

Euskarak baditu tipologikoki ezaugarri batzuk inguruko hizkuntzetatik berezi egiten dutenak. Hala, 'buru azkena' duen hizkuntza dela esaten da, sintagma-buruak eskuineko posizioan jartzeko joera duelako. Euskararen sintagma egiturari erreparatzen badiogu, marka morfologikoak sintagma hori osatzen duen azken osagaiak daramatzala ohartuko gara. Kasu marka horiek, ordea, ez zaizkio beti buru berari eransten: zenbaitetan, sintagma horretako buruari erants dakizkioke, esaterako *Edozein lotsagabek egiten du ...* eta beste batzuetan determinatzaileari, adibidez *Lotsagabe batek egiten du ...*

Hala, aztertu dugun corpus zatiaren analisian aurre egin beharreko arazoetako[22] bat hori izan da. Ondoko adibide hauetako izen-sintagmen azterketak arazo horien isla dira.

(24) *Edozein lotsagabek egiten du egun telebista-programa bat.*

detmod (- , lotsagabek, edozein)
ncsubj (erg, egiten, lotsagabek, lotsagabek, subj)
arg_mod (- , egiten, lotsagabek, subj)

(24) adibidean 'ncsubj'-ari dagokion etiketan eremuak errepikatuta agertzea deigarria gertatzen da; horrek, ordea, badu bere arrazoia. Izen-sintagmetan oinarritzat edo burutzat izena hartzen badugu, markaketa desberdina izango da (25) adibidean ikus daitekeen bezala; batean ergatiboduna (24) izango da eta bestean (25) ez.

(25) *Lotsagabe batek egiten du egun telebista-programa bat.*

ncsubj (erg, egiten, lotsagabe, subj)
detmod (- , lotsagabe, batek)
arg_mod (- , egiten, lotsagabe, subj)

Horri irtenbidea eman nahian eta dependentzia eredua apur bat ezaugarri horietara moldatuz, erlazio gramatikalaren deskribapenean eremu bat gehitzea erabaki dugu, kasua daraman hitzarena hain zuzen. Horrekin 'lotsagabe batek' kasuan, 'lotsagabe' da gunea eta 'batek' osagaia gehituko genioke, berak daramalako kasua.

[22] Argitu beharra dago, arazoa benetan ez dela euskararena, baizik eta hitzean oinarritutako MG analizatzailearena.

(26) *Lotsagabe batek egiten du egun telebista-programa bat.*

ncsubj (erg, egiten, lotsagabe, **batek**, subj)
detmod (- , lotsagabe, batek)
arg_mod (- , egiten, lotsagabe, subj)

5.5.2 pro izeneko kategoria isilak

Euskararen aditz-flexioa oso aberatsa da. Aditz trinko zein aditz laguntzaileak berarekin komunztadura egiten duten kasuei (absolutiboa, ergatiboa eta datiboa) buruzko informazioa ematen du. Hala, esaldi batean kasu hauei dagokien sintagmaren bat agertzen ez bada ere, aditz laguntzaileak horren berri ematen du (*pro-drop* izeneko hizkuntzen ezaugarria), beraz kontuan hartu behar da elementu hori aztergai dugun esaldiko aditzari dagokiola.

Adib.: (27) *Begietatik igarri nionan ez zela bizi.*

ncmod (abl, igarri, begietatik, begietatik)

arg_mod (- , igarri, begietatik, mod)

ncsubj (erg, igarri, pro, pro, subj)

arg_mod (- , igarri, nik, subj)

nczobj (dat, igarri, pro, pro, zobj)

arg_mod (- , igarri, hari, zobj)

auxmod (- , igarri, nionan)

...

5.5.3 Koordinazioa

Koordinatzen diren osagaiak edo elementuak maila berekoak direla adierazteko, *"head"* edo buru imaginario bat proposatu dugu, ondoko adibidean ikus daitekeen bezala.

