Desanbiguazio morfologikoa, azterketa sintaktikoaren lehen urratsak eta aplikazioak Murriztapen Gramatikaren eredu konputazionala jarraituz

Itziar Aduriz, Jose Mari Arriola, Arantza Diaz de Ilarraza
IXA taldea
Lengoaia eta Sistema Informatikoak Saila

Euskal Herriko Unibertsitatea
Informatika Fakultatea, E-20080 Donostia

Laburpena

Artikulu honen bidez, *Constraint Grammar*, (CG) (Murriztapen Gramatika) izeneko formalismo sintaktikoa aurkeztu nahi da; bere oinarriak eta helburuak zein diren azaldu, eta batez ere, euskararen tratamendu sintaktiko automatikoari ekiteko oinarri gisa formalismo hau aplikatzeko egindako urratsak. Horrez gain, Euskal Hiztegiko (EH) aditzen adibideak aztertzeko egindako analisi sintaktikoa deskribatuko dugu. Analisi hori egiteko euskararako garatu dugun Murriztapen Gramatika hartu da abiapuntutzat aditzen azpikategorizazio-lanetarako laguntza automatikoak eskaintzeko.

1. Sarrera

Linguistika konputazionala diziplinarteko arloa da eta adimen artifizialeko ikerkuntzarekin badu harremanik. Bi motibazio nagusi ditu: teknologikoa eta linguistikoa. Lehenengoari dagokionean, konputazio-sistema adimentsuak garatzea du helburu, hala nola, datu-baseei hizkuntza naturalez galdetzeko interfazeak, itzulpen automatikorako sistemak, testuen analisirako *parser*rak, ahotsaren tratamendurako tresnak, etab. Bigarrenean, alderdi linguistikoari erreparatzen zaio, eta gizakiok hizkuntza naturalaren bidez komunikatzen garenez, komunikazio-tresna horren ulermen sakonagoa erdietsi nahi da. Bi arrazoi nagusi hauek kontuan harturik, linguistika konputazionalaren xedea, honako hau da: hizkuntz ulermenaren eta sorkuntzaren teoria konputazional ulergarri, taxutu eta linguistikoki motibatua eraikitzea.

Lengoaia Naturalaren Prozesamenduaz (LNP) hitz egiten da Linguistika konputazionalaren arloan, lengoaia naturalaren tratamendu automatikoaz aritzerakoan.

Lengoaia naturalaren tratamendu automatikoari ekiteko analisi sintaktikoa edota *parsing*a ezinbesteko urratsa da. Zaila da, oinarritzat analisi sintaktiko sendorik ez duten aplikazio konputazional aurreratuetan pentsatzea. Horrela bada, 70. eta 80. hamarkadetan LNParekiko interesa areagotu egin zen, eta zehazki alderdi sintaktikoari zegokiona.

Euskararen sintaxiaren tratamendu automatikorako bi hurbilpen nagusi landu dira IXA[1] taldearen baitan: batean, baterakuntza-formalismoak hartu dira oinarritzat, eta, bestean, Murriztapen Gramatikaren bideari ekin zaio. Hain zuzen ere, txosten honetan ildo horretatik sintaxiaren eremuan egindako oinarrizko lanak eta aplikazioak izango ditugu mintzagai.

2. Analizatzaileak edota parserrak

Esaldi bat sintaktikoki analizatzen denean, egituraren bat esleitzen zaio. Egitura horrek esaldiko osagai linguistikoak errepresentatzen ditu, eta beraien arteko harreman gramatikalak azalarazten ditu. Lan hori guztia modu mekanikoan gauzatzen duten algoritmoak *parser* izenarekin ere ezagutzen dira. Beraz, *parser* edota analizatzaile sintaktikoa esan dezakegu. Bestalde, *parsing* terminoaren bitartez analisi sintaktikoari egiten zaio erreferentzia.

Ondoren sintaxiaren prozesamenduan dauden joera nagusien ikuspegi orokorra aurkeztuko dugu, gure hurbilpenean aukeraturiko formalismoa bera hobeki ulertzearren.

2.1 Deskribapen linguistikoetan oinarritutako analizatzaileak

Deskribapen hauek teoria gramatikaletan dute oinarria, esate baterako: *Lexical Functional Grammar* (LFG), *Generalized Structure Grammar* (GPSG), *Head Phrase Structure Grammar* (HPSG), *Government and Binding* (GB), etab. Linguistikoki interesgarrienak diren esaldiez arduratzen dira batez ere, eta ez horrenbeste testu errealez. Lantzen dituzten errepresentazio-ereduak handinahikoak izaten dira

[1] http://www.ixa.ehu.es

normalean. Honelako deskribapenetan oinarritutako *parser*ek kale egingo dute, maiz, egunkarietan edo testu teknikoetan aurki daitezkeen esaldien aurrean. Beste arazo bat da ezagutzen dituzten esaldietarako hainbat interpretazio ematen dituztela, eta hauetako zein den egokiena erabakitzeke geratzen dela.

Hala ere, esan beharra dago badirela aplikazioak testu errealei begira garatutakoak, eta oinarri gisa honelako teoriak hartu izan dituztenak. Esaterako, *Xerox Linguistic Enviroment -ak (*XLE) (Kaplan eta Newman, 1997) erraztasunak ematen ditu LFGn oinarritutako estaldura zabaleko gramatikak eraikitzeko, informazio lexikala eta morfologikoa kanpoko iturrietatik eskuratuz.

2.2 Probabilitatean oinarritutako teknikak

Hurbilpen probabilistikoan (Black eta beste, 1993), aurreko joeran gramatikariek egiten zituzten lanak modu automatikoan egiten dira. Eta hurbilpen honek indar handia hartu du azken hamarkadan. Sinplifikatzearren estatistika hutsa erabiltzen dute. Hala ere, ez da egia osoa aldez aurretik ezagutza linguistiko oinarrizkorik ez dagoenik. Kontua da, gramatikak garatzeko orduan eskuzko lan gramatikal minimoa egiten dela. Hau da, ezagumendu linguistikoa probabilitateen bidez ateratzen da etiketaturiko corpusetatik abiatuz (corpus hauei *treebank* edo *parse bank* deitzen zaie). Estrategia hori jarraituz, sistema probabilistiko gehienetan azaleko analisia egiten da, etiketatzaileetan adibidez: hitz bakoitzaren kategoria sintaktikoa igarri behar da. Ildo horri jarraituz etiketatzileek nolabaiteko muga gaindiezinak dituzte. Esaterako, %95-97 inguruko neurriak (Voutilainen 1994, Brill eta Wu 1998) agertu dira zenbait hizkuntzatarako. Sintaxi osoari begira, zenbaki horiek onartu ezinak dira, esaldi askotan errore bat egotea suposatuko lukeelako. Estatistika hutsa erabiltzeak arazoak izan ditu testuinguru mugatuetan gertatzen ez diren fenomenoak tratatzeko. Eta, bestalde, teknika horien bidez erdietsitako analisiek beste oztopo bat dute, emaitzak horiek linguistikoki interpretatzea ez baita erraza.

2.3 Probabilitateetan eta gramatikan oinarrituriko hurbilpenak konbinatzen dituztenak

Linguistek idatzitako gramatiketan, maila altuko gertaera linguistikoak deskribatu dira gehienbat, sintagmak zein esaldi osoak konbinatzeko, baina arreta gutxiago eskaini zaio esaldi errealetan agertzen den zenbait fenomenori, egitura jakin baten maiztasuna

kasu. Horregatik metodo probabilistikoak eta ezagutza linguistikoa lotzeko saioak egin dira, bakoitzaren abantailak biltzeko asmoz. Oro har, hurbilpen honetan gramatikaren erregelak linguistek idazten dituzte, baina hauen aplikazioa ezagutza estatistikoan oinarritzen da. Ezagutza estatistiko hau etiketaturiko corpus edo *parser*ak probatzeko corpus handietatik atera da. *Parsing*a lan bikoitza legez ulertuko dute:

1) *Parser*ak posible diren aukera guztiak emango ditu.

2) Aukeretarik zein den hoberena edo egokiena erabakiko dute.

Adibidez *IBM/Lancaster Approach* (Black, Garside eta Leech (arg.), 1993).

Hurrengo puntuan ikusiko dugu zehatzago probabilitateak eta gramatikak baliatzen dituen *CG* formalismoa.

2.3.1 Constraint Grammar (CG) formalismoaren filosofia

Formalismo hau desanbiguazioan, morfologikoan zein sintaktikoan oinarritzen da. Desanbiguazio-prozesu hori murriztapen-erregela multzo baten bitartez egiten da, zeinak, ondoren azalduko diren beste elementu batzuekin batera, gramatika osatzen duten.

80. hamarkadan Fred Karlsson-ek sortua da eta hona hemen gramatikaren printzipio nagusiak (Karlsson et al., 95)[2]:

- CG formalismoa hizkuntzarekiko independentea da, analisi morfologikoan oinarritzen da eta helburua edozein testu analizatzea du.
- Erregelen bidez, analisi morfologikoaren eta funtzio sintaktikoen desanbiguazioa aurrera eramaten ditu CGk. Erregela horiek, testuinguru jakin batean zuzenak/egokiak ez diren ahalik eta interpretazio[3] gehienak kentzen dituzte. Horregatik esaten da murriztailea dela.

Formalismoaren filosofia azaltzen duten sententzia nagusiak hauek dira:

1) Helburua: analizatzailearen helbururik behinena anbiguotasuna ebaztea da, forma baten interpretazioen artean, analisi zuzenak/egokiak aukeratzea. Honekin batera, aurretik morfologikoki analizatutako testu baten analisi sintaktiko azalekoa ematea da helburua.

---

[2] Liburu honetan azaltzen den eskema jarraitu dugu CGPren ezaugarri nagusiak aipatzeko.

[3] Hitz-forma baten analisi morfologiko posible bakoitzari *interpretazio* deitzen zaio.

2) *Parser*rei zuzendutako gramatiken helburua (eta CG horietako bat da) ez da esaldien gramatikaltasuna adieraztea, analisi guztiei irtenbide bat ematea baizik, ahalik eta interpretazio oker/ez-egoki gehienak kentzea. Aztertzeko dagoen testuko elementu orori emango dio irtenbideren bat; alde horretatik sendoa dela esango dugu.

