Jakintza-arloa: Hizkuntzalaritza

# Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira: aditz-adibideen analisia murriztapen-gramatika baliatuz, azpikategorizazioaren bidean

Egilea: **JOSE MARI ARRIOLA EGURROLA**

Urtea: 2000

Zuzendaria: JOSEBA ABAITUA ODRIOZOLA

Unibertsitatea: UPV/EHU

ISBN: 978-84-8438-142-6

## Hitzaurrea

Hitzaurre honetan 2000. urtean aurkeztu nuen tesi-lanaren nondik norakoak kokatzen saiatuko naiz. Euskal Herriko unibertsitatean “Bikain Cum Laude” kalifikazioa lortu zuen Joseba Abaitua eta Xabier Artola doktoreen zuzendaritzapean burututako *Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira*. *Aditz-adibideen analisia Murriztapen Gramatika baliatuz, azpikategorizazioaren bidean* tesiak. Epaimahaian Fred Karlsson, Patxi Goenaga, Beñat Oiharzabal, Irene Castellón eta Xabier Arregi jaun-andreak egon ziren.

Euskal Herriko hizkuntzalaritzaren tradizioan hizkuntzaren azterketa konputazionala ez zen erabat berria: sintaxia Abaitua (1988), morfologia Urkia (1997), esaterako. Ikertzaile batek baino gehiagok jorratua zuen bide hori. Aipatzekoa ere bada nire tesi-egitasmoaren motibazioak eta helburuak eskuartean nituen lanetatik etorri zirela. Eta lan horiek talde-lana zutela oinarri, IXA taldearena. Hala ere, garbi dago arlo hori ez zegoela erroturik gure artean. Beraz, gure lana hasten ari zen diziplina baten emaitza dugu. Izan ere, tesiaren oinarrizko motibazioa ondorengo proiektu eta egitasmoentzat beharrezko izan zitezkeen baliabideak eta metodoak ezartzea zen: baliabide lexikalen berrerabilpena eta aditzen argumentu-egitura zehazten laguntzeko bideak eskaintzea, alegia.

Tesi-proiektuaren hipotesia izan zen *Euskal Hiztegia*-ko (EH) adibideak baliagarriak izango zirela aditzen azpikategorizazioa lantzen laguntzeko. Hortaz, erronka hauxe genuen: zenbateraino izango ginen gai gure metodo eta tresnen bidez hiztegiko informazioa erdiautomatikoki erauzteko.

EH hiztegian aztertzeko genuen xedea ez zen artean euskal hizkuntzalaritzak izan dituen ohiko egitekoen modukoa eta arrazoia begien bistakoa da. Hala ere, 80. hamarkadan hizkuntzalaritza konputazionala aurreratuago zegoen tokietan hedatuta zegoen hiztegiak helburu konputazionalarekin erabiltzea, bereziki, euskarri magnetikoan dauden hiztegiak (Machine Readable Dictionnaries, MRD). Hiztegi horiek baliabide lexikal oparotzat jotzen dira, lexikografoen hainbat urtetako lanen ondorioz informazio lexikal eta semantiko aberatsa baitute.

Hiztegi horiek aztertzerakoan bi alderdi bereizi behar ditugu: i) prestatze-lana, diziplina beregaina zena hein handi batean eta ii) analisirako metodologia.

Adibideak analizatzeko erabili genuen metodologia azalekoa izan arren, atalik erakargarriena dela esango nuke. Batetik, hiztegiko adibide guztiak analizatzeko gai izan ginelako, eta, bestetik, analisi horiek erdiesteko erabili zen analisi-hurbilpenak geroko lanetan ere izan duelako jarraipenik. Seguruenik, gaur berriz hasi beharko banu, garai hartako hasiera-hasierako asmoak murriztu egingo nituzke, eta azaleko analisi sintaktikoaren alde horri garrantzi gehiago eman ez ezik, gai honi helduko nioke bakarrik. Nik uste dut gaur egun ere badela tesi baterako baino, hiru edo lau tesirako adinako gaia ere.

Bukatze aldera, esan beharra dago honelakoetan lortutako emaitzei eman ohi zaiela garrantzirik handiena. Gurean, prestatze-lanaren atalari dagokionean, esan daiteke emaitza onak izan zirela, eta, ez genuela denbora alferrik galdu; nahiz eta, askotan, oso esker txarrekoa zen lan horrek besterik ematen bazuen ere. Aldiz, adibideen analisien emaitzak ez ziren hain onak izan. Bazen zer hobetu eta jarraitu behar ez den biderik ere. Hurbilpen hartako ondorioetan saiatu nintzen horiek guztiak jasotzen, hartara, ondorengo lanetan aprobetxatzekoak izan zitezen.

Osotasunean, lan hau hizkuntzalaritza konputazionalekotzat har daiteke, baina emaitzak eta ondorioak aztertzerakoan, euskararen ezaugarri soziolinguistikoak ere aintzat hartu behar dira. Baliabide lexikalak genituen helburu, eta, era berean, horiek eskuratzeko alde metodologikoak lantzea. Garbi dago, tesiko garaitik hona, asko aurreratu dela baliabideetan: corpusak, hiztegiak, gramatikak, besteak beste. Hala ere, euskarak duen baino egoera arruntagoa eta osasuntsuagoa duten hizkuntzen artean ere ingelesa da nagusi hizkuntzalaritza konputazionalaren arloan. Beraz, euskarak atzean gelditu nahi ez badu, badugu zer egin!

eman ta zabal zazu

euskal herriko unibertsitatea
universidad del país vasco

EUSKAL FILOLOGIA SAILA

# *EUSKAL HIZTEGIA*-REN AZTERKETA ETA EGITURATZEA EZAGUTZA LEXIKALAREN ESKURATZE AUTOMATIKOARI BEGIRA

## Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean.

**Jose Mari Arriola Egurrolak**

Euskal Filologian Doktore titulua eskuratzeko aurkezturiko

**TESI-TXOSTENA**

Gasteiz, 2000ko apirila.

eman ta zabal zazu

euskal herriko unibertsitatea
universidad del país vasco

EUSKAL FILOLOGIA SAILA

# *EUSKAL HIZTEGIA*-REN AZTERKETA ETA EGITURATZEA EZAGUTZA LEXIKALAREN ESKURATZE AUTOMATIKOARI BEGIRA

## Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean.

Jose Mari Arriola Egurrolak Joseba Abaitua eta Xabier Artolaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Euskal Filologian Doktore titulua eskuratzeko aurkeztua.

Gasteiz, 2000ko apirila.

# eskerrak ematen

Eskerrik asko IXAko guztiei, azken finean lan hau guztiona baita. Bereziki, Igeldoko Perlari, Nafar Irribarretsuari eta Donostiar Finari.

Antxoko Alternantzia Bikainari, Vikingo Gorriari eta Begoñako Torpedoari.

Luzaroan pairatu nauen sorgin katalanari (força MG!)

Pisukide ohiei: ollun popi akat!!!

BEKADUN guziei: eutsi goiari!

# eskerrak emoten

Hasi eta Amatxuko kuadrilliai: segi holantxeik parrandai eutsitxen!

Familixakuai ez takat berbaik birdan moure eskertzeko, beti ixango naz sorretan eta!

Amaitzeko, AUPA MARI!

Nerea inprimategia.

Ondarroako Unibertsitatea, 2000ko apirila.

Egilea baliatu da, lan honetan, Gipuzkoako Kutxaren diru-laguntzaz eta Eusko Jaurlaritzako bekaz.

# AURKIBIDEA

# SARRERA ETA AURKEZPEN OROKORRA

# I. *Euskal Hiztegia*-ren[1] azterketa erdiautomatikoa: motibazioa eta helburuak.

## I.1 Motibazioa.

Tesi-proiektu honek bi motibazio nagusi izan ditu: (1) existitzen diren baliabide lexikalen berrerabilpena, hauetan gorderik dagoen informazioaz aprobetxatuz Euskararen Datu-Base Lexikalaren (EDBL) aberasketari eragiteko eta (2) aditzen argumentu-egitura zehazten laguntzeko bideak eskaintzea, eta ahal den heinean azaleko argumentu-egitura erdiestea analizatzaile sintaktikoetan integratzeko.