Adib.: (28) *Ni eta zu etorri gara*.

head

Ni eta zu etorri gara.

lot (eta, izen_EMEN, ni)
lot (eta, izen_EMEN, zu)

5.5.4 Osagai ez jarraiak

Orain arteko Ixa taldean garatutako gramatika konputazionalean, bere barruan hartzen dituen fenomenoetan ondoz ondoko osagaiak lotzen dira beti. Oraingoan, berriz, beste urrats bat egin dugu eta ez jarraiak diren osagaiak aztertzera iritsi gara; horren adibidea (3) esaldia dugu.

5.5.5 Elipsia

Elipsia nola tratatu behar den erabakitzeak buruhauste bat baino gehiago eman du hizkuntzalaritza konputazionalean. Guk azterketa honetan, lanean aurrera egin ahala gehiago zehaztu beharko badugu ere, elipsia markatzea erabaki dugu, horretarako *e* markaz baliatuz. Hor dugu, esaterako, koordinazio mailan gertatzen den elipsiaren adierazgarri den (8) adibidea.

5.5.6 Konparaziozko perpausak

Honela aztertzea erabaki dugu:

Adib.: (29) *Nik zuk baino diru gehiago irabazten dut.*

ncsubj (erg, irabazten, nik, nik, subj)

arg_mod (- , irabazten, nik, subj)

ncsubj (erg, e, zuk, zuk, subj)

arg_mod (- , irabazten, zuk, subj)

ncobj (abs., e, PRO, PRO, obj)

arg_mod (- , irabazten, diru, obj)

lot_konp (baino, diru, e)

ncmod (- , diru, gehiago, gehiago)

ncobj (abs, irabazten, diru, gehiago, obj)

arg_mod (- , irabazten, diru, obj

auxmod (- , irabazten, dut)

5.5.7 Esaldi anbiguoak

Eskuz etiketatzen ari garen corpusean gertatzen diren esaldi anbiguoak aztertzea erabaki dugu.

Adib.: (30a) *Altxa eta jardinera garraiatu ninan Timo.*

ncsubj (erg, altxa, pro, pro, subj)

arg_mod (- , altxa, nik, subj)

ncobj (abs, altxa, pro, pro, obj)

arg_mod (- , altxa, hura, obj)

lot (eta, aditz_EMEN, altxa)

lot (eta, aditz_EMEN, garraiatu)

ncmod (ala, garraiatu, jardinera, jardinera)

arg_mod (- , garraiatu, jardinera, mod)

ncsubj (erg, garraiatu, pro, pro, subj)

arg_mod (- , garraiatu, nik, subj

auxmod (- , garraiatu, ninan)

ncobj (abs, garraiatu, Timo, Timo, obj)

arg_mod (- , garraiatu, Timo, obj)

(30b) *Altxa eta jardinera garraiatu ninan Timo.*

ncsubj (abs, altxa, pro, pro, subj)

arg_mod (- , altxa, ni, subj)

lot (eta, aditz_EMEN, altxa)

lot (eta, aditz_EMEN, garraiatu)

ncmod (ala, garraiatu, jardinera, jardinera)

arg_mod (- , garraiatu, jardinera, mod)

ncsubj (erg, garraiatu, pro, pro, subj)

arg_mod (- , garraiatu, nik, subj

auxmod (- , garraiatu, ninan)

ncobj (abs, garraiatu, Timo, Timo, obj)

arg_mod (- , garraiatu, Timo, obj)