3) Hizkuntzarekiko eta programazio-kodeketarekiko independentea da.

4) Gramatika eta lexikoiak testu-motari egokitzen ahal zaizkio.

6) Gramatikaren oinarria murriztapen-erregelek osatzen dute, baina ezaugarri probabilistikoa duten elementuak ere erabiltzen dira; erregelek irtenbide egoki bat ematen ez dutenean, gramatikari sendotasuna emanez.

7) Analisi morfologikoa eta lexikoa dira oinarri eta mami.

8) Hiru zatitan laburbil daiteke gramatikaren zeregina:

- testuinguruari lotutako desanbiguazio morfosintaktikoa;

- esaldi barruko eta esaldien arteko mugen esleipena;

- azaleko funtzio sintaktikoen esleipena.

9) CG murriztailea da, anbiguotasun morfologikoa eta sintaktikoa ebaztea duelako helburu. Hau da, testuinguru jakin bati ez dago(z)kion analisia(k) ebaztea.

10) Anbiguotasuna hitz mailan adierazten da. Beraz, analisi-unitatea hitza da

11) Analisi sintaktikoak hitz bakoitzari funtzio bat esleituko dio: lehenik datu-basetik funtzio sintaktikorik gabe datozen hitzak osatuko dira eta ondoren desanbiguazio sintaktikoari ekingo zaio. Halaber, hitzen arteko erlazioaren berri ere emango du. Hala ere, analisia azalekoa eta lineala da, hau da, arbolarik edo egitura hierarkikorik ez da zuzenean sortzen.

12) Erregelen oinarriak hauexek dira: gramatikaren eta corpusen azterketek ematen duten informazioa.

13) Abstrakzio teorikoari dagokionez, berriz, CGP osatzen duten erregelen maila baxuagoa dela esan daiteke, sintaxi formalaz aritzen diren beste teorien

erregelekin konparatzen badugu; esate baterako: Government and Binding teoriakoekin edota Generalized Phrase Structural Grammar-ekoekin (GPSG).

14) Erregelak bata bestearekiko independenteak dira.

15) Aurreprozesua oso inportantea da. Zer lan egin behar du aurreprozesadoreak? tokenizazioa, hots, item-ezagutzailea[4], paragrafoen mugen esleipena, hitz anitzeko unitateen ezagutzea, etab.

16) SGMLko kodeketa estandarra erabil daiteke erregeletan.

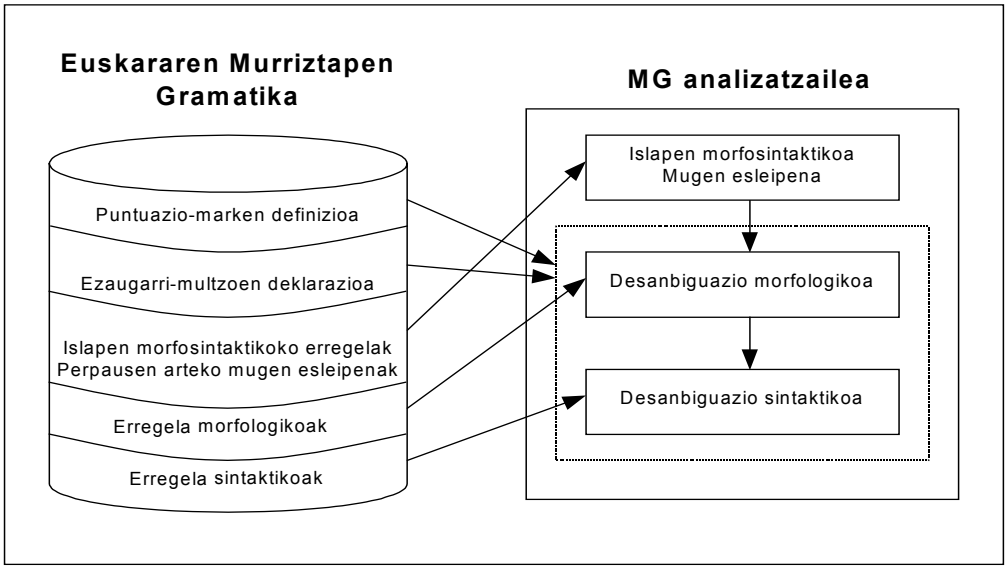
3. Euskararako desanbiguazio gramatika: EUSMG

Aipatu ezaugarri orokorrak kontuan harturik, euskararen tratamendu sintaktikoari ekiteko CG formalismoa aukeratu dugu. Erabakiaren egokitasuna teorikoki frogatua gelditu bada ere, euskararako egin dugun aplikazioa ikusiko dugu ondoren. Honetan, printzipio orokorrak jarraitu baditugu ere, erabaki partikular batzuk hartu behar izan baititugu.

Gramatika honek sei atal dauzka eta *parser*rak atal guztiak bertan daudela egiaztatzen du, hutsik badaude ere (*NIL* markarekin).

Atalen zehaztapenak azalduko ditugu segidan. Aurretik, ordea, III.1 irudian gramatikaren arkitektura modu argiagoan ikusten ahal da.

[4] Tokenizazioa token edo item ezagutzea da. Hau da, analizatzaile morfologikoak sarrera gisa erabiliko dituen unitateetan bereiztea da. Unitate horiek, esate baterako, lexikoak eta puntuazio-markak izan daitezke.



III.1 irudia.- MG analizatzailearen arkitektura.

3.1 Esaldiak mozten dituzten puntuazio-marken definizioa

Puntuekin batera, puntu eta koma, galdera-ikurra, etab. izango dira hemen definituko direnak: DELIMITERS = "<$.>" "<$;>" "<$?>" "<$!>".

3.2 Ezaugarri-multzoen deklarazioa

Erregelen testuinguruak adierazteko elementu gramatikalak erabiltzen dira (kategoria, kasua, mugatasuna, etab.). Askotan antzeko ezaugarriak dituzten elementuak multzoka daitezke eta erregeletan erabili ahal izateko aurretik definitu behar dira. Multzo horiek atal honetan definitzen dira.

3.3 Funtzio sintaktikoen islapen-erregelak (morphosyntactic mappings)

Islapenaren funtzioa informazioa gehitzearena da. Normalean ezaugarri morfologiko eta sintaktikoen arteko harremanak islapen-erregelen bidez adierazten dira. Islapenak interpretazio morfologiko bati funtzio sintaktiko bat esleitzen dio. Datu-basetik ez datozen funtzio sintaktikoak esleitzeko erabiliko dira. 83 islapen-erregela daude gramatikan.

Islapen-erregelek honako formatua dute:

<eragilea, etiketa sintaktikoa, *TARGET* hitza, helburua, *IF* hitza, testuinguruko baldintzak>

- `MAP (@-JADNAG) TARGET (ADI) IF (0 BURU) (1 ADL);`

  `# Adibidea: Basoan biziaz aberetu EGIN dira`

Posizioa zenbaki baten bitartez adierazten da (ik. beherago desanbiguatzeko erregelak azaltzean honi buruz esaten dena). Zenbakia positiboa edo negatiboa izan daiteke, eskuina edo ezkerra adierazteko. "0" posizioa berriz, aztertzen ari garen hitzari berari egiten dio erreferentzia.

Hau guztia ikusita, honela parafrasea daiteke goian jarri dugun islapen-erregela: *islatu @-JADNAG funtzio sintaktikoa ADI(tz) ezaugarria duten formei, baldin eta forma bera burutua bada eta eskuinetara aditz laguntzailerik badu.*

3.4 Perpausen arteko mugen esleipena

Goian azaldu diren islapen-erregelen bidez eramaten da aurrera perpausen arteko mugen esleipena. MUGA hitza da agintzen den kasuan esleitzen den ezaugarria.

3.5 Desanbiguatzeko erregelak

Erregela-mota bera erabiltzen da desanbiguazio morfosintaktiko nahiz sintaktikorako. Erregelek fenomeno orokorrak eta partikularrak tratatzen dituzte. Eremu hauez osatuak daude:

<(domeinua) eragiketa, helburu-interpretazioa, *IF* hitza, tratatzen ari den hitzaren baldintzak, testuinguruko baldintzak>

Ikus ditzagun elementuok adibide honetan:

- `(1) REMOVE (ADI) IF (0 ADJ) (NOT -2 DET)(-1 ZERO) (1 DET)`

  `# Adibidea: Bizitoki JAKIN bat ez zutela ...`

Erregela hau horrela parafrasea daiteke:

Ezabatu aditzaren interpretazioa, baldin eta tratatzen ari garen forma adjektiboa ere bada eta ezkerretara, bi hitzetara, ez badu determinatzailerik; ezkerretara hitz batera morfema gabeko elementurik badu eta eskuinetara, posizio batera, determinatzailerik badu.

Desanbiguazio-erregelak ataletan bana daitezke ziurtasun-mailaren arabera. Euskararako egin dugun gramatikan lau atal bereizi dira: lehenengoak erregela morfosintaktiko ziurrenak jasotzen ditu; bigarrenak, ziurtasun-maila txikiagoko erregela morfosintaktikoak; hirugarrenak, ordea, erregela sintaktiko ziurrak jasotzen ditu eta laugarrenak, azkenik, erregela sintaktiko ez-ziurrak eta oro har, behin behinekoak direnak. Gainera, ataletan banatze honek nolabaiteko ordena jartzen du gramatikan, bestela, horrenbeste erregelarekin nekez lortuko litzatekeena (Sánchez, 1997).

Euskararako gramatikan 1.113 desanbiguazio-erregela daude: lehenengo sekzioan 672; bigarrenean 45; hirugarrenean 289 eta laugarrenean 107.

3.6 END

Azken atala da eta END hitzak osatzen du.

4. EUSMGren oinarriak eta iturriak

4.1 Euskararen Datu-Base Lexikala: EDBL

Prozesadore morfologikoaren muina den lexikoa datu-base batean antolatuta dago, EDBLn (Alegria et al., 1997). Ezinbestekoa da eskala errealeko proiektu aplikatu bati ekitean datu linguistikoak datu-base batean taxuz egituratuak izatea.