Beraz, gure lanaren zeregin nagusia informazio lexikalaren eskuratzearena dugu. Lengoaia Naturalaren Prozesamenduaren (LNP) komunitatean 1980ko hamarkadan baliabide lexikal zabal eta aberatsen beharra zabaldu zen. Alderdi teorikotik, teoria linguistiko erabilienak (segur aski Chomskyrengandik hasita, 1970ean) lexikora lerratzen dira. Aplikazioen ikuspegitik, LNPrako aplikazio errealak garatzeko, ezinbestekoa zen lexiko zabalak edukitzea. Lexikoi konputazionalak dira LNPko aplikazioetako osagai garrantzitsuenetarikoak, eta beren baitan jaso beharreko informazioen artean azpikategorizazioarena nabarmentzen da; bereziki aditzei dagokienean.

Bestalde, ordura arte erregela konplexu eta ugarien bidez deskribatzen ziren fenomeno linguistiko askok jatorri lexikala zutela jabetzean, lexikoa hitz-zerrenda laua izatetik informazio

---

[1] Sarasola I.. *Euskal Hiztegia.* Kutxa Fundazioa: Donostia, 1997.

aberats eta konplexua zuen sistema izatera pasa zen. Gauzak horrela, lantaldeak lexiko horiek eskuz eraikitzen hasi ziren. Kodetu beharreko informazio kopurua itzela da eta pertsona-urte askotako ahalegina eskatzen du, proiektu erraldoi gutxi batzuen esku dagoena (adibidez, CYC, EDR edo WordNet proiektuak[1]). Eskuzko kodeketaren osagarri gisa, lexikoiak[2] edukiz betetzeko laguntza automatiko edo erdiautomatikoak ere bilatu izan dira, eta horrekin arreta bestelako baliabide lexikalen tratamendura zuzendu zen, hots, corpus eta hiztegietara.

Aditzen azpikategorizazioari gagozkiola, informazio hau LNPko aplikazioetarako lexikoietan landua izatea oso garrantzitsua da. Esaterako, berriki garatu diren teoria sintaktiko gehienak aditzen argumentu-egiturari buruz lexikoian gorderiko informazioan oinarritzen dira egitura sintaktikoak eraikitzeko. Beraz, analizatzaile sintaktiko sendoak izateko aditzen argumentu-egiturari buruzko informazio zehatza beharko da. Lexikoietako aditzak informazio honekin hornitzerakoan sortzen dira arazoak, hau da, aditz edota aditz-mota bakoitzeko dauden azpikategorizazio aukerak eskuz lantzea eta mantenimendua egitea oso garestia da. Arazo hau arintzeko asmoarekin azpikategorizazio automatikoaren laguntza hartuko da irtenbidetzat linguistika konputazionalaren (LK) arloan —ikus § IV.1—. Hona hemen, (Kuhn eta beste, 1998:89)-n azpikategorizazioa eskuratzeko beharrari buruz esaten dena:

> "However, for verbs (and for relational adjectives and nouns), an idiosyncratic assignment of subcategorization frames for each lemma is indispensable if the grammar is supposed to produce accurate and deep analyses, as they are required, e.g., in high-quality machine translation. Now, unless the domain to be covered by the system is restricted radically, it is impractical to hand-encode the required subcategorization lexicon; thus the success of broad-coverage symbolic parsing depends to a considerable degree on ways of automating the construction of high-quality lexical resources."

LNPko ikertzaileek gero eta behar handiagoa izan dute lexikoi konputazionalak corpusetatik abiatuz eraikitzeko, jada dauden datu-base lexikal "estatikoetatik" abiatu beharrean (Pustejovsky eta Boguraev, 1994). Hainbat saio izan dira lexikoietan azpikategorizazioari dagokion informazioa corpusetatik eskuratzeko, besteak beste (Brent, 1991a; Manning, 1993; Briscoe eta Carroll, 1997) aipa litezke.

Gure proiektuaren xedeetariko bat, lehenago aipatu dugun bezala, lexikografoari edota hizkuntzalariari aditzen azpikategorizazioa lantzeko laguntza eskaintzea dugu. Kontuan izan behar da, argumentu-egitura zehazteak denbora askoko lana eskatzen duela, eta ez nolanahikoa gainera. Hau da, oso zaila da aditz batek izan ditzakeen erabilera guztiez jabetzea, aditza

---

1 CYC, 13 urtetik gora iraun duen proiektu erraldoi honetan, pertsonok dugun sen ona ezagutza-base batean islatu nahi izan dute. EDR, *Japan Electronic Dictionary Research Institute* sortu zuten Japonian 1986an, japoniera eta ingelesaren tratamendu automatikorako lexikoa eraiki zezaten. Wordnet, ingeleserako sare semantikoen artean ezagunena-edo dugu. Oso erabilia da LNP inguruko ikerkuntzan eta edozeinek eskura dezake Internet bidez.

2 Lexikoi terminoa erabiltzen da Lengoaia Naturalaren Prozesamenduaren arloan informazio lexikalaren biltegiei edota hiztegiei erreferentzia egiteko.

erabiltzen deneko testuinguruak izan gabe. Nabaria da arlo honetan, lexikografo zein linguistentzat oso garrantzitsuak direla azpikategorizazioari buruzko informazioa corpusetatik eskuratzen laguntzen duten tresnak garatzea. Hona, Sussane Gahl-ek (1998:428) dioena:

> "There is a need for a tool that can (1) find evidence for subcategorization patterns and (2) determine their frequencies in large corpora ..."

Gure azterketak ere bide hori jorratzeari ekin dio. Horrela bada, gure zeregin nagusiari erantzuteko, abiapuntu gisa hiztegi arrunt baten MRD (*Machine Readable Dictionary*) bertsioa hartu dugu, Euskal Hiztegiarena (EH), gure ustez jada existitzen diren baliabide lexikalen berrerabilpena garrantzitsua baita. Hala ere, jakin badakigu irtenbide honek ere badituela bere gorabeherak. Alde batetik, MRDetatik informazioa eskuratzean dauden arazoak —ikus § III.2—, eta bestetik adibideen beraien egokitasuna azpikategoria kontuetan aurrera egiteko. Arazo horiei gehitu behar zaizkie behin analisia erdietsi ondoren, informazioa eskuratzeko orduan sortzen diren arazoak. Arazo hauek, VI. kapituluan azalduko dira azaleko patroien ebaluaketarekin batera —ikus § VI.4—.

MRDez den bezainbatean, oso erabiliak izan dira jakintza eskurapenerako prozesuetan, definizio-eremuak oparo aztertu dituzte erlazio lexiko-semantikoak erauzteko asmoarekin. Badira probetxugarri gerta daitezkeen beste eremu batzuk ere, eta hauetarik adibideena nabarmentzen da. Azpimarratzekoa da hiztegi-ezagutzen iturririk aberatsenetako bat adibideetan datzala. Kontuan hartzeko moduko ezagutza garrantzitsua, hizkuntz ezagutza batik bat. Bertan aurki daitezkeen informazioen artean, kasu-erlazioen argumentuena litzateke ohargarrienetakoa (Wilks eta beste,1989).

EH[1]ren MRDa abiapuntu egokia dela uste dugu, bai EDBLren aberasketa orokorrerako bai aditzen azpikategorizazioaren gaiari heltzeko. Badakigu lan honek ezin diola helburu orokorreko argumentu-egiturak lortzeko erronkari erantzun, baina uste dugu ezerezetik hasi beharrean lan honen bidez eskaintzen den informazioak lagun dezakeela helburu orokorreko proiektuetan. Hau da, hemendik eskuratutako informazioa corpusetan egin beharreko azterketetarako lagungarri gertatuko delakoan gaude. Beraz, bi iturriok konbinatzearen aldekoak gara.