6. ESKUZKO MARKAKETAN LAGUNTZEKO TRESNAK

Eskuz egindako lan honen abantaila nagusia hizkuntzalarien ezagutza guztia erabili ahal izatea da, baina arazorik handiena horrek eskatzen duen lan ikaragarria dugu, testuetan dagoen informazioa zabala eta ia bukaezina delako. Horrez gain, corpus handiagoak erabiltzen diren neurrian, handitu egiten da hizkuntzalariak errorea egiteko probabilitatea, edo antzeko fenomenoak modu desberdinean tratatzeko arriskua. Hori dela-eta, etiketatze sintaktikoan erabiliko ditugun etiketak definitzearekin batera, corpusaren etiketatze automatikoa helburu izango duen tresna informatikoa ari gara lantzen. Honekin, batetik, etiketatzaileek etiketatze lana erraztuko dien tresna bat izatea lortu nahi da, eta bestetik (testu)-editoreak ikasketa sistema bat edukitzea, hala markaketa lana aurrera doan heinean ezagutza berria erants edo gehitu dakion (Brants *et al.*, argitaratzear). Tresna hau mendekotasun-markaketak egiteko erabiliko da. Ezagutzaren gehitze honek, etengabeko hobekuntza, etiketatzearen eginkortasuna eta sendotasuna, eta prozesuaren azkartasuna dakartza berarekin.

7. ONDORIOAK

Ikerlan honetan *treebank*-a edo sintaktikoki etiketatutako corpusa eraikitzeko lehen urratsak egin dira. Aukera desberdinak aztertu ondoren, dependentzia erlazioetan oinarritutako formalismoa jarraitzeko erabakia hartu dugu, euskara bezalako hurrenkera askea duten hizkuntzentzat aproposena delako batetik, eta bestetik, metodo erraza eta intuitiboa izateaz gain, eskema honen malgutasunari esker, beste mota bateko etiketak sar ditzakegulako, esaterako, paper tematikoei dagozkienak. Etorkizunean helduko diogun lanerako urrats garrantzitsua da hau.

Era berean, orain arte egin den analisi sintaktikoaren hobekuntza egin da, azaleko egiturak aztertzetik sakonekoetara pasa gara, gunea eta modifikatzailearen arteko erlazioa esplizitu eginez. Azterketa honetan sortutako zenbait arazori (osagai ez jarraiak, koordinazioa, konparaziozko perpausak, etab.) irtenbidea bilatu diegu.

Bukatzeko, sintaktikoki etiketatutako corpus honen beharra azpimarratu behar da, taldean osatzen jarraitzen dugun euskararako *parser*-aren ebaluaziorako eta garapenerako balio baitu. Hartara, landu dugun sintaxiaren sendotasuna zenbaterainokoa den azter daiteke. Bestalde, corpusaren gainean lan egiteak landuriko metodologia ebaluatzeko balioko digu, zer-nolako emaitzak lortzen diren ikusi ahal izateko. Izan ere gogoan izan behar dugu ebaluazioak berebiziko garrantzia duela LNPko aplikazioen alorrean.

GLOSATEGIA

analizatzaile morfologikoa

Hitzak zati morfologikoetan banatu eta morfemen arteko loturak bideratzen dituen tresna.

analizatzaile sintaktikoa

Hauen zeregina testuetako osagai sintaktikoak ezagutzea da: hitz isolatuez osatutako sekuentzietan elkarrekin lotuta dauden egitura sintaktikoak (perpausa, izen-sintagmak, aditz-sintagmak, izenlagunak, etab.) ezagutzen ditu.

baterakuntza-formalismoak

Hitzek osatzen dituzten unitate handiagoak, ezaugarri-egituretan oinarritutako herentzia mekanismoez lortzen dituzte.

corpusa

Hizkuntza bateko testu-multzo bat da, helburu zehatz batekin sortua.

datu-basea

Informazioa modu egituratuan gordetzeko euskarri informatikoa.

desanbiguazio morfosintaktikoa

Hitz-forma baten osaera morfologiko posibleetatik, testu-zatian dagokion interpretazioa esleitzeko beharrezkoa den prozedura.

esleitu

Esate baterako, funtzio sintaktikoak esleitzen dira, hau da, hitz bati funtzio sintaktiko bat jarri edo eman egiten zaio.

etiketatzailea

Hitzari testu-zatian dagokion interpretazioa esleitzen dion tresna, bere baitan prozesu desanbiguatzaileak baliatzen dituena. Horrez gain, funtzio sintaktikoak esleitzeko ahalmena du, eta testua sintaktikoki etiketatzea ahalbidetzen du. Hala, etiketatzaileak analizatzaile sintaktiko partzialak dira neurri handi batean.

flexioa

Lemari informazio morfosintaktikoa ematen dion atzizkia.

lematizatzailea

Hauen zeregina corpus bateko testu-hitz bakoitzari bere lema eta kategoria esleitzea da.