Hasieran, bi mailatako formalismoaren bidez morfologia tratatzeko sortu bazen ere, EDBL gaur egun euskararen tratamendu automatikorako datu-base lexiko orokorra da, Lengoaia Naturalaren Prozesamenduaren arlo askotan funtsezko ezagumendu-oinarria den aldetik. Horrexegatik, mota guztietako informazioa biltzen da bertan: morfologikoa eta sintaktikoa, semantikoa oraindik ez badago ere (adieren arteko banaketa, alegia), homografo-identifikatzaileak egoteak nolabaiteko hurbilpena adierazten du (Agirre et al., 1994; Aduriz et al., 1998a).

Hala ere, inportanteena informazio lexikala da. Laurogei hamar mila sarreratik gora ditu orain EDBLk (80.000 inguru) eta kopuru horri arrazoizkoa deritzogu, izan ere beste hizkuntzetako aplikazioetan erabiltzen diren lexikoiekin konparatzen badugu sarrera-kopuru antzerakoekin lan egiten dutela konturatuko gara.

EDBL osatzeko erabili den deskripzio lexikografiko-linguistikoari gagozkiola, estandar lexikografikoen eta arau linguistikoen aginduetara makurtu garela esan behar da, aplikazio konputazionalak ekar ditzakeen eskakizunetara baino. Orokorra den aldetik, datu-basea aplikazio bat baino gehiagotara zuzenduta dago, eta beraz, ezin makur gaitezke bakoitzak eskatzen duen berezitasunetara[5].

Sarrera bakoitzari dagokion informazioa eremuetan biltzen da, esate baterako: forma kanonikoa, bi mailatako forma (morfologian erabiliko den ereduari egokitua), itsats dakizkiokeen morfemei buruzko informazioa (jarraitze-klasea), homografo identifikadorea, iturburua, adibidea(k), etab.

4.2 Corpusak

CG formalismoa korronte enpirikoaren barruan kokatzen da, analizatzaile honen oinarrizko puntuetako bat corpus errealekiko joera izanik (Karlsson et al., 1995:17). Joera hori bi puntutan zehaztu daiteke: corpusak, gramatikarekin batera, informazio linguistikoaren iturri dira, desanbiguazio-erregelen sorkuntza-prozesuan batez ere; eta bestetik, aplikazioetarako eta tresnetarako probaleku ezinbestekoak dira, sistemen zehaztasuna neurtzeko.

Goian aipatutako erabilerez gain, hurbilpen estatistikoetan derrigorrezko tresna bilakatu dira. Esate baterako, etiketatze-lanetan erabiltzen diren eredu markoviarretan eta Bayes-en ereduetan, corpusak ezagumendu-iturri gisa erabiltzen dira.

Euskaltzaindiaren bermea duten bi corpus daude euskaraz, historikoa –Orotariko Euskal Hiztegia (OEH)– eta egungo euskararen corpus erreferentziala (XX. mendea) –Egungo Euskararen Bilketa-lan Sistematikoa (EEBS[6])– eta biak Euskaltzaindiko lexiko-finkapenerako proiektuaren barruan kokatzen dira[7].

---

5 Gure kasuan, datu-basetik murriz zitekeen anbiguotasuna, potentzialki anbiguo izango diren formei kategoria jakin bat ezarriz, homografo diren adjektibo/izenen kasuan, esate baterako (ENCGGn egin duten bezala: ik. Karlsson et al. (1995:94-95)).

6 EEBS corpusa UZEIn lantzen ari dira. UZEI eta Donostiako Informatika Fakultatearen arteko harremana IXA taldearen hasieratik dago linguistika konputazionaleko ikerketaren barruan.

7 "Si bien ambos corpus recogen documentos escritos, la gran diferencia radica en que el histórico recoge obras completas, exhaustivamente despojadas, mientras que el actual es estadístico: interesa más la variedad léxica que la calidad de las obras. De ahí el nombre de la referencia, ya que muestra el euskera que se utiliza hoy." (Urkia, 1998).

EEBS (Urkia & Sagarna, 1991) idatzizkoaren gaineko corpus estatistikoa da, eta testu-motei begira, orekatua[8]. Urtero eguneratzen da corpus hau eta egun 4.000.000 inguru hitz lematizatu ditu.

Guk EEBS corpus orekatuaren zati batzuk erabili ditugu gure lanerako. Erregelak egiteko prozesuan zati bat erabili dugu eta beste bat erregelen zehaztasuna neurtzeko.

Ikusten denez, euskaraz (eta euskara bezalako hizkuntza minorizatu eta minoritarioentzat) oso corpus gutxi dago eskura. Ingeleserako, berriz, baliabide ugari daude: Brown Corpus edo Penn Treebank modukoak. Corpusaren tamaina oso garrantzizkoa da, izan ere, gertaera linguistikoen deskribapen zabala izateko, handia behar du derrigorrez. Horrela, corpusak ugaltzen direnean, berauetan oinarritutako ikerketok sakontasunean eta zehaztasunean irabaziko dute.

4.3 Analizatzaile morfologikoa

CG formalismoaren filosofiaren oinarrizko puntuei buruz aritu garenean, hauxe esaten genuen: a*nalisi morfologikoa eta lexikoa dira oinarri eta mami*. EUSMGren analisi morfologikoa analizatzaileak ematen du, eta lexikoa EDBL datu-basean dago bildua.

EDBL datu-basearen ezaugarriak ikusi ditugu aurreko puntuan. Ikus dezagun orain analizatzailearenak.

Euskararen morfologiaren tratamendu konputazionalari ekin zitzaionean, ereduen artean euskararen ezaugarriak adierazteko gehien balio zezakeenen artean bilatu zen. Aukera bat baino gehiago zegoen orduan (ikus Alegria (1995) eta Urkia (1997)) eta proba batzuen ondoren, azkenik Koskenniemi-ren bi mailatako morfologia (Koskenniemi, 1983).

Bi mailatako ereduaren ezaugarri garrantzizkoenak aipatuko ditugu orain laburki Alegriaren eta Urkiaren lanetan oinarrituta (Alegria, 1995; Urkia, 1997): a) eredu orokorra da, edozein hizkuntzetan aplikatzen ahal dena, ezagutza linguistikoa eta algoritmikoa bereizten dituelako; b) hitzen analisi morfologikorako zein sintesirako balio du; c) azaleko eta sakoneko sistema lexikoak ongi bereizita daude, horregatik ez dago alomorfoak erabili beharrik; d) fonologia sortzaileko berridazketa erregelen ordez

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

erregela paraleloak erabiltzen ditu, sistema kontzeptualki zein konputazionalki errazago bihurtuz.

Bi mailatako formalismoaren osagai garrantzitsuak lexiko-sistema, morfotaktika eta erregela morfofonologikoak dira.

Aipatzekoak diren beste puntu batzuk, analizatzailearen sendotasuna eta malgutasuna dira, Koskenniemi-ren bi mailatako morfologiaren sistemari egin zitzaizkion hiru hobekuntzen ondorio direnak. Batetik, erabiltzailearen lexikoa integratzeko aukera eman zen; bestetik, analizatzaileari sendotasuna emateko, bi mailatako eredu bera erabiliz, forma ez-estandarrak tratatu ziren, dagokion estandarrari lotua dagoena[9]; eta lexikorik gabeko analisia fonologiarako ez ezik, testuen analisirako ere erabili zen.

Izan ere, analizatzaileak ezagutzen duen hitz bakoitzerako, aurretik daukan informazio guztia ekartzen du bueltan, morfema bakoitza zatitua agertzen dela. Askotan, gainera, anbiguoa den analisi-multzo bat osatzen da, adibide honetan ikusten den bezala:

```
((forma "bide")
 ((analisi 1)
  ((lema "bide")((SAR bide)(KAT IZE)(AZP ARR))))
 ((analisi 2)
  ((lema "bide")((SAR bide)(KAT PRT)(MDL ZIU))))
 ((analisi 3)
  ((lema "bide")((SAR bide)(KAT IZE)(AZP ARR)))
  ((morf "0")((SAR 0)(KAT DEK)(KAS ABS)(MUG MG)(FS1 @OBJ)(FS2 @SUBJ)(FS3
@PRED))))
```

Analisi-aukera bakoitzean, informazio morfologikoa (kategoria, azpikategoria, kasu-mugatasuna, ...), eta sintaktikoa ere ager daitezke (funtzio sintaktikoak). Hauetako funtzio sintaktiko batzuk hasieratik daude definituta datu-basean; beste batzuk, ordea, islapen-erregelen bidez esleitzen dira, aurrerago ikusiko dugun bezala.

Ikusten denez, hizkuntzaren deskripzioan datza bueltan etorriko den anbiguotasunaren izaera, deskripzio horretan hartutako erabaki eta irizpide jakin batzuen ondorio baita. Deskribapen linguistikoak baldintzatzen du, erremediorik gabe, ondorengo emaitza. Horregatik, bai analizatzailearen definizioan, eta bai datu-baseko lexikoa lantzerakoan, fin jokatzea komeni da.

---

9 Bi mailatako ereduari jarraituz egin den euskararen morfologiaren deskribapen orokorra euskara estandarrari dagokio. Gerora egin zen forma ez-estandarren deskribapen hau.

Beraz, analizatzaile morfologikoaren emaitzaren gainean egingo dugu lan. Hori izango da gure lanaren oinarria edo sarrera, testu bat analizatu ondoren azaleratzen delako anbiguotasunaren arazoa.

5. Anbiguotasunaren azterketa

5.1 Aztergaia mugatuz

Hizkuntzaren edozein eremutan *anbiguotasunak* komunikazioaren alterazioa ekartzen du berarekin, izan ere, horrelakoetan, esaldi bat, hitz bat, etab. modu batean baino gehiagotan interpretatzen ahal denean gertatzen da. Izatez da hizkuntza anbiguo.

> "Que la anbigüedad es connatural al lenguaje común –a lo que llamamos lengua a secas– en cualquiera de sus variadísimas especies es un hecho tan conocido que no hace falta apelar a refinadas técnicas dialécticas y retóricas para traer a los incrédulos al buen camino. (...). La ambigüedad es, sin lugar a dudas, uno de los universales más patentes del lenguaje natural (...)." (Michelena, 1972).