Dena den, adibideen "egokitasunaz" hitz egiterakoan autore guztiak ez datoz bat, eta maizen egozten zaien akatsa da lexikografoek jasoriko adibideetan testuetan agertzen diren erabilera asko falta direla, edota beti adibide berberak jartzen dituztela. Hori dela eta, corpusetara jotzeari egokiago deritzote. Halaxe diote, besteak beste, Briscoe eta Carroll-ek (Briscoe eta Carroll, 1997:356):

---

[1] Hemendik aurrera, *Euskal Hiztegia*-ren ordez EH laburdura erabiliko dugu.

> "These problems suggest that automatic construction or updating of subcategorization dictionaries from textual corpora is a more promising avenue to pursue."

Hala eta guztiz ere, gure hipotesia da hiztegietako adibideak egokiak direla ditugun helburuei erantzuteko. Hau da, gure azterketak ikerketa sakonagoetarako laguntza eskaini nahi du, eta horrez gain, adibideen autoritatea kontuan izanik, uste dugu hauetan gordetzen den informazioa baliagarria izan daitekeela oinarrizko informazioa erdiesteko edota corpusetan aurrerantzean azterketa sakonagoak gidatzeko.

Adibideak ez dira nolanahiko testuak ezpabere aukeraturikoak ditugu; "tipikoak", hots, tradizio lexikografikoan oinarrituta oso informazio aberatsa eskaintzen digute hitz bati dagokion erabilera arrunt eta funtsezkoena deskribatuz. Hona hemen, Ibon Sarasolak (Sarsola, 1997:XXIII) EHren sarreran dakarrena adibideen eginkizunei buruz idazterakoan:

> "Alegia, adibideen bitartez definizioak argitzen eta osatzen, hizkuntz ereduak finkatzen eta maizeneko erabilerak jasotzen saiatu naiz."

Bestalde, gauza bat da aditz batek erakusten dituen azaleko egitura sintaktikoak jasotzea, eta oso bestelakoa da aditz baten argumentu-egitura zehaztea. Hots, guk landutako metodologiaren bidez, aditzen adibideetan aditz bakoitzak inguruan dituen sintagmak eta aditz-kateak bereiziko ditugu. Baina, horiek aditzaren argumentu diren ala ez linguistak edota lexikografoak erabaki beharko du.

Zergatik hartu dugu hiztegi hau gure proiektuaren oinarri? Gure kasuan, euskarazko hiztegia behar genuen noski, hots, euskarazko adibideak dituen hiztegia. Hiztegiaren egileak sarreran dioenari erreparatuz gero, ohar gaitezke adibideetan datzala, hain zuzen ere, hiztegi honen altxorrik preziatuena, tradizio literarioan oinarrituta baitaude. Eta gainera, oso garrantzitsua da kontuan izatea corpus zabalago batetik aukeratuak izan direla. Hain zuzen ere, hiztegiei maiz egozten zaien errua da gehiago oinarritzen direla lexikografoen intuizioetan testu errealetan baino. Hona hemen, horri buruz Ulrich Heid-ek (1994:459) zer dioen:

> "Many dictionaries, for both natural language processing (NLP) and human use, are based more on lexicographers´ introspection than on "real text" as it occurs in newspapers, books, spoken discourse, etc. Only recent work in British lexicography (cf. work by Sinclair 1991, Atkins/Filmore 1991), and a few dictionary projects for other languages (e.g. Den Danske Ordbog) are based on text corpora and accompanied by methodological work on corpus use in lexicography."

Hiztegiek hizkuntzaren eredu bat definitzen dute. Lexikografoak saiatzen dira eredu hori hizkuntzaren deskripzio zabalenaren eta neutroenaren bidez egiten. EHko adibideek hizkuntzaren corpus-eredu "bitxi" bat osatzen dute. "Bitxia" diogunean, adierazi nahi dugu adibide horiek alde batetik corpusetik atereak direla. Eta beste aldetik, adibideak lexikografoak hautatuak direnez, nolabait bere intuizioak ere biltzen dituela adibideen corpus honek. Hau da,

nolabait intuizioak eta corpusa elkartzen dira adibideetan. Hala ere, garbi dago gaur egungo joera nagusia, bai lexikografian bai hizkuntzalaritzan corpusaren aldekoa dela.

Bestalde, EH gure lanerako iturri gisa hautatzearen arrazoiekin jarraituz, ordenagailuz irakurgarria (MRD) izatea ere ezinbestekoa genuen. EH euskarri magnetikoan izateak ateak zabaldu dizkigu lexikografia konputazionalaren eremuan oparo jorratu den bide honi ekiteko. Gure lanean, LNPko sistemen osagai lexikalak eraikitzeko iturburua dugu EHren euskarri magnetikoan dagoen bertsioa. Hau da, hiztegia oso izaera bereziko testu-corpustzat hartzen dugu, jakin badakigun arren testu-corpus hutsa baino askoz ere gehiago dela, hitzen esanahiari dagokionean. Esan dezakegu "ezusteko" iturri lexikal gisa darabilgula, berez giza erabiltzaileari zuzenduriko hiztegia baliabide lexikal gisa baliatu nahi baitugu, iturri jori honetan gordetzen den informazioaz aprobetxatuz.

Aldeko eta kontrako arrazoiak kontuan izanik, EHko aditzen adibideak aztertzea izan dugu xede. Azter zitezkeen adibideen artean, aditzena aukeratzearen arrazoi nagusiak honako hauek dira:

- Eskura daitekeen informazioari erreparatuz: argumentu-egituraren berri jaso nahi dugu, hots, ahal den heinean aditz bakoitzak zein argumentu eskatzen duen zehaztuko da. Informazio hau oso inportantea izango baita hurrengo aplikazioetarako, bereziki analisi sintaktikorako. Adibidez, aditz batek ea mendekorik hartzen duen ala ez jakitea oso inportantea da parsing-a garatzerakoan, informazio hori zehaztuta izanez gero habiaturiko esaldien kopurua jaitsi egingo bailitzateke. Mota honetako informazioa ez da azaltzen hiztegi elektroniko gehienetan, baina oso garrantzitsua da LNPn garatzen diren analizatzaile sintaktiko (parser) desberdinetarako.

- 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, honezaz 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. Hona hemen, Ibon Sarasolak (Sarasola, 1997:XXIII) zer dioen:

  > "Adibideok, gehienbat, idazle hoberenetarik hartu dira."

  Horrez gain, esaera zaharren ondare bikaina ere adibideen artean dugula ezin ahaztu daitekeena da.

Ordea, esan beharra dago corpusetan oinarritzeak maiztasuna kontuan hartzeko ematen duen aukera ezin izango dugula baliatu. Corpusek hiztegietako 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; Basili eta beste 1994).

Hiztegiko adibideen aldeko arrazoi horiek kontuan izanik, lehenago aipatu dugun bezala, azterketaren abiapuntua da hiztegiko adibideak lagungarri gertatuko direla aditzen oinarrizko azpikategorizaziorako; hala ere, hau frogatzeke dagoen zerbait da. Zer esan nahi dugu "frogatzeke" dagoela diogunean? Esan nahi duguna da, erabilera jatorraren gordailuaren aurrean egon arren, ez dakigula zenbateraino izango garen kapaz gure metodologiaren bidez informazio hori guztia jasotzeko, eta horrekin batera, lexikografoaren iritziz jasoriko adibideek agian ez dietela erantzuten LNPko ikuspegitik adibide hauetan aurkitu nahi genituzkeen informazio-motei.