Lengoai Naturalaren Prozesamendua (LNP)

Ordenagailuaz baliatuz hizkuntza tratatzea helburu duen ikerkuntza-arloa; nahiz eta batzuetan hizkuntzalaritzako ikuspuntua garrantzitsua denean batez ere, Linguistika Konputazionala ere esaten zaion.

Murrizpen Gramatika (MG)

Analizatzaile morfologikoaren emaitzak, hitz bakoitzaren analisi posible guztiak ematen ditu. Horren gainean, bada, aplikatzen da Murrizpen Gramatika (MG) deituriko formalismoa. Honek, informazio linguistikoan oinarrituz, desanbiguazio morfologikoa burutzen du.

sintaxi partziala

Sintaxi partzialak analisi tradizionalaren informazioaren zati bat, ez guztia, aztertzen du; hau da, azaleko egituretatik abiatuta zenbait erlazio sintaktiko adierazten ditu. Sintaxi partzialean erabiltzen diren teknikek fidagarritasuna eta sendotasuna dute helburu, sakontasuna eta osotasuna neurri batean galduz.

tag

etiketa

tagging

Zenbait markaketa linguistiko, hala nola hitzei kode bereziak atxikitzea bere zenbait ezaugarri adierazteko, markaketa gisa baino gehiago etiketatze (*tagging*) gisa ezagutzen dira; eta ezaugarriei egokitzen zaizkien kodeei etiketa (*tag*) esaten zaie.

token

Testu unitatea.

Objektu-mota jakin bat izendatzeko hartutako hitz edo ikurra.(*Informatika hiztegia*. UZEI. 1993)

treebank

Zuhaitz itxura duten sintaktikoki analizatutako corpusak.

" Corpus anotado con información de estructura de frase" (McEnery *et al.,* 1996)

ERREFERENTZIAK

Atal honetan, ikerlanean zehar erreferentziatu diren obrak aurkituko dira.

Abeillé A., Clément L. & Kinyon A. "Building a Treebank for French" in *Building and Using syntactically annotated corpora*, Abeillé, A. Ed. Kluwer, Dordrecht (argitaratzear)

Aduriz I., Aldezabal I., Alegria I., Ezeiza N. & Urizar R. (1996). "Del analizador morfológico al etiquetador/lematizador: unidades léxicas complejas y desambiguación". SEPLN'96. Sevilla

Aduriz I. (2000). *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Alegria I. (1995). *Euskal morfologiaren tratamendu automatikorako tresnak*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Alegria I., Artola X., Sarasola K. & Urkia M. (1996a). "Automatic morphological analysis of Basque". *Literary and Linguistic Computing*. *11 (4)*. Oxford University Press.

Arriola J.M., Artola X., Maritxalar A. & Soroa A. (1999) "A Methodology for the Analysis of Verb Usage Examples in a Context of Lexical Knowledge Acquisition from Dictionary Entries" in *Proceedings of Linguistically Interpreted Corpora (EACL'99)*, Bergen (Norvegia)

Arriola J. M. (2000). *Euskal hiztegiaren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Basili R., Pazienza M.T. & Zanzotto F.M. (2000) *Customizable Modular Lexicalized Parsing*. IWPT´2000, Trento.

Böhmova A., Panemová J. & Sgall P. (1999). "Syntactic Tagging: procedure for the Transition from the Analytic to the Tectogrammatical Tre Structure" *in Proceedings of the Second Workshop on Text, Speech and Dialogue*, Marianske Lazne, República Checa.