Gai honi buruz asko idatzi da, dudarik gabe, eta ikuspuntu diferenteetatik planteatu da arazoa. Izan ere, oso eremu zabala denez, mota askotako alterazio linguistikoak egon daitezke definizio horren barruan.

Alderdi konputazionaletik, *parsing*, edo analisi sintaktikoarekin lotzen da arazoa gehienbat. Aurretik egindako analisi morfologiko-morfosintaktikotik datorren informazioak sortzen duen anbiguotasuna da tratatzen dena.

Hizkuntzaren azterketan anbiguotasunaren arazoa eremu guztietan agertzen ahal bazaigu ere, are latz eta korapilatsuago bihurtzen da ordenadore bidezko azterketetan. Izan ere, aurretik ordenadorean pilatutako informazio guztia (lexikoa nahiz morfosintaktikoa) prozesatu, eta ematen digu inongo mugarik gabe, askotan pentsatu ezinezko emaitzak sortuz[10]. Honek egiten du anbiguotasuna LPNeko arazo nagusietako bat izatea, batez ere analisi sintaktikoaren mailan.

Gure aztergaia, beraz, alderdi konputazionalean sortzen den anbiguotasunaren eremuan kokatzen da eta bestelako bereizketetan sartu aurretik, azaleko sintaxiaren esparruan mugituko garela esan beharra dago. Horrek baldintzatuko du anbiguotasuna tratatzeko modua, inondik ere hurbilpen teorikoak egiten duenarekin konpara ezin daitekeena.

[10] Nork pentsatuko luke lehenengo kolpetik, aditzaz gain *zituen* forma *zitu* izenaren genitibo plurala dela?

Horrela, anbiguotasun semantiko eta pragmatikoa alde batera utziko ditugu eta gure aztergaia anbiguotasun gramatikal lokala izango litzateke (morfologikoa ere deitua) eta konkretuki, morfosintaxiari eta funtzio sintaktikoen alorrei dagokiena. Beti ere kontuan hartuz hitz mailan mugitzen garela (zuriunetik zuriunera doan segida), eta berau desanbiguatzeko testuinguru hurbila hartzen dugula aintzakotzat. Horrela, kategoria mailako anbiguotasunak tratatuko ditugu (*omen* hitza partikula, izen arrunta eta aditza izan daitekeelako), morfosintaxi mailakoak (*etxeak* forma absolutibo mugatu plurala eta ergatibo singularra izan daitekeelako) eta sintaxi mailakoak ere, *etxeak* hitza subjektu, objektu eta predikatibo izan daitekeelako.

Batek baino gehiagok pentsa dezake forma horiek testuinguru jakinetan ez direla anbiguo. Hori da hain zuzen ere gramatika honen eginbeharra, alegia, testuinguru jakinetan dagokion kategoria, kasua edo funtzioa aukeratzea. Pentsatu behar da formak modu isolatuan analizatzen direla, testuingurua kontuan hartu gabe eta urrats honetan desanbiguazioa egin behar da, testuingurua kontuan hartuta.

5.2 Anbiguotasun-motak

Guretzat maila honetan anbiguotasuna nola ulertzen den ikusi ondoren, anbiguotasun morfosintaktikoa lau multzotan sailkatzea proposatzen dugu:

1. Kategoriala: adjektibo eta adberbioen artean dagoen anbiguotasuna; aditzoin, adjektibo eta adberbioen artekoa (adibidez *bizkor* formak lau analisi ditu: lehenengo analisian adberbioa dugu, bigarren eta laugarrenak adjektiboak dira eta hirugarrena aditzoina); aditz laguntzaileen eta trinkoen artekoa; etab.
2. Deklinabide-atzizkiei dagokiena: absolutibo mugatu plurala eta ergatibo mugatu singularraren artean dagoen anbiguotasuna adibide bat litzateke; *–ko* atzizkiaren anbiguotasuna: leku-genitiboa, deskribatzailea, banatzailea; etab.
3. Mendeko atzizki/aurrizkien ingurukoa: *-la* atzizkiaren edo *bait-* aurrizkiaren arteko balioen arteko anbiguotasuna sartuko litzateke multzo honetan, esate baterako.
4. Aditzen aspektu eta modu-denborari dagokiona: adibidez, zenbait aditzetan, aditzoinaren eta partizipioaren arteko anbiguotasuna aurkituko dugu hemen (*egon*, *joan*, etab.).

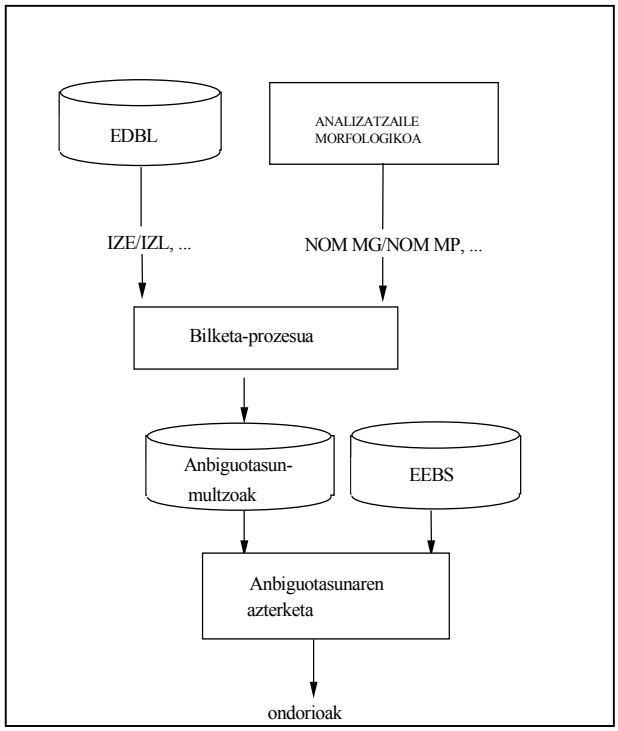
5.3 Erabilitako metodologia

Desanbiguazioa aurrera eramateko urratsez arituko gara metodologiari eskainitako atal honetan.

Gure helburua desanbiguazio orokorra egitea da edozein testu tratatu ahal izateko. Horregatik, anbiguotasunaren arazoaren azterketa orokorra egin dugu, fenomeno linguistiko nabarmenenak kontuan hartuz. Lan honetan lehenengo iturria analizatzaile morfologikoa dugu, EDBLrekin batera. Bi iturri horietatik lehenengo anbiguotasun-multzoak ateratzen dira.

Ondoren, multzo horien benetako agerpenaren berri emateko helburuarekin, corpusaren gaineko azterketak egiten dira, eta ondorioz arazo bakoitzaren tamaina eta garrantzi erreala ezagutzen da.

Grafikoki azaltzeko, hau da anbiguotasuna detektatzeko jarraitzen den prozeduraren eskema:



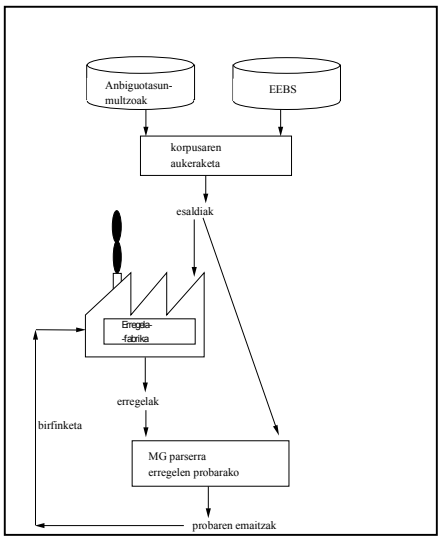
V.1 irudia.- Anbiguotasuna detektatzeko prozesua.

Beraz, testu errealen gaineko azterketak egitea, anbiguotasunaren emankortasuna aztertzeko oso baliagarria izateaz gain, fenomenoaren orokortasuna ikusteko ere garrantzitsua izango da. Era berean, gero erregeletara itzuliko diren testuinguruaren ezaugarriak ere azaltzen hasten dira.

Ondoren, corpusaren eskuzko desanbiguazioa egiten da. Eskuz desanbiguatutako corpusa prest dagoelarik, aurretik detektatutako anbiguotasun-arazoak konpontzeko desanbiguazio-erregelak sortzen dira. Erregelak osatzeko prozesuan, 14.000 hitzeko

corpus bat erabili da. Behin eta berriro probatzen dira erregelak, erroreak konponduz eta informazioa gehituz, hau da birfinduz, testuak zuzen desanbiguatuak egon arte.

Irudi honetan ikusiko dugu desanbiguatze-prozesua:



V.2 irudia.- *Desanbiguazio-prozesua.*

5.4 Anbiguotasunaren neurketa

Deskribatu berri dugun anbiguotasun morfosintaktikoak, estatistikoki zer tamaina hartzen duen ikusiko dugu ondoren. Bi multzotan bereziko ditugu datuok: alde batetik anbiguotasun kategorialari dagozkion kopuruak eskainiko ditugu, eta ondoren, analizatzaile morfosintaktikoak ematen duen gainontzeko informazio guztia kontuan hartuko da emaitzak ateratzeko. Hau da, deskripzioan zehaztu diren beste hiru anbiguotasun-motak ere sartuz (morfema ez-askeena, menderagailuak barne eta aditzena).

Horretarako, 14.000 testu-hitzeko corpusa hartu dugu oinarri gisa. Benetako corpus librea denez, forma estandarrez gain, hitz ezezagunak agertzen dira testuetan (EDBLn sartuta ez daudenak[11]), baita forma estandarren aldaerak ere (forma dialektalak, esate baterako) (Alegria, 1995; Ezeiza, 1997). Analizatzaileak forma guztiei ematen die irtenbideren bat: analisi estandarraren bidez, aldaeren analisiaren bidez eta lexikorik gabeko analisiaren bidez (Ezeiza, 1997).

11 "Analizatzen ez diren hitzen erdien ingurua ez dira ezagutzen dagokion lexikoan ez dagoelako (…). Lexikoan ez egotearen arrazoiak desberdinak izan arren, (…) askotan ezin da edo oso zaila gertatzen da lema guzti horiek lexiko orokorrean egotea, testuaren edota idazlearen erabilpen espezifikoak baitira askotan." (Alegria, 1995).