## I.2 Helburuak.

Aurreko puntuan aipatu dugun bezala, EHren MRDa hartu dugu azterketaren oinarri gisa. Autore anitzek seinalatzen du LNPko sistemen osagai lexikalak eratzeko hiztegi arruntez baliatzeak duen interesa (Alshawi, 1989; Byrd, 1990; Wilks eta beste, 1990, Richardson, 1997). Horrela bada, euskarri magnetikoan dauden hiztegiez (MRD) baliatzea izan dute helburu joan deneko azken hamarkadan lan egiten ziharduten ikertalde askok. Tesi-proiektu hau ere arlo honetan koka daiteke, abiada helburu nagusi hauekin eman zaiola:

- EH hiztegiaren analisia bera: EH ordenagailu bidez aztertu ahal izateko egin behar den oinarrizko urratsa dugu hau. Kontuan harturik EH giza erabiltzaileari zuzenduriko hiztegia dela, lehenbiziko lana prozesamendu automatikoari heldu aurretik hiztegi-testuaren egituratzea izan da. Jakina, lan hau burutzeko ezinbestekoa da MRD bertsioa izatea. Lan honen fruitua, analisiaren emaitza eredu estandar baten proposamenetara egokitu dugu: *Text Encoding Initiative-ko* (TEI[1]) formatu estandarrera. Horrela bada, euskara aztergai edo lantresna duen edonorentzat baliagarri izango da. Analisi honi esker, interesgarri deritzogun informazio-eremuak aztertu ahal izango ditugu ordenagailu bidez, gure kasuan adibideak.

1 http://www.uic.edu/orgs/tei/index.html

- Euskararen tratamendu automatikorako Datu-Base Lexikala (EDBL) aberastea: EDBL dugu IXA taldeko[1] lanen emaitzetariko bat. EDBL, lexikoiari buruzko informazio-biltegi erraldoia eta euskararen prozesamendu automatikoan zeregin desberdin askotarako oinarri lexikala da. EHko azterketatik ateratzen diren informazio-mota anitzez hornitu nahi dugu EDBL sendotu eta aberastu dadin.

- EHko aditzen adibideen azterketa: azpimarratzekoa da hiztegi-ezagutzen iturririk aberatsenetariko bat adibideetan datzala. Kontuan hartzeko moduko ezagutza garrantzitsua, hizkuntz ezagutza batik bat. Bertan aurki daitezkeen informazioen artean, kasu-erlazioen argumentuena litzateke ohargarrienetarikoa (Wilks eta beste, 1989). Tesi honetako helburu zehatza da adibideetarik oinarrizko argumentu-egitura ahal den heinean eskuratzea eta informazio hau EDBLren osaketarako baliatzea.

- Aditzen azpikategorizazio-lanetarako laguntza eskaini: oinarrizko argumentu-egitura eskaintzeaz gain, alderdi metodologikotik hau erdiesteko erabilitako bidea baliagarria izan daitekeelakoan corpus handiagoetan aplikatzeko, edota erdietsitako informazioa balia daitekeelako corpusean egin nahi diren azterketetarako abiapuntu gisa.

- Azaleko analisi sintaktikoa landu: adibideak aztertzeko adibideen analisi sintaktikoa egin behar izan dugu. Azaleko analisia burutu dugu, *Constraint Grammar*[2] formalismoa jarraituz, euskararako garatu dugun Murriztapen-gramatika (EUSMG) baliatuz —ikus § IV.8—. Eta analisi hori beti hartu ahal izango da analisi sakonago bat egiteko abiapuntu eta laguntzatzat.

- Hiztegigintzarako laguntza: hiztegigintzan diharduen ororen jakinmina piztea ere balitzateke gure asmoetako bat, sinetsita baikaude, edozein dela ere hiztegi mota, etorkizunari begira hiztegiok era kontsistente eta ahalik eta zabalduenean (hots, formatu estandarrean) edukitzeak daukan garrantzi handiaz.

---

1 IXA taldea Donostiako Informatika Fakultateko ikertaldea da. Euskararen tratamendu automatikorako oinarriak finkatzea du helburu. Helburu horri erantzuteko LNP arloko ikerkuntza garatzea ezinbestekoa izan du. Jorraturiko ikergaiak eta sorturiko tresnen berri izan nahi duenak jo beza amarauneko helbide honetara: http://www.ixa.ehu.es.

2 Murriztapenetan (constraint-based) oinarritutako formalismo gramatikal gisa ezagutuak dira, bateratze-gramatiken multzoan ere maiz kokatzen diren mota askotariko formalismo gramatikalak, hala nola, Functional Grammar (FUG), Head-Driven Phrase-Structure Grammar (HPSG), Lexical Functional Grammar (LFG), Categorial Unification Grammar (CUG), Tree Adjunction Grammar (TAG), etab. Gure kasuan, euskararako garatu dugun gramatikak Constraint Grammar formalismoa jarraitzen du, baina ez du zerikusirik aurrerago aipatu dugun baterakuntzan oinarritzen diren gramatika horiekin.

# I.3 Proiektuaren aurkezpen orokorra.

Motibazioak eta helburu nagusiak azalduta, proiektuaren aurkezpen orokorra egingo dugu atal honetan. Ideia nagusia da, ditugun baliabideak erabiliz (nagusiki EHren MRDa eta Euskararako Murriztapen-gramatika (EUSMG)), informazio lexikala eskuratzea bai EDBL aberasteko bai analisi sintaktiko osoagoa egiteko. Hau da, azpikategorizazioari buruzko informazioa eskuratu nahi dugu. Eta, kontuan izan behar da ez dugula hasieratik azpikategorizazioari buruzko informaziorik. Baina, lortzen den informazioa erabili nahi dugu gure analizatzaile sintaktikoetan, horrela tresna horrek ahalmen handiagoa izango luke analisi sakonagoak egiteko eta berriz ere azpikategorizazioaren eskuratzeari ekiteko (prozesu hori *bootstrapping* lexikal gisa ezagutzen da). Horrela bada, zenbait autorek informazio lexikalaren eskurapenerako jarraitu duten bideari ekin nahi diogu. Besteak beste, Brent (1997:25)-n ditugu, aipatu *bootstrapping*-a lantzen duten zenbait autore:

> "Both Siskind and Brent and Cartwright developed approaches to bootstrapping problems that do not involve importing analyses of input utterances from another linguistic domain. These new approaches involve extracting a tiny bit of linguistic knowledge from unanalyzed inputs, thereby extracting more knowledge from them."

Aurrerago esan dugun bezala, EHren MRDa hautatu dugu "ezusteko" baliabide lexikal gisa informazio lexikalaren eskurapenari ekiteko. Hiztegi hori euskarri magnetikoan egotea oso garrantzitsua da, baina ordenagailuen erabilerak ez du soilik euskarri aldaketa ekarri, baizik eta LNPko teknikak erabiltzeko aukera ematen digu hiztegi horretan gorderik dagoen informazioaz probetxatzeko. Horretarako, lehendabizi, aipaturiko helburuei erantzuteko EH ordenagailu bidez aztertu ahal izateko bete behar diren oinarrizko urratsak egin ditugu. Hauen ondorik, adibideak aztertzeko moduan prest izan ditugunean, LNPko teknikak garatuz eta gure baliabideetan oinarrituz, aditzen adibideetan aztergai ditugun aditzek zer-nolako sintagmak eta aditz-kateak azaltzen dituzten jaso dugu. Gure lanaren bidez, aditz bakoitza zer-nolako sintagmekin eta aditz-kateekin azaltzen den jartzen dugu eskura, baina lehenago aipatu dugun bezala, azken finean horiek argumentu diren ala ez linguistak erabaki beharko du. Hau da, erabat automatikoki ezin dugu erabaki argumentu diren ala ez. Beraz, hein batean bederen, garatu nahi duguna adibideetarik ezagutza linguistikoa eskuratzeko metodo bat da. Metodo honek, aditzen erabilera erakusten duten adibideetatik argumentu-egituari dagokion informazioa erdiesteko bideak ematen ditu. Gure hurbilpenak argumentu-egituraren eskuratze erdiautomatikoari jarraitzen dio. Hona zer dioten (Kuhn eta beste 1998:89)-n:

> "The ideal acquisition method is a semi-automatic one with very high precision in the automatic phases combined with an efficient scheme of post-editing by a human lexicographer (Eckle-Kohler, 1998)."