Bosco C., Lombardo V., Vassallo D. & Lesmo L. (2000). "Building a Treebank for Italian: a Data-driven Annotation Schema" in *Proceedings of the Second Conference on Language Resources and Evaluation* (LREC2000), Atenas

Brants T., Skut W. & Uszkoreit H. "Syntactic Annotation of a German Newspaper Corpus" in *Building and Using syntactically annotated corpora*, Abeillé, A. Ed. Kluwer, Dordrecht (argitaratzear)

Burnage G. & Dunlop D.(1993). "Encoding the British National Corpus" *in Aarts et al eds*. (1992) *English language corpora: design, analysis and explotation*, Amsterdam Rodopi, pp. 79-95

Carroll J., Briscoe T. & Sanfilippo A. (1998b). "Parser evaluation: a survey and a new proposal". In *Proceedings of the International Conference on Language Resources and Evaluation*, 447-454. Granada (Spain)

Carroll J., Minnen G. & Briscoe T. (1999). *Corpus Annotation for Parser Evaluation.* Proceedings of Workshop on Linguistically Interpreted Corpora, EACL´99, Bergen.

Charniak E. (1996). "Tree-bank Grammars" Technical report CS-96-02

Chomsky, N. (1957). *Syntactic Structure*. Mouton

Euskaltzaindia (1987). *Euskal Gramatika: Lehen urratsak-I (Eranskina)*. Euskaltzaindia, Bilbo.

Gojenola K. (2000). *Euskararen sintaxi konputazionalerantz*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Järvinen T. & Tapanainen P. (1997). *A Dependency Parser for English*. Technical Report, nº TR-1, Department of General Linguistics. University of Helsinki.

Karlsson F., Voutilainen A., Heikkila J. & Anttila A. (1995). *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Kübler S. & Hinrichs E.W. (2001). "From Chunks to Function-Argument Structure: A Similarity-Based Approach" in *Proceedings of the ACL01*.

Lin D. (1995). "A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1420-1425. Montreal (Canada)

Marciniak M., Mykowiecka A., Przepiórkowski A & Kupsc A. "Construction of an HPSG treebank for Polish" in *Building and Using syntactically annotated corpora,* Abeillé, A. Ed. Kluwer, Dordrecht (argitaratzear)

Marcus M., Santorini B. & Marcinkiewicz M. (1993). "Building a Large Annotated Corpus of English: the Penn Treebank" *Computational Linguistics.*

McEnery T. & Wilson A. (1996). *Corpus Linguistics*. Edinburgh University Press.

Montemagni S., Barsotti F., Batista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M., Saracino D., Zanzotto F., Mana N., Pianesi F. & Delmonte R. "Building the Italian Syntactic-Semantic Treebank" in *Building and Using syntactically annotated corpora,* Abeillé, A. Ed. Kluwer, Dordrecht (argitaratzear)

Moreno A., Grishman R., López S., Sánchez F. & Sekine F. (2000). "A Treebank of Spanish and its Application to Parsing" *in Proceedings of the Second Conference on Language Resources and Evaluation* (LREC2000), Atenas.

Oflazer K. (1999a). "Dependecy parsing with an Extended Finite State Approach" ACL'99, Maryland.

Oflazer K., Zeynep D., Tür H. & Tür G. (1999b). "Design for a Turkish treebank" *Proceedings of Workshop on Linguistically Interpreted Corpora*, at EACL´99, Bergen.

Sampson G. (1995). "English for the Computer The SUSANNE corpus and Analytic Scheme" Clarendon Press, Oxford.

Shieber S.M. (1986). *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, 4 zenbakia, Stanford.

Skut W., Krenn B., Brants T. & Uszkoreit H. (1997). "An Annotation Scheme for Free Word Order Languages", Fifth Conference on Applied Natural Language Processing (ANLP´97), Washington, DC, USA, 88-95.

UZEI (1993). *Informatika hiztegia*. Elkar.