	Anb. kategoriala	Anb. Morfosintaktiko osoa
Estandarrak	%46.34	%80.09
Aldaerak	%32.89	%81.25
Lexikorik gabe	%57.95	%95.88
Batez beste	**%37.80**	**%65.75**

V.3 taula.- *Anbiguotasunaren neurketa*

Taulan ikus daitekeenez, forma ez-estandarren interpretazio-kopurua erabilera estandarrena baino handiagoa da.

Anbiguotasun kategorialean 20 etiketa orokor sartzen dira: 17 orokor eta beste hiru etiketa, elipsia bezalako kasu bereziak etiketatzeko. Hori dela eta, testuan aurkitzen diren hitzen %37.80 anbiguoa da, hau da, hitz bakoitzak analisi-lerro bat izan beharrean bat eta erdi baino gehixeago ditu.

Anbiguotasun morfosintaktiko osoa kontuan hartzen badugu, ordea, kopurua bikoiztu egiten da, hitzen %65.75a anbiguo izatera iritsiz, hau da, forma bakoitzak 2,81 interpretazio ditu. Izan ere, analizatzailearen informazio osoa dago hemen, kasua, numeroa, mugatasuna, aditzen kasuan modua, denbora, etab.

Anbiguotasun kategoriala eta morfosintaktiko osoaren arteko diferentzia handi hau ez da bakarrik euskararen kasuan gertatzen. Oro har, hizkuntza eranskarietan eta flexiboetan, morfologia aberatsagoa dutenez, kategoriala ez den anbiguotasun-mota hau gehiago gertatzen da, ingelesa edo gaztelania bezalako hizkuntzetan baino.

Euskaraz anbiguotasun morfosintaktikoaren arazoa handia dela garbiago ikusten da beste hizkuntza batzuetako datuekin konparatzen badugu (Karlsson et al., 1995:23). Esate baterako, finlandieraz %11,2koa da. Suedieraz, hebraieraz bezala, handiagoa da, %60a bietan. Espainolak %43 ingurukoa du eta ingelesak %35ekoa[12].

6. Erregela morfologikoak eta sintaktikoak.

Hizketa bizian, anbiguotasun baten aurrean, hiztunak (eta hartzaileak) baditu bideak sortzen diren anbiguotasunak hausteko, hala nola, azentuazioa, testuingurua, etab.

[12] Zaila da beste hizkuntzetan dagoen anbiguotasunarekin konparatzea, kasu bakoitzean oinarrizko etiketak erabat desberdinak erabiltzen direlako eta oinarri-testuak ere izaera desberdinekoak dira. Erabat konparagarriak izango liratekeen datuak erkatzeko, oinarri-testuak eta etiketa-sistema antzekoa erabili beharko lirateke (Márquez, 1999).

Hizketa bizian ditugun baliabide horien zeregina egitera datoz erregela hauek, eta guk hiztunok mekanikoki egiten dugun mezuen desanbiguazioa egitera.

Zer ulertzen dugu *desanbiguazio* hitzarekin? Forma batek eduki ditzakeen ulertzeko aukeren artean bat, egokia, hartzea:

> "In Constraint Grammar, disambiguation does not mean "bring out all alternatives" but rather "pick the appropriate alternative(s) by discarding one or more inappropriate ones". The Constraint Grammar notion of morphological disambiguation is functionally similar to the notion "homograph separation" (...)." (Karlsson et al., 1995).

Helburu horretan erabiltzen diren desanbiguazio-erregelok gramatika batean antolatuta daude. Gramatikaren diseinuari buruzko xehetasunak hala nola erregelen sintaxiari buruzkoak, gorago aztertu ditugu. Ondoren, beraz, ez dugu gramatika bere gordinean azalduko[13].

Honekin batera, murriztapen-erregelak gramatika-erregelen ondorio direnez, anbiguotasun-kasu bakoitzean, erregela-multzo batetik eratortzen diren gramatika-erregelen antzeko printzipioak eratorriko ditugu.

6.1 Erregelak eta printzipioak

Gramatikaren atalean 1.113 erregela daude. Horiek guztiak kontuan harturik idatzi dira printzipio teorikoak. Gehienetan erregelak orokorrak izango dira, anbiguotasun-talde osoari erreferentzia eginez. Beste batzuetan, ordea, partikularrak izango dira, anbiguotasun-talde horretako hitz konkretuei erreferentzia egiten baitiete.

Desanbiguazio-erregelen aurkezpena adibide batekin egingo dugu: printzipio nagusi bat azalduko dugu eta ondoren, printzipio teoriko horri dagozkion erregeletatik bat ikusiko dugu. Erregelarekin batera, honen adibidea[14] agertuko da.

Goian ikusitako anbiguotasunaren adibide bat konpontzeko erregela jarriko dugu ondoren: *bizkor* forma, *bizkortu* aditzaren aditzoin izan daiteke, adjektibo eta adberbio. Hori tratatzen duen printzipio teorikoa hau dugu:

---

[13] (Aduriz et al., *2000*) barne-txostenean gramatika osoa azalduta dago.

[14] Erregelen adibideak corpus errealetatik ateratakoak dira kasurik gehienetan.

*Aditzoinek, perifrasietan, *edin eta *ezan (ADL1) motako aditz laguntzaileak hartzen dituzte ondoan, ezker-eskuin, esaldi-motaren arabera. Partizipioek berriz, izan eta *edun motakoak hartzen dituzte.*

Honi dagokion erregeletako bat honakoa dugu:

- `SELECT (ADOIN) IF (0 ADJ-ADB) (1C ADL1);`

Adibidea: azkar *bizkor* zaitezen

Erregela hau aplikatu ondoren, *bizkor* bezalako formak, alegia, anbiguotasun hori dutenak, adjektiboaren eta adberbioaren interpretazioak ezabatu ondoren, aditzoinaren analisiarekin gelditukо dira.

6.2 Emaitzak

VII.1 taulak desanbiguazio kategorialaren emaitzak erakusten dizkigu. Datuok ateratzeko, probetarako eta erregelak egiteko aurretik erabili ez den 10.000 hitzetako corpusa hartu dugu oinarritzat:

	Hitzeko analisiak	Anbiguotasuna	Interpretazio zuzeneko hitzak
Sarreran	1.50	%37.80	%100
Irteeran	1.18	%14.12	%99.12

VI.1 taula.- *Desanbiguazio kategorialaren emaitzak*

Hurrengo taulak, ordea, desanbiguazio morfosintaktiko osoaren datuak erakusten dizkigu, oinarri-corpus bera erabiliz:

	Hitzeko analisiak	Anbiguotasuna	Interpretazio zuzeneko hitzak
Sarreran	2.81	%65.75	%100
Irteeran	1.76	%33.28	%97.51

VI.2 taula.- *Desanbiguazio morfosintaktiko osoaren emaitzak*

Datuok erakusten digute desanbiguazio-gramatikaren sendotasuna eta ahalmena testu errealak tratatzerakoan. Desanbiguazio morfosintaktikoan, ia erdiraino jaitsi da anbiguotasun-tasa: sarreran %65.75etik %33.28ra. Hau da, sarreran hitzeko 2.81 analisi egotetik 1.76 egotera. Desanbiguazio-prozesu honetan, interpretazio zuzenak %97.51etan mantendu dira.

Desanbiguazio-kategorialeko datuak are hobeak dira. Anbiguotasuna, 1.50 analisi hitzeko izatetik, ia bat izatera pasa gara (1.18). Ehunekotan emanda, sarrerako anbiguotasuna %37.80koa bada, irteerakoa %14.12koan gelditzen da. Gainera, %99.12ko interpretazio zuzenak mantentzen dira.

Errore-tasa anbiguotasun kategorialean erabat onargarria dela iruditzen zaigu (0.8). Handiagoa da anbiguotasun morfosintaktiko osoa kontuan hartzen badugu. Izan ere, badakigu forma ez-estandarren agerpenak asko igotzen duela anbiguotasun-tasa eta gramatika hizkuntza estandarrerako idatzia egoteak, askotan erroreak eginarazten ditu.

Bestalde, desanbiguatu gabe geratzen den portzentaje hori, neurri handi batean informazio eskasiarengatik (azpikategorizazioaren gaia) geratuko litzateke. Horri gehitu behar zaizkio morfosintaxitik inola ere desanbiguatu ezin diren kasuak, semantikoak edo pragmatikoak direlako.

Ezin ahantz daiteke, bestetik, sarrerako anbiguotasuna egiten den deskripzio linguistikoaren baitan dagoela. Lan honen oinarri den datu-basea etengabe eguneratzen denez, askotan ez datoz bat gehitzen den anbiguotasuna eta hori desanbiguatzeko egin behar den ahalegina. Testu, deskripzio eta arazo errealekin lan egiten honek etengabeko eguneratzea dakar berarekin. Honek garbi uzten du gure lanaren ziklikotasuna eta datuetara gehiegi makurtu ezina.

7. Analisi sintaktikoa

Analisi morfosintaktikoari dagokion urratsa egin ondoren, analisi sintaktikoaren atala azalduko dugu ondokoan. Atal horrek bi pauso nagusi ditu: lehenengoan, hitz-forma orori posible dituen funtzio-etiketa[15] sintaktiko guztiak esleitzen zaizkio. Funtzio-etiketa sintaktikoak hitzei esleitzen zaizkie ezaugarri morfologikoak esleitu zaizkien modu berean. Ondorengoan, anbiguotasun sintaktikoen ebazpena burutuko da. Funtzio sintaktikoak adierazteko funtzio-etiketa sintaktikoak erabiltzen dira. Eta funtzio-etiketa sintaktiko bat baino gehiago agertzen denean hitz batean, orduan anbiguotasun sintaktikoaz hitz egingo da. Adibidez:

Txakurrak (@OBJ @SUBJ)

[15] Funtzio-etiketa sintaktikoek @ ikurra daramate aurretik

Estrategia nagusia anbiguotasuna murriztea da, aurreko puntuan azaldutako bera alegia. Horretarako, testuinguruan oinarritutako murriztapen sintaktikoak baliatzen dira. Murriztapen-erregela sintaktikoen aplikatzearen helburua hitz bakoitza funtzio-etiketa sintaktiko bakarrarekin eta zuzenarekin uztea izango da. Hala ere, esan beharra dago ebatzi ezin daitezkeen anbiguotasun sintaktikoen kasuan murriztapen-erregelek anbiguotasuna bere horretan utzi behar dutela. Esaterako, ondoko esaldian letra-molde beltzez nabarmentzen ditugun hitzak desanbiguatzeke geratu beharko lirateke: "**Txakurrak** (@OBJ @SUBJ) **egunkariak** (@OBJ @SUBJ) ahoan zekartzan".