Hala ere, esan beharra dago, badirela bestelako hurbilpenak ere. Esaterako, (Meyers eta beste, 1994)-n bost irizpide eta bost heuristiko proposatzen dira argumentuak bereizteko, eta, sei irizpide eta bi heuristiko adjuntuak bereizteko. Lan horretan argumentu eta adjuntuen arteko bereizketa estandarizatzeko asmoa zuten, eta aipatu irizpideok eskuz aplikatu zituzten. Horietariko asko ezin dira automatikoki landu, semantikan oinarritzen baitira. Adibidez, hautazko argumentuak ez baldin badira azaltzen, argumentu horiek nezkez atzeman daitezke automatikoki, horrelakoetan interpretatu edo ulertu egin beharko dira.

Ohargarria da, bestalde, aditz bakoitzaren inguruan dauden sintagmak eta aditz-kateak ezagutzeko atala garatzean, azaleko sintaxi gisa ezagutzen den arloa landu dugula. Horretarako, EUSMGren analisiaren irteera abiapuntutzat hartu, eta analisi horretako funtzio sintaktikoetan oinarrituz sintagmak eta aditz-kateak ezagutzeko gramatikak idatzi ditugu. Lan hau burutzerakoan, azaleko sintaxiaren arloari heldu diogu —ikus § 1.4.2—, eta horrez gain, erronka bat dugu, hots, ikustea zenbateraino den egokia EUSMG oro har, eta, batez ere, aditzen argumentu-egiturak erdiesteko.

Esan beharra dago proiektuaren sorrera ez dela hutsetik abiatu. Tarte bat egingo diegu ondoan proiektuak aurrera egiteko baliatu ditugun oinarrizko tresna nagusien azalpenari:

- Euskararen Datu Base Lexikala (EDBL), (Agirre eta beste, 1995). EDBLk gaur egun 73.000 sarrera biltzen ditu. Arloko lan desberdinetan beharrezko diren lexikoien euskarri eta iturri. Horrela bada, analizatzaile morfologikoaren irteeran edota analisi sintaktikoan aurkeztuko diren ezaugarriak EDBLn egongo dira adieraziak.

- Analizatzaile Morfologikoa (Alegria eta beste, 1996) eta (Urkia, 1997). Bi mailatako morfologia[1] jarraitzen duen analizatzaile morfologikoa tresna sendoa dugu. Hitz-forma orori posible dituen interpretazio guztiak eta hauei dagokien informazioa esleitzen dizkio. Emaitza gisa izango dugu hitzaren zatikatze morfologikoa, eta morfema bakoitzari lexikoian dagozkion ezaugarriak: kategoria, azpikategoria, mugatasuna, numeroa, kasua, etab.

- CG-2 parserra (Tapanainen, 1996) baliatu dugu landu dugun euskararen gramatika (EUSMG) aplikatu ahal izateko. Parser honek *Constraint Grammar* formalismoa (Karlsson eta beste, 1995) jarraitzen duten gramatikak aplikatzeko balio du. Formalismo sintaktiko hau analisi morfologikoan oinarritzen da, eta ezaugarrien artetik azpimarratzekoak dira: anbiguotasun morfosintaktikoaren arazoari ematen dion erantzuna eta testu errealak analizatzeko helburua izatea.

---

[1] Koskenniemi K.. Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production. Doktore-tesia, Helsinkiko Unibertsitatea. Pubs. no. 11. 1983.

Aurreko atalean azaldu dugun bezala, helburu nagusia, batetik, ahal den neurrian, aditzen oinarrizko argumentu-egitura jasotzea da; eskuratzen den informazio hori Euskararen Datu-Base Lexikalean (EDBL) egokitzeko adierazpide egokia aurkituz. Eta, bestetik, informazio hori eskuratzeko baliabideak lantzea dugu. Era berean, ez da ahaztu behar informazio hori eskuratzeko euskararen analisia (parsing-a) ere garatzen dela. Beraz, helburuei begira eta lehenago aipatu ditugun tresnak baliatuz, ondorengo puntuetan zerrendatzen direnak dira eginiko lanak:

1) Batetik analisia egitea bera, hots, gramatika konputazionalki lantzea —ikus § IV.8. eta V.1.4tik V.1.7ra bitartekoak—.

2) Bestetik ezagutza lexikala eskuratzearena, hau da, aditzen argumentu-egitura, ahal den neurrian, zehaztea —ikus § VI.3—.

3) Analisiaren emaitza bera (ikus C eranskina), hots, sintaktikoki etiketaturiko corpusa erdiestea. Corpus hau erabilgarria izan daiteke esate baterako, subjektu bezala gauzatu daitezkeen paper tematikoak aztertzeko, horrela bada subjektu etiketaz galdetzea (@SUBJ) interesgarriagoa suertatuko zaio hori aztertzen ari den linguistari izen-sintagmaz galdetzea baino (zeinak objektu zuzenak zein zeharkakoak, subjektuak, etab. biltzen dituen).

   Hautapen-murriztapenak (*selection restrictions*) eskuratzeko autore anitzek erakutsi dute sintaktikoki analizaturiko corpusen erabilgarritasuna, hala nola, beste batzuen artean: Grishman & Sterling (1992), Poznanski & Sanfilippo (1993), Resnik (1993), Ribas (1994).

4) Adibideen analisiaren emaitza Testuak Markatzeko Lengoaia Estandar eta Orokorrez (*Standard Generalized Markup Language,* SGML) kodeturiko ezaugarri-egituren bidez errepresentatzea. Horrela, adibideen analisien emaitza testu huts izatetik formatu aberatsago batera moldatzean, hauetan dagoen informazioa eskuratzeko bideak irekitzen dira; gure lanaren emaitzaren berrerabilgarritasuna bermatuz —ikus § V.1.7—.

5) Galdeketa-sistemaren diseinua. Galdeketa-sistemaren bidez, analisietatik eskuratu nahi dugun informazioa jasotzeko eta ikerketa errazteko bidea landu dugu —ikus § V.1.8—.

Tesi-proiektu honen motibazioak eta helburuak deskribatu ondoren, garbi dago tesi hau Lengoaia Naturalaren Prozesamenduaren (LNP) edota Linguistika Konputazionalaren arloan kokatzen dela. Nahiz eta euskal hizkuntzalarien artean hedatuegi ez dagoen ikerrarlo bat izan, hizkuntzaren industria oso bat sortzen ari da, ordenagailuaz baliatuz hizkuntza tratatzea helburu duena. Hizkuntzaren teknologiaz hitz egiten da dagoeneko. Teknologia horren oinarrian

ikerkuntza dago, hizkuntzaren tratamendu automatikoaren arloko ikerkuntza, alegia. Esan beharra dago, alor hau batez ere ingelesaz arduratu dela, eta ingelesa ezean hizkuntza nagusiez kezkatu dela.

Oro har, arlo honek aurrera egin dezan hizkuntzalari eta informatikoen arteko elkarlana sendotu beharra dago. Hau da, ez dadila gerta linguistikaren esparruan arlo hau nolabait isolatuta egotea edota, zenbait kasutan, linguistikak espero zitekeen baino eragin ahulagoa izatea LNPn. Horrekin batera, jende gehiago jarri beharra dago lanean, oinarrizko lan asko baitago egiteko; hizkuntzari bereziki lotutako lana, hizkuntzaren arloko ikerkuntza eskatzen duena.

Bestalde, lan-martxa lastertu beharra dago, alor honek daraman abiada izugarriaren aurrean atzean ez baldin badugu geratu nahi.

Amaitzeko, argi utzi nahi dugu, lan hau diziplinartekoa dela, informatikarien eta hizkuntzalarien jakintza-arloak elkarren osagarri baitira, lan egiteko ikuspegiak zabalduz eta ezagutza aberastuz. Gure iritziz, uste dugu linguistontzat asko dagoela irabazteko ordenagailuak hizkuntzen azterketarako eskaintzen dituen aukeretatik. Hona hemen John Sinclair-ek (1991:379) zer dioen:

> "The advent of computers has improved the quality of many scientific disciplines in recent years, but in none of them is the effect so profound as it will be in the study of language. For linguistics will see quite new methodologies and argumentations, and the relationship between speculation and fact will alter sharply."