7.1 Desanbiguatze sintaktikoa

Erregela sintaktikoen xedea hitz-forma bakoitza funtzio-etiketa sintaktiko bakarrarekin uztea da. Hori erdiesteko erregela sintaktikoak ditugu, aurreko desanbiguatze-erregelen funtzionamendu berdina dutenak. Desberdintasuna izango da ezaugarri morfosintaktikoen artean erabaki beharrean, funtzio-etiketa sintaktikoekin lan egiten dutela. Hala ere, erregela multzo hauek, morfologikoak eta sintaktikoak, elkarrekin badute harremanik. Hain zuzen ere, desanbiguatze morfosintaktikoa burutu ondoren aplikatzen baitira murriztapen-erregela sintaktikoak.

Murriztapen-erregela sintaktikoek tratatu beharreko funtzio-etiketa sintaktikoak *Euskararako murriztapen-gramatika: lehen urratsak* (Aduriz eta beste, 1996)-n azaltzen direnetan oinarritzen dira. Eta azkenak (Aduriz, 2000) tesi laneko eranskinean ikus daitezke. Edozein modutan, funtzio-etiketa sintaktikoek CG (*Constraint Grammar*) formalismoaren filosofia jarraitzen dute, hau da, funtzio-etiketadun dependentzia-sintaxia[16] (*functionally labelled dependency syntax*) dugu oinarri. CG formalismoa jarraituz esaldian hitzek dituzten funtzio sintaktikoak eta beraien arteko interdependentziak adieraziko ditugu. Hala ere, analisi sintaktikoaren emaitzak ez du sintagma-egitura espliziturik, ez baitugu zehazten sintagma tipoko osagaien hierakiarik. Hori dela eta, analisi-estrategia hau azaleko sintaxiaren baitan kokatzen da (*shallow syntax* edota *partial parsing* gisa ere ezaguna da). Eta analisi hori beti hartu ahal izango da analisi sakonago bat egiteko abiapuntu eta laguntzatzat. Orokorrean azaleko

[16] Azterketa gramatikaletan tradizio luzea du dependentzia sintaktikoaren kontzeptuak aro grekolatinotik. Hurbilagoko formalizazioak teoria sintaktikoan aipatzerakoan, Tésniere (1959), Hays (1964) eta Mel'cuk (1988) aipa daitezke, besteak beste, dependentzia-sintaxiaren suspertzaile gisa arlo teorikoan.

sintaxiaren terminoa erabiltzen da ohiko *parser*ren irteerako analisiak bezain osoak ez diren analisiez aritzerakoan.

Aipatu ditugun hitzen arteko azaleko interdependentzia horiek adierazteko funtzio-etiketa sintaktikoak bi motakoak dira: modifikatzaile edota beren buruarekiko noranzkoa adierazten dituztenak, eta nagusiak. Modifikatzaile-etiketek beren burua zein noranzkoan dagoen adierazten dute. Adibidez, @IZLG> etiketa izango dugu bere eskuinetara dagoen izen bat modifikatzen duten izenlagunentzat, eta etiketa honek adierazten du modifikatzen duen burua eskuinetara dagoela (*mendiko* (@IZLG>) *tontorretik* (@ADLG)). Funtzio sintaktiko nagusiek, hots, buruei dagozkien etiketek, perpauseko osagai tipikoak errepresentatzen dituzte, hala nola: subjektua, objektua, zehar objektua, etab. Oro har, funtzio hauek gramatika tradizionaletik oso gertu daudenak direla esan genezake. Funtzio-etiketa sintaktikoei buruzko kontsiderazio orokor horien ondorik, funtzio-etiketa sintaktikoak lau multzotan bil daitezkeela esan beharra dago: funtzio-etiketa sintaktiko nagusiak, izen-sintagma barruko dependentzia sintaktikoak adierazteko funtzio-etiketak, aditzen funtzio sintaktikoak, eta, azkenik bestelako funtzio sintaktikoak.

Gorago ikusi dugun bezala, erregela morfologiko eta sintaktikoen arteko harreman estua kontuan izanik, funtsezkoa izango da desanbiguatze morfologikoan eginiko lana ondorengo urratsean funtzio sintaktiko zuzena hautatu ahal izateko.

Ondoren aipatuko duguna, anbiguotasun sintaktikoaren arazo bat da: absolutibo singularrak, pluralak eta mugagabeak, subjektu, objektu eta predikatibo funtzioak izan ditzakete. Anbiguotasun-arazo honen aurrean desanbiguatze-erregela bat baino gehiago sortu da. Adibidez:

- `REMOVE (@OBJ)(0C ABS)(NOT *-1 NORK)(*-1 (NR_HU)) (1 (PUNT_PUNT));`

  `# Adibidea: Eta bertan agortu zen haren ODOL-JARIOA.`

Erregela hau horrela parafrasea daiteke: *Forma anbiguoa ez da objektua izango, esaldi horretan nor saileko aditza egonik (NR_HU), nork motako laguntzailerik ez badago (NORK) eta forma anbiguoaren eskuinetara puntua badago, hau da, esaldia bertan bukatzen bada.*

7.2 Oinarrizko egitura sintaktikoen ezagutzea

Lehenago aipatu dugun azaleko analisitik abiatuz, areago jo dugu zenbait oinarrizko egitura sintaktiko edota zati (sintagmak eta aditz-kateak) ezagutzeko lana burutuz. Zati[17] horiek funtzio sintaktikoen etiketek adierazten dituzten harreman sintaktikoei esker ezagutuko dira. Harreman sintaktiko horiek inplizituak dira, eta oinarrizko egitura sintaktiko horiek ezagutzerakoan agerian uzten ditugu.

Egitura sintagmatikorik adierazita ez egon arren, deskribapen linguistiko horretan inplizituki daude adieraziak elementuen arteko harremanak, eta informazio horretan oinarrituz aditz-kateak eta sintagmak atzeman ditzakegu. Horretarako, hitz-forma ororen anbiguotasun morfosintaktikoa zein sintaktikoa ebatzita egotea komeni da, zatien osaketa burutu ahal izateko. Horregatik, murriztapen sintaktikoen aplikazioaren ondorik geratzen diren etiketa sintaktikoetan oinarritzen gara. Batez ere, etiketa sintaktiko nagusi eta modifikatzaileen arteko bereizketan dago gakoa urrats honi ekiterakoan. Hori dela eta, funtzio sintaktiko nagusi eta modifikatzaileei erreparatuz dagoen anbiguotasuna txikia izatea komeni da.

Baldintza horiek kontuan izanda, lehendabizi, aditz-kateen osaketarako erregelak definitu ditugu, eta aditz-kateak bereiztu ondorik, hauen inguruan dauden sintagmak ezagutu ditugu batik bat (Arriola eta beste, 1999; Arriola, 2000). Aditz-kateak definitzeko aditzen funtzio sintaktikoak eta aditz-kateko partetzat hartu ditugun modalitatea eta egiatasuna adierazten duten partikulak baliatu ditugu. Elementu horietan oinarrituta, aditz-kate jarraiak eta gehienez ere hiru osagai dituzten aditz-kate ezjarraiak ezagutu ditugu.

Sintagmei dagokienean, sintagma-etiketa esleitzeko arau nagusia honako hau da: elementuak lotzen joatea harik eta funtzio sintaktiko nagusidun bat aurkitu arte. Hau da, modifikatzaileek ezin dute beraiek bakarrik sintagmarik osatu; modifikatzaile direnez, beti beste elementuren bati eragiten diote. Modifikatua den elementuari modifikatzailearen burua ere esaten zaio, eta funtzio sintaktiko nagusidun etiketa

[17] Lanean bereizi ditugun zatiak edota *chunk*ak sintagmak eta aditz-kateak dira. Oro har, azaleko sintaxiko lanetan esaldia *chunk*etan banatzen dela esaten da. *Chunk* terminoa Abney (1997)-ren arabera honela definitzen da: gune edo buru baten inguruan osatzen den zentzu sintaktikoa duten elementuen segida.

duenez, sintagma bukaera etiketa edota elementu bakarreko sintagmaren etiketa har dezakete.

Aditz-kateekin zein sintagmekin, zati horiek ezagutu ahal izateko, horien hasierak, bukaerak eta elementu bakarrekoak ezagutzeko etiketak[18] baliatu ditugu. Adibidez:

```
"<Haurgintzaren>" %SIH
   "haurgintza" IZE ARR DEK GEN MG AORG HAS_MAI @IZLG>
   "haurgintza" IZE ARR DEK GEN NUMS MUGM AORG HAS_MAI @IZLG>
"<nekeak>" %SIB
   "neke" IZE ARR DEK ERG NUMS MUGM @SUBJ
"<ez>" %ADIKATHAS
   "ez" ADB ADO @ADLG
"<du>"
   "*edun" ADL A1 NOR_NORK NR_HU NK_HU @+JADLAG
"<abaildu>" %ADIKATBU
   "abail" ADI SIN ASP PART DA-DU NOTDEK @-JADNAG
"<$.>"
   PUNT_PUNT
```

VII.2 Irudia.-Zatiak ezagutu ahal izateko etiketak dituen adibidea.

Goiko adibide horretan ezarri diren zatiak ezagutzeko etiketei esker bi oinarrizko zati ezagutu ahal ditugu: Haurgintzaren nekeak eta ez du abaildu.