## I.4 Landutako metodologiaz.

Aurreko ataletan aurkeztu ditugun motibazioei eta helburuei erantzuteko landu dugun metodologiak bi parte nagusi ditu:

### I.4.1 Hiztegiaren prestatze-lana.

Euskal Hiztegiaren azterketa eta egituratzea burutzeaz arduratu gara, ezagutza lexikalaren eskuratze automatikoari begira. Horrela bada, MRDetatik informazioa eskuratzerakoan egin beharreko urratsak burutu ditugu —ikus § III.2—. Azpimarratu beharra dago, urrats hau ezinbestekoa dela ondoren egin dugun azterketa ahalbideratzeko. Horrez gain, Euskal Hiztegian gorderik dagoen informazio lexikala kontsultatu eta aztertu nahi duen ororen eskura jartzeko, eredu estandarrera egokitu dugu hiztegia SGMLz errepresentatuz eta TEIko kodeketa-gidalerroen gomendioak jarraituz.

## I.4.2 Azaleko sintaxia.

Aurreko atalean esan dugun bezala, metodologia lantzerakoan azaleko sintaxiaren arloa ere lantzen dugu. Aditzen azpikategorizaziorako informazio bila aritu baikara *Constraint Grammar-eko* (CG) formalismoa baliatuz EUSMG lantzeko. Abney-k (1997:129)-n*Constraint Grammar* (CG) formalismoa azaleko sintaxiaren alorrean kokatzen du:

> "Voutilainen (Karlsson et al., 1995) describes a partial parser, ENGCG, that is very similar in operation to the constraint-grammar tagger. Lexical and morphological analysis assigns a set of possible syntactic function tags to each word, in addition to part of speech."

Azaleko sintaxia edota *shallow syntax* (*partial parsing* gisa ere ezaguna da) terminoa erabiltzen da orokorrean, ohiko parserren[1] irteerako analisiak bezain osoak ez diren analisiez aritzerakoan. Kontua da, testu errealak erroreak dituztelako, eta lexikoien eta gramatiken ezosotasunagatik, oso zailak direla analizatzen. Horrez gain, esaldien luzerak eta gramatiken anbiguotasunak ere arazoak sortarazten dituzte. Arazo horien aurrean, *shallow parsing* edota *partial parsing*-eko teknikek nahiago dute benetako testuetatik informazio sintaktikoa jasotzea eraginkortasunarekin eta ziurtasunarekin, analisi osoa eta sakona erdiestea baino. Ideia da informazio sintaktiko gutxirekin jaso daitezkeen egitura sintaktikoak jasotzea, informazio lexikal gehiago behar duten egitura sintaktikoak jasotzeke utziz.

Azaleko analisia egiterakoan ez da sintagma-egiturako zuhaitzik erdiesten. Azaleko analizatzaile batek zenbait sintagma-egitura (adibidez izen-sintagmak) identifika ditzake, hauen barne-egiturarik eta esaldian duten funtzio sintaktikorik zehaztu gabe. Parserrak arazo guztiak ebazteko gaitasunik ez baldin badu, analisi partzialak emateko aukera izango du.

Lehenago aipatu dugun bezala CG formalismoa (formalismo honen aurkezpen orokorra IV. kapituluan egingo dugu, —ikus § IV.4—) jarraituz landu dugu EUSMG. Eta, EUSMGko azaleko analisitik abiatuz, areago jo dugu sintagmak eta aditz-kateak ezagutzeko lana burutuz[2]—ikus § V.1.6—. Hau da, EUSMGko analisi sintaktikoaren emaitzak ez du sintagma-egitura espliziturik, zeren eta CG formalismoa jarraituz, CGk esaldian hitzek dituzten funtzio gramatikalak eta beraien arteko interdependentziak adierazten ditu. Ez du esplizituki zehazten sintagma tipoko osagaien hierarkiarik.

---

[1] 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* gisa ezagutzen dira. Beraz, *parser* edota analizatzaile sintaktikoa esan dezakegu. Bestalde, *parsing* terminoaren bitartez analisi sintaktikoari egiten zaio erreferentzia.

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

Baina, errepresentazio horretako funtzio sintaktikoen etiketetan oinarrituz, zenbait egitura sintagmatiko esplizitu egin ditugu. Horrela bada, sintagmak eta aditz-kateak jasotzeko gai gara. Eta kate horiek jasotzea baliagarria izango da azpikategorizazioari buruzko informazioa erdiesteko. Informazio hori eskuratzeko garatu dugu azaleko sintaxiaren atala, eta urrats hori aurrera eramatean *bootstraping* lexikala burutzen dugu. Garaturiko bide honen ebaluazioa egingo dugu V. kapituluan —ikus § V.1.6.3 eta § V.1.6.4—. Urrats honen ebaluazioa oso garrantzitsua da azken kapituluan ikusiko ditugun emaitzak ulertzeko.

## I.5 Txostenaren eskema.

Ondokoan, txosten hau bi parte nagusitan dago banatuta.

Lehenbizikoan, II. kapituluan, eta laburki bada ere, LNPren eta Linguistika Konputazionalaren nondik-norakoak aurkeztu ondoren, LNPrako osagai lexikalak sortzeko premia azpimarratuko dugu. III. kapituluan azalduko ditugu EH iturri lexikal gisa aprobetxatzeko bete behar izan diren oinarrizko urratsak. Kapitulu horretan deskribatuko dugu ondorengo azterketak burutzeko ezinbestekoa den lan hau egiteko jarraituriko urratsak: iturriaren ezaugarrien azterketa, etiketatzea, analisirako gramatikaren idazketa, analisia bera eta analisiaren emaitza TEI formatura egokitze-lana. EH TEI formatura egokitu ondoren, interesatzen zaigun adibideen eremua aztertuko dugu, aditzen adibideena.

Bigarren partean, EHko aditzen adibideen azterketaren aurrekariak aurkeztuko ditugu IV. kapituluan. Horrela, bada, azpikategorizazioa zenbait teoria sintaktikotan eta LNPko zenbait lexikoitan nola tratatzen den ikusi ondoren, azpikategorizazio automatikoaren markoa gaingiroki deskribatuko dugu. Ondorengo puntuan, argumentu-egituraren kontzeptua oso zabala izanik, gure ustez argumentu-egituratzat zer har dezakegun azalduko dugu, horretarako zenbait autoreren iritzietan oinarrituz. Jarraian, gain-gainetik bada ere, analisi sintaktikorako joera nagusiak deskribatuko ditugu. Ondoren, aditzen adibideen azterketarako metodologian oso garrantzitsua den formalismoaren aurkezpen orokorra egingo dugu, hots, *Constraint Grammar* (CG) formalismo sintaktikoarena. Eta IV. kapitulu honi bukaera emateko, CG formalismoaz baliatuz euskararako landu dugun Murriztapen-gramatika (EUSMG) deskribatuko dugu.

Hurrengo kapituluan, V.enean, azterketa burutzeko garaturiko metodologia deskribatuko dugu. Horrez gain, garaturiko metodologiaren ebaluazioa egingo dugu kapitulu honetan.

EHko aditzen adibideen azterketarako landuriko metodologiaren bidez lorturiko emaitzak aurkeztuko ditugu hurrena, VI. kapituluan. Hala ere, kapitulu honetan emaitzak ez ditugu emango oso-osorik, baizik eta multzo esanguratsuenak eta horiek deskribatzeko zenbait adibide. Beraz, emaitza guztiak ikusi nahi dituenak, jo beza C eranskinera.

Txostenaren azken partean berriz, etorkizuneko lana eta ondorio azpimarragarrienak izango ditugu mintzagai —VII. eta VIII. kapituluetan—. Bertan azalduko ditugu, lan honen ondoren ikergai geratutakoak eta, gure ustez, aurrera egiteko bide interesgarrienak diratekeenak.

Bestalde, txosten honen eranskinak aparteko liburuki batean bildu dira:

- A eranskina: hiztegiaren egitura jasotzen duen DCG (*Definite Clause Grammar*) gramatika.
- B eranskina: kategoria-sistema, analisiak ulertzeko laburtzapenen azalpena, lokuzioen multzoa, zatien osaketarako gramatikak eta EHko DTDa (*Document Type Definition*[1]).
- C eranskina: adibideen analisien emaitzak.

[1] Dokumentu-motaren nozioa SGMLk (Standard Generalized Markup Language) ezartzen du: dokumentu bakoitza mota batekoa izango da eta dokumentu mota batek dokumemtu-multzo bat definituko du. Dokumentu-mota hau DTD (*Document Type Definition*) delako fitxategian definitzen da eta dokumentu guztiek DTD bati esleituak egon behar dute.

# LEHEN PARTEA: EUSKAL HIZTEGIA TEI GIDALERROEN ARABERA ERREPRESENTATZEKO PROZESUA

## II. Hiztegi arruntak Lengoaia Naturalaren Prozesamendurako (LNP) baliabide lexikal gisa.

Egin dezagun kontu giza erabiltzaileari zuzenduriko hiztegi arrunt bat daukagula MRD formatuan eta horren baitan gordetzen den informazioaz baliatu nahi dugula LNPrako osagai lexikalak aberasteko. Behinolako burutazioa da hori, tesi-proiektuaren motibazioak eta helburuak zehazteari ekin genionekoa. Hastapenekoa izanagatik, ordea, burutazio horrek ondo bereizten ditu tesi-proiektuaren bi alde: batetik iturri lexikalaren prestatze-lana, bestetik informazio lexikalaren eskuratze-lana. Tesiko lehen partean, proiektua kokatzen deneko esparru orokorra aurkeztuko dugu lehenbizi, eta ondoren, prestatze-lanez jardungo dugu. Bigarren partean, aldiz, informazio lexikala eskuratu ahal izateko garaturiko azaleko sintaxiaz eta hiztegiko aditzen adibideei aplikatu diogun metodologiaz arduratuko gara batik bat.

Lehen partea bi kapitulutan banatu dugu. Lehenengoan, II. kapituluan, oro har, Euskal Hiztegiaren (EH) azterketa kokatzen deneko esparrua aurkeztuko dugu. Horrela bada, kapitulu honetan, eta laburki bada ere, lehenbizi linguistika konputazionalaren bilakaera aurkeztuko dugu, 50. eta 60. hamarkadetatik hona arlo honek lexikoarekiko jarreran izandako aldaketa nabarmenduz eta LNPrako tamaina errealeko (ez jostailuzko) osagai lexikalen premia azpimarratuz. Ondoren, II.2 atalean lexikografia konputazionala zertan den azalduko dugu —ikus II.2.1 puntua—, arlo honetan kokatzen baita gure lana. Eta jarraian, informazio lexikalaren eskurapenaren —ikus § II.2.2— eta errepresentazioaren —ikus § II.2.3— gaiei

lotuko gatzaizkie. II.3 atalean baliabide lexikalek gaur egun duten garrantziaren lekuko zenbait proiektu aipatu eta hauen sailkapen bat jasoko dugu. Horrez gain, II. 4 atalean, bereziki, halako iturrien estandarizazioaren premiaz arituko gara. Horrela bada, labur-labur *Text Encoding Initiative*-ren (TEI) ikuspegi orokorra, eta, batez ere TEIko hiztegigintzarako proposamenak aurkeztuko ditugu. Giza erabiltzaileari zuzenduriko hiztegi elebakar zein eleanitzak kodetzeko oinarrizko etiketa-multzoa proposatzen da TEI gidalerroetan. Hauetan batez ere kodeketa-arazo orokorrak planteatzen dira, kontuan izan behar baitira hiztegien konplexutasuna (bai tipografikoki bai egitura aldetik) eta erabiltzaile desberdinen interes gatazkatiak.

Azkenik, II.5 atalean MRDez baliatzeari zer deritzoten linguistika konputazionaleko zenbait ikertzailek, eta MRDetatik zer-nolako etekinak atera daitezkeen erakustearren, MRDez baliatuz eginiko zenbait lanen berri emango dugu.

Bigarrengoan, III. kapituluan, aldiz, LNPrako "ezusteko" baliabide lexikal honen prestatze-lanez arituko gara.

## II.1 Linguistika konputazionala: sarrera gisakoa.

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari Lengoaia Naturalaren Prozesamendua (LNP) esaten zaio, nahiz eta batzuetan, hizkuntzalaritzako ikuspuntua garrantzitsua denean batez ere, Linguistika Konputazionala (LK) ere esaten zaion. Bi motibazio nagusi biltzen dira bere baitan: teknologikoa eta linguistikoa. Lehenengoari dagokionez, konputazio-sistema adimentsuak garatzea du helburu, hala nola, datu-baseei lengoaia naturalez galdetzeko interfazeak, itzulpen automatikorako sistemak, testuen analisirako parserrak, ahotsaren tratamendurako tresnak, etab. Bigarrenean, alderdi linguistikoari erreparatzen zaio, eta gizakiok elkarrekin harremanetan jartzeko erabiltzen dugun hizkuntzaren ulermen sakonagoa erdietsi nahi da. Bi alderdi hauek kontuan harturik, linguistika konputazionalaren xedea honako hau da: hizkuntz ulermenaren eta sorkuntzaren teoria konputazional ulergarri, taxutu eta linguistikoki motibatua eraikitzea.

Jakina, linguistika teorikoak eta psikolinguistikak ere badute alderdi linguistikoari dagokion kezka hori. Dena dela, kezka bera izan arren, hurbilpen, metodo eta helburuei erreparatuz gero, zenbait desberdintasun aurki ditzakegu hiruen artean:

- Helburuei dagokienean, psikolinguistikak gizakien prozesua den hizkuntza naturalaren ulermena eta sorkuntza aztertuko ditu. Linguistika teorikoak, aldiz, teoria gramatikal dotore, murriztua eta unibertsal linguistikoen berri emango duena du helburu. Linguistika konputazionalak, berriz, funtzionatzen duen sistema eraiki nahi du, hots, egitura linguistikoa eraginkortasun konputazionalarekin prozesatzea du jomuga.

- Metodo edota tresnei dagokienean, psikolinguistika esperimentazioan oinarritzen da. Linguistika teorikoa hiztunen konpetentzia aztertzeaz arduratzen da bereziki, eta datuak erdiesteko, batez ere, introspekzioa du iturri nagusi. Ondorioak metodo deduktiboen bidez erdiesten ditu. Linguistika konputazionala, berriz, erabilera linguistikoan zentratzen da, komunikazio-egoera errealetatik datuak erdietsiz. Ikertzeko metodo deduktiboak zein induktiboak baliatzen ditu. Oro har, esan dezakegu linguistika teorikoa teoriaren dotoretasunaz arduratzen dela gehiago, eta linguistika konputazionala, aldiz, sistemaren eraginkortasunaz. Horrela bada, eraginkortasunari begira, fenomenoen deskribapen oso esplizitoak, eta formalki zein konputazionalki ahalik eta sinpleenak garatuko dituzte linguista konputazionalek.

Dena dela, linguistikaren baitan, teoriak garatzeko datu objetiboetatik abiatzen diren hurbilpenek gero eta indar handiagoa dute. Ildo honetatik, aurrerapen handiak lortu dira ordenagailuen erabilerari esker, esate baterako, corpusean oinarritutako linguistikan.

Bestalde, (Sparck, 1996:14)-n esaten denez, oro har, linguistikaren, eta bereziki lingusitika teorikaren eragina LNPn oso ahula izan da. Horrez gain, artikulu horretan, Sparck Jones-ek esaten du informazio-teknologien eragina linguistikan, LKtik kanpo osoa zaila dela aurkitzen, eta uste duela linguistikak asko duela irabazteko arlo konputazionaletik:

> "..., there is much for linguistics to gain from looking both at how computation does things and at what it finds."