8. Aplikazioak: Euskal Hiztegiko adibideen azterketa

Landu dugun azaleko sintaxia aplikatzerakoan Euskal Hiztegiko (Euskal Hiztegia, EH) aditzen adibideetan aditz bakoitzak inguruan dituen sintagmak eta aditz-kateak bereizi nahi ditugu. Baina, horiek aditzaren argumentu diren ala ez linguistak erabaki beharko du. Aplikazio honek bi motibazio nagusi izan ditu: (1) existitzen diren baliabide lexikalen berrerabilpena, eta (2) aditzen argumentu-egitura zehazten laguntzeko bideak eskaintzea, eta ahal den heinean azaleko argumentu-egitura erdiestea analizatzaile sintaktikoetan integratzeko. Adibideen autoritatea kontuan izanik, uste dugu hauetan gordetzen den informazioa baliagarria izan daitekeela oinarrizko informazioa erdiesteko edota corpusetan aurrerantzean azterketa sakonagoak gidatzeko.

[18]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.

Azter zitezkeen adibideen artean, aditzena aukeratzearen arrazoi nagusiak honako hauek dira:

- Eskura daitekeen informazioari erreparatuz: argumentu-egituraren berri jaso nahi dugu, informazio hau oso inportantea izango baita hurrengo aplikazioetarako, bereziki analisi sintaktikorako.
- Analisi zuzenduagoa egiteko aukera: hiztegian aditzei egokitzen zaien informazioen artean, zein tipotako laguntzaile-mota hartzen duen esaten digun informazioa dagoelako (da, du, da-du, dio, ...) eta, honetaz gain, aditza bertan dagoelako segurantza.
- Corpus orokorrarekin alderatuz: arestian esan bezala, adibideak corpus berezitu gisa har ditzakegu corpusetik atereak direlako, eta maiztasun txikiko aditzak aztertu nahi ditugunean corpus handietara jo beharrik gabe, adibideetan bertan aurki genezake aditz hauei buruzko oinarrizko informazioa. Eta, gainera, aukeratuak izateak corpus orokorrekoek ez duten autoritatea ematen die, nolabait. Esan beharra dago, ordea, corpusetan oinarritzeak maiztasuna kontuan hartzeko ematen duen aukera ezin izango dugula baliatu. Corpusek hiztegiko adibideek ez bezala, estatistika baliatzeko aukera ematen dute. Hala ere, ez da kontua bakarrik estatistikoki esanguratsuak diren esaldien azterketa egitea. Esaldi guztiak dira interesgarri aditz baten portaera aztertzeko, ez bakarrik maizen gertatzen direnak. Bestalde, jakina da aditzek portaera desberdina izan dezaketela domeinuaren arabera (Jensen 1991; Basilii eta beste 1993).

Hiztegiko adibideen aldeko arrazoi horiek kontuan izanik, azterketaren abiapuntua da hiztegiko adibideak lagungarri gertatuko direla aditzen oinarrizko azpikategorizaziorako. Hala ere, erabilera jatorraren gordailuaren aurrean egon arren, ez dakigu zenbateraino izango garen kapaz gure azaleko sintaxiaren bidez informazio hori guztia jasotzeko.

Azaleko sintaxia garatzerakoan eta adibideen gainean aplikatzerakoan, azterketaren arrakastaren alde izan ditugun ezaugarrien artean, honako hauek azpimarratuko genituzke:

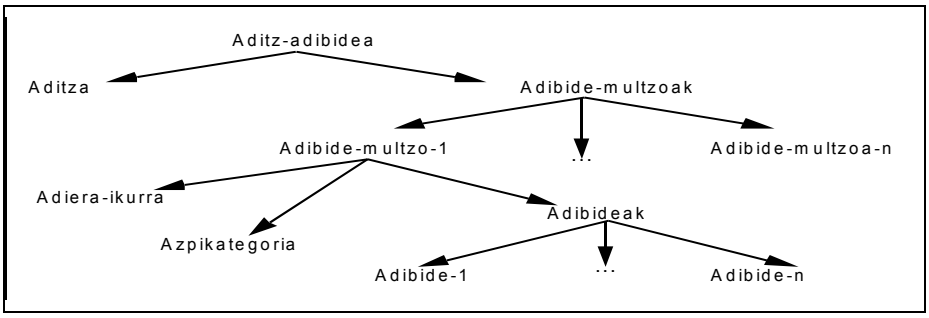
- %98,65 ez da anbiguo aditz-kateen eta sintagmen osaketarako darabilgun

irizpidearekiko. Edota, bestela esanda, %1,35ean gertatzen da nagusi/modifikatzaile anbiguotasuna.

- Esaldien luzera (6,44 hitz ditugu esaldiko), beraz, esaldi ezkonplexuak espero ditugu.
- Eta hiztegian azaltzen den laguntzaile-mota txertatu dugula aditzen analisi morfologikoan.

Gogora dezagun aplikazioaren helburu nagusia, aztertzen ari garen adibideetako aditz-sarreren inguruan dauden sintagmak eta aditz-kateak aditzen azterketarako baliagarri irizten diegun ezaugarriekin jasotzea dela. Beraz, zati horiei dagozkien etiketak gaizki esleituta azaltzen diren adibideak baztertuak izango dira. Hau dela eta, 13.089 adibidetatik (2.929 aditzi dagozkienak), 11.616ri (2.833 aditzi dagozkienak) aplikatuko zaizkie estandarizatzeko urratsa (hots, *Standar Generalized Markup Language* estandarraren arabera moldatuko dira adibideak) eta ondorengo galdeketa-sistemarena. Azken urrats hauetatik at geratzen diren adibideak (1.473) eskuz aztertu beharko dira.

Adibideetarik jasoko ditugun ezaugarriak gain-gainek bada ere ikusi ondorik, adibideak multzokatzeko irizpideak aurkeztuko ditugu. Lehenengo erabakia izan zen, adibideen analisiak, testu huts direnak, egitura aberatsago batera egokitzeko xedearekin zuhaitz-egitura batean gordetzea. Hau da, adibideak eta adibideen analisiaren emaitzetatik interesatzen zaigun informazioa zuhaitz modura errepresentatu ditugu. Eta, zuhaitz horren antolamendua definitzeko adibide-zuhaitzaren ezaugarri-egitura definitu dugu (ikus Arriola et al., 1999; Arriola, 2000). Bestalde, adibideak multzokatzeko irizpide nagusia adiera-ikurrarena izango dugu (adiera-multzoa, ñabardura, etab. Barne). Hain zuzen ere, ikur hauen arabera laguntzaile-mota (zuhaitzean "azpikategoria" gisa azaltzen dena) aldatu egin baitaiteke. Ikus dezagun grafikoki, aditzen adibideak errepresentatzeko zuhaitzean nola antolatzen ditugun adibideok:



VIII.1 irudia.- Adibideen antolamendua erakusten duen zuhaitza.

Adibideen antolamendu orokor horri eta SGMLratze urratsari esker, adibideen analisien corpusa, testu huts dena, egitura aberatsago batera egokitu dugu. Hartara, bertan jasotzen den informazioa modu errazago batez ustiatzeko.

Ustiapenaren emaitza gisa, aditz bakoitzeko azaleko patroien multzoa jaso dugu automatikoki, eta multzo horietako bakoitzean ditugun aditzen adibideak identifikatzeko gako bat definitu dugu. Ikus dezagun adibide baten bidez, gako horretan jasotzen den informazio-mota. Adibidez: *bultzatu-A0.-DU-2*:

- aditz-partizipioa: aztergai dugun aditzaren partizipioa. Adib. *bultzatu*.
- adiera-ikurra: zein adiera, azpiadiera edota ñabardurari dagozkion aditz horren adibideak. Adib. *A0*
- laguntzaile-mota: hiztegian duen laguntzaile-mota: DA, DU, DIO, ZAIO, DA-DU. Adib. *DU*
- adibide-zenbakia: aditz horren zenbatgarren adibidea den. Adib. *2*.

Adibidez hona hemen *bultzatu* aditzerako jaso ditugun azaleko patroiak:

```
**************************************************
bultzatu, bultza, bultzatzen.
**************************************************
bultzatu-A0.-DU-1          DU.@SUBJ_ERG-@OBJ_ABS.
bultzatu-A0.-DU-2          DU.@OBJ_ABS.MP+
bultzatu-A0.-DU-3          DU.@ADLG.
bultzatu-A0.-DU-4          DU.@SUBJ_ABS-@OBJ_ABS @PRED_ABS.MP-
bultzatu-A0.-DU-5          DU.@OBJ_ABS-@OBJ_ABS-@ADLG_ABZ-@OBJ_ABS-
                           @OBJ_ABS-@ADLG.MP+
bultzatu-N1.-DU-1          DU.@SUBJ_ERG.MP+
bultzatu-N1.-DU-2          DU.@OBJ_ABS.
bultzatu-N1.-DU-3          DU.@SUBJ_ERG-@ADLG_ALA.
bultzatu-N1.-DU-4          DU.@SUBJ_ERG-@OBJ_ABS.MP-MP+
bultzatu-N1.-DU-5          DU.@ADLG_ABZ-@OBJ_ABS.
bultzatu-N1.-DU-6          DU.@OBJ_ABS.
```

Aditz bakoitzerako ematen dugun azaleko patroi horretan lehenbizi hiztegian duen laguntzaile-mota azaltzen da, eta jarraian funtzio sintaktiko/kasu bikoteak[19] ; eta adibide horretan bestelako aditz-katerik azaltzen baldin bada, MP ikurraren bidez adierazten dugu (+ mendekoa / - ez-mendekoa izan den ala ez).

Aditzen sailkapenean eman ditugun azaleko patroi horien ebaluazioa egitea oso garrantzitsua da patroi horien fidagarritasuna neurtzeko. Horretarako, ausaz hartutako lagin bateko aditz bakarreko adibideen gaineko patroi bakoitzeko aztertu zen ea funtzio sintaktiko/kasuei erreparatuz, zenbat kasutan asmatzen zen eta zenbatetan huts egiten zen patroia osatzen duen funtzio sintaktiko/kasua. Baina, horretaz gain eskuz analizatutako laginean agertzen ez diren, eta analisi automatikoaren bidez markatzen diren funtzio sintaktikoen berri ere ematen dugu. Ebaluazioaren emaitzak, sailkatze automatikoa eta eskuzkoa erkatzearen ondorio dira. Eta hastapeneko ebaluatze horretan erdietsitako emaitza zehatzak (Arriola, 2000) lagin txiki bati dagozkionak direnez ezin dira orokortu, baina joera gisa ezaugarri hauek azpimarratuko genituzke:

- Oro har, ikus daiteke patroian subjektuak edo objektuak bakarrik hartzen duenean parte, asmatze-tasa txikiagoa izaten dela beste funtzio batekin konbinatzen direnean baino.
- Zehar objektuak parte hartzen duen patroiak fidagarriagoak dira.