LNPak izan ditzakeen hurbilpen nagusiak ikusi ondoren, LNPak izan duen bilakaera izango dugu mintzagai, labur-labur, hurrengo puntuan.

## II.1.1 LNParen bilakaera.

Hastapenetan, lengoaia naturalaren prozesamenduaz (LNP) arduratzen zirenak (50 eta 60ko hamarkadetan), aplikazio zehatzetara mugatzen ziren batez ere, aplikaziotik aplikaziora helburuak aldatuz. Bi aplikazio-multzo nagusi nabarmentzen ziren :

1. Alde batetik gizaki eta ordenagailuaren arteko komunikazioa errazten dutenak :
   - datu-baseen galdeketa-sistemak.
   - elkarrizketarako interfazeak.
2. Giza komunikaziorako aplikazioak :
   - testuen eduki-araketa.
   - testu-edizioa.
   - itzulpen automatikoa.
   - ahozko idazmakina.

Sistema konputazional gehienek jostailuzko lexikoiak lantzen zituzten, oso aplikazio-domeinu konkretuei lotuak eta sarrera kopuru murritzekoak. Askotan zerrenda soilak baino ez ziren izaten. (Boguraev eta Briscoe, 1989:1)-n esaterako, hau diote:

> "Knowledge of words underlies these tasks, yet until very recently dictionaries (or lexicons, as linguists usually call them) for natural language processing systems have by and large been the poor sisters of computational linguistic research."

Bestalde, oro har, teoria linguistikoek sintaxia eta erregela gramatikaletan jartzen zituzten beren indarrak.

70 eta 80ko hamarkadetan LNParekiko interesa areagotzeaz gain, azpimarratzekoa da epe horretan hurbilpen-aldaketa gertatu zela. Hau da, alderdi linguistikoan arreta handiagoa jarri zen. Hasieran, alderdi linguistikoak ez zuen garrantzi handirik, eta aurrerago aipatu dugun legez, aplikazioetara lerraturik zegoen linguistika konputazionala. Horrela bada, garaturiko hainbat sistema aplikazio espezifikoetarako baino ez ziren baliagarri. Honen ondorik garaturiko beste bi joera nagusi ere aipatu beharrekoak dira. Batean, ordenagailuaz baliatuko dira modelo linguistiko teorikoak frogatzeko. Hots, teoriak sortarazitako sistemak ditugu, eta garatu izan dira teoria desberdinak frogatzeko; beste batzuen artean: gramatika transformazionalak (*Transformational Grammars*) (Friedman, 1969), Montague-ren gramatikak (Friedman, 1978), *Generalized Phrase Structure Grammars* (GPSG) (Evans, 1985, Phillips eta Thompson, 1985). Proposaturiko eredu gramatikalak sortu behar lituzkeen esaldiak sortzen dituen konprobatu nahiko dute ikuspegi hau lantzen dutenek. Ondorengoan, joera nagusia (egun ere dirauena) corpusetan oinarritzean datza. Hauen artean ere ikuspegi desberdinak aurki ditzakegu, jakina, baina denak bat datoz hizkuntza aztertzeko corpusak ordenagailuez baliatuz ikertzerakoan.

Gaur egungo joera ordea, hastapenetakoarekin alderatuz gero, erabat aldatu dela esan dezakegu. Linguistika teorikoaren zein LKren egungo joera hizkuntz ezagutza gramatikaren arlotik lexikoarenera lerratzea baita. Teoria linguistikoan eragin handiena izan duten formalismoak (UG, LFG, HPSG, etab.) erregela gramatikalak erraztera jotzen dute, eta lexikoa muina izango dute. Alderdi teorikoari dagokionez, segur aski Chomsky-k eman zion abiada joera honi (Chomsky, 70). Ildo beretik jardungo dute aplikazioei loturiko LNPko arlokoek ere. Hau da, sistema errealetarako ezagutza lexikala eskuratzea ezinbestekotzat jotzen da laborategiko saioak gaindituz arlo honetan aurrera egin nahi bada. LNPko sistemek neurri errealeko osagai lexikalak behar dituzte, beraien aplikazio-eremua hedatu eta sendotzeko. Baina osagai lexikal horiek eskuz egitea hain da lan handia, ezinezko baita ia. Horrela bada, LNPko aplikazioen problemarik larriena lexikoi konputazionalak hornitzeko ezagutza lexikalaren eskuratze-prozesuak garatzean datza. Gauzak horrela, LNPrako lexikoien eraikuntzarako laguntza automatikoak garatzea eta dauden baliabide lexikalez baliatzea dira harturiko irtenbide nagusiak.

Bestalde, hizkuntzen teknologiako aplikazioak diseinatzerakoan, ikertzaile asko datu estatistikoek gidaturiko metodoetara lerratu da. Zenbait hamarkadatan, egitura kognitiboak eta giza hizkuntzaren erabiltzailearen prozesuen azterketatik teknologiak aurrera egin zezakeelako itxaropena izan ondoren; gizakiek sorturiko datu linguistikoetan eta hizkuntzaren teknologiak prozesatu beharrekoetan jarri dute ikusmira. Horrez gain, azken urteetan azkar garaturiko arlo berezi bat, hizkuntza mintzatuaren zein idatziaren prozesamendurako, optimizazio-tekniken erabilpenarena da, batez ere, metodo estatistikoekin batera. Optimizazio-metodoak posible diren soluzio multzo batetik soluzio edota soluziorik hoberena aurkitzeko erabiltzen dira, horretarako zenbait ebaluazio-irizpide aplikatuz.

## II.1.2 LNPko sistema baten osagaiez[1].

Modu eskematiko batez adierazita, LNPko sistema batek, honako modulu hauek ditu edo izan ditzake:

- Ezagutza fonetikoa eta fonologikoari dagokiona. Ahotsaren tratamenduaz aritzerakoan, bi sistema nagusi garatzen dira: hizketaren ezagumendua (*Speech Recognition, SR*), eta sintesia edo sorkuntza.

- Ezagutza morfologikoa. Aplikazio batzuetan ez da beharrezkoa, adibidez ingeleserako askotan ez da kontuan hartua izaten (hala ere, bada bestela pentsatzen duenik ere). Baina, morfologia aberatsa duten hizkuntzen prozesamenduan oso garrantzitsua da, esate baterako, euskara, suomiera, etab.

- Ezagutza sintaktikoa: esaldien egitura ezagutzeaz arduratzen da.

- Ezagutza semantikoa: analizaturiko esaldiei esanahia ematen zaie.

- Testuinguruari dagokiona, pragmatika gisa ezagutzen dena. Berez linguistikoa ez den, eta igorpen linguistikoen prozesamenduan eta interpretazioan eragina duten informazioez arduratzen da. Bi atal bereiz daitezke:

  1) Diskurtsoaren ezagutza: hemen lehenago igorri diren esaldien interpretazioez arduratzen dira. Anaforari dagozkion arazoak, eta denborari dagozkion ezaugarriak tratatzen dira beste batzuen artean.

  2) Munduaren ezagutza: hizkuntza bateko hiztunek elkarren artean komunikatzerakoan, munduari buruz duten ezagutza kontzeptual guztia hartu behar da kontuan. Horrelako

---

[1] Joerak azkar ari dira aldatzen, eta jasotzen diren atalak tradizionalki jasotzen direnak dira. Bereziki, atal hau lantzeko ondoko liburu hau baliatu dugu: *Survey of the State of the Art in Human Language Technology*, Edited by Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Batista Varile, Annie Zaenen eta Antonio Zampolli. Cambridge University Press. 1998. Autore askoren artean idatzitako liburu honetan LNPri buruzko ikuspegi zabala jasotzen da.

ezagutzak esaldietan esplizituki adierazten ez den, eta bistan den informazioa ulertzeko balio du.

Jakina, goian aipatu ditugun modulu horietako informazio guztia ez da beharreazkoa lexikoi espezifiko baterako. Lexikoi