Edozein modutan ere, etorkizunera begira ebaluazio-sistema sendoagoa lantzearen premia nabari geratu da.

9. Aditz-kateen eta sintagmen osaketaren ebaluazioa

Azaleko sintaxiaren atala aplikatzerakoan, aditzen sailkapenerako urrats oso garrantzitsua eman dugu. Hau da, sarrerako aditz-katearen inguruan dauden sintagmak eta bestelako aditz-kateak esplizituki adierazi ditugu zatiak markatzeko etiketen bidez. Urrats horren ondoren, ditugun adibide guztietatik (13.089), 400 adibideko lagina osatu dugu ausaz. Lagin honen gainean eskuzko azterketa burutu dugu bi ezaugarriri erraparatuz nagusiki:

1) Esleituriko aditz-kate nahiz sintagma-kate etiketak ongi esleituta dauden.

[19] Funtzio sintaktikoak eta kasuak azpimarra batez lotzen dira. Eta funtzio sintaktiko/kasu bikoteak bereizteko marratxoa baliatu dugu.

2) Aditz-kate nahiz sintagma etiketa behar duen elementuren bat etiketarik gabe dagoen. Kate-etiketa behar duten elementuak aurreko puntuetan ikusi ditugun sintagma eta aditz-kateak osatzeko parte hartzen duten elementuak dira. Beraz, bigarren puntu honetan etiketatzeke azaltzen diren zenbait elementu ez ditugu aintzat hartuko ebaluaziorako. Hau da, elementu horiek ezin dira ebaluatu, ez baitugu elementu horiek zati gisa etiketatzeko erregelarik garatu. Beste batzuen artean, honako elementu hauek geratuko lirateke garaturiko etiketatze-erregeletatik kanpo: lokailuak, juntagailuak, erlatibozkoak, hitz anitzeko unitate lexikalak, etab. Horrez gain, argi gera bedi ezagutzen ditugun kateak, aditz-kate ezjarraiak salbu, kate jarraiak izango direla.

Lehenengo puntuari dagokionez, 84 adibide baztertu behar ditugu horietan sintagma edota aditz-kateren bat gaizki osatuak baitaude. Beraz, %79 ongi etiketaturik daudela esan genezake. Gaizki etiketatze horren arrazoi nagusiak, honako hauek dira:

- Adibideetan geratzen den anbiguotasuna. Zatiak markatzeko estrategia funtzio sintaktikoetan oinarritzen denez, funtzio sintaktikoen anbiguotasuna izango dugu arazo-iturri. Baina, anbiguotasun guztiek ez dute eraginik zatiak markatzeko urratsean. Anbiguotasun kaltegarria izango dugu hitz batean funtzio sintaktiko nagusi bat eta funtzio sintaktiko eznagusi bat ditugunean. Lehenago aipatu dugunez, anbiguotasun hori oso txikia da, ehuneko bira ere ez baita iristen.
- Desanbiguatze-erroreak. Atal honetan, funtzio sintaktiko desegokiak hautatzerakoan suertatzen direnak hartuko ditugu kontuan. Hauek baitira zatiak osatzeko urratsari eragiten diotenak.
- Hitz ezezagunen arazoa. Hitz ezezagunak ditugu EDBLn sarrerarik ez duten hitzak. Hitz hauek ere analizatu egiten dira lexikorik gabeko lematizazioari esker. Kontua da, honelakoetan analisi zuzena asmatzea zailago suertatzen dela.
- Sintagma koordinatuak. Hauetarako baditugu zenbait erregela, baina gehienetan halako egituretan akatsak aurkitzen ditugu. Beraz, erregela horiek birfindu eta hobetu egin beharko dira.
- Postposizio-egiturak. Zenbait postposizio landuak baditugu ere, atal hau

gehiago osatu beharra dugu, horietako asko ez baitugu ezagutzen, eta oso garrantzitsua izan daitekeelakoan baikaude aditzen portaera aztertzeko.

- Deskribapen sintaktikorako etiketa-multzoan aurreikusteke dauden egiturak. Esate baterako, *-ik ena* (*Arbolarik ederrena* ...) moduko egiturak harrapatzeko deskribapen sintaktikorako baliatu dugun etiketa-multzoan aldaketak egin beharko lirateke.

Lotzeke geratzen diren elementuak batez ere postposizio-egiturak, sintagma koordinatuak eta aditz-kate ezjarraituak ditugu. Hauek atzemateko baditugu zenbait erregela, baina azaldu zaizkigun kasu hauek harrapatzeko hobetu egin beharko lirateke. Dena dela, gure xedea batez ere kate jarraituak ezagutzean datza.

10. Ondorioak

Azaleko sintaxiaren bidetik, euskararen murriztapen-gramatika garatu dugu, testu errealetako hitzen desanbiguazioa lortzeko. Gramatika horren ekarpen nagusiak anbiguotasun morfosintaktikoaren azterketa sistematikoa eta desanbiguazio morfosintaktikorako zein sintaktikorako erregelen zehaztapena dira.

Horrez gain, aditzen sailkapenak erdiesten laguntzeko, azaleko sintaxia aplikatu dugu EHko aditzen adibideen gainean, aditz-kateak eta hauen inguruan dauden sintagmak ezagutuz. Azaleko analisi horretatik analisi sakonago batera jauzia egiteko azpikategorizazioaren informazioa ezinbestekotzat jotzen dugu, besteak beste.
Nolabait, "gurpil zoro" batean gaudela dirudi. Hau da, batetik, azpikategorizazioaren informazioa erdiesteko sintaxiaren alorra sendotu beharra ikusten dugu, eta bestetik, hori hobetzeko ezinbestekotzat jotzen dugu azpikategorizazioaren informazioa.

Bestalde, aurreko atalean aurkezturiko arazo horiekin topo egin arren, eta, lortu dugun informazioa azalekoa izan arren, uste dugu informazio hori lagungarri izan daitekeela bai analisi sintaktikoan aurrera egiteko bai aditzen azterketan aurrera egiteko. Horretarako, lorturiko informazioa lexikoian integratu beharko litzateke gure tresna sintaktikoetan aplikatu ahal izateko.

Bibliografia

Abney S. (1997). *Corpus-Based Methods in Language and Speech Processing*, Steve Young and Gerrit Bloothooft (Eds.).

Aduriz I. (2000). *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N. & Urizar R. (1996a). "EUSLEM: A lemmatiser/tagger for Basque". *Proceedings of EURALEX'96.* Göteborg, Sweden, Part 1, 17-26.

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A. & Insausti J. M. (1998). "EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque". *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98),* Granada.

Agirre E., Arregi X., Arriola J. M., Artola X., Insausti J. (1994). *Euskararen Datu-Base Lexikala (EDBL)* . Barne-txostena UPV/ EHU / LSI / TR8-94.

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

Alegria I., Artola X. & Sarasola K. (1997). "Hizkuntzaren tratamendu automatikoa". *JAKIN* 102, 61-82.

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

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". *Proceedings of the LINC.* 1-7. Bergen, Norvegia.

Basilii R., Pazienza M.T. and Velardi P.(1993). Acquisition of Selectional Patterns in Sublanguages. Machine Translation, vol. 8, 175-201.

Black E., Garside R. & Leech G. (1993). *Statistically-Driven Computer Grammars of English: The IBM / Lancaster Approach*. Black, Garside & Leech (eds.), Rodopi, Amsterdam.

Brill E., Wu J. (1998). Clasifier Combination for Improved Lexical Disambiguation. COLING-ACL'98, Montreal.

Euskaltzaindia. (1993). *Euskal Gramatika Laburra: Perpaus Bakuna*. Euskaltzaindia, Bilbo.

Ezeiza N. (1997). *EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa*. Tesina-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

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

Hays D. C. (1964). “Dependency theory: a formalism and some observations”. *Language* 40, 511-525.

Jensen K. (1991). *A Broad-Coverage Natural Language Analysis System*, in M. Tomita (ed.), Current Issues in Parsing Technology, Kluwer.

Kaplan R.M. eta Newman P.S. (1997). Lexical Resource Reconciliation in the Xerox Linguistic Enviroment. *Proc. of a Workshop on Computational Enviroments for Grammar Development and Linguistic Engineering*, Madrid.

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

Koskenniemi K. (1983). *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. Phd. thesis, University of Helsinki.

Laka I. (1998). *A Brief Grammar of Euskara, the Basque Language*. HTML-ko dokumentua. Euskararako Errektoreordetza, Euskal Herriko Unibertsitatea.

Màrquez L. (1999). *Part-of-Speech Tagging: A Machine Learning Approach based on Decision Trees*. Doktoretza-tesia, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

Mel'cuk I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Mitxelena L. (1972). “De la ambigüedad sintáctica”. *Revista Española de Lingüística*. 1972-2, Madrid, 237-247.

Sánchez F. (1997). *Análisis morfosintáctico y desanbiguación en castellano*. Doktoretza-tesia, Departamento de Lingüística, Lenguas Modernas, Lógica y Filosofía de la Ciencia. Universidad Autónoma de Madrid.

Tesnière, L. (1959/1966) Elèments de Syntaxe Structurale, 2. arg. Errebisatua, Paris, Klincksieck.

Urkia M. & Sagarna A. (1991). “Terminología y Lexicografía Asistida por Ordenador. La experiencia de UZEI”. *Actas del VII congreso de la SPLN*. Donostia. (Urkia, 1998).

Urkia M. (1997). *Euskal Morfologiaren Tratamendu Automatikorantz*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Voutilainen A. (1994). *Designing a Parsing Grammar*. Publications of the Department of General Linguistics, 22. University of Helsinki.

Zabala I. & Odriozola J.C. (1994). ““Adjektiboen” eta “adberbioen” arteko muga zehatzik eza”. *ASJU* (XXVIII-2). Donostia. 525-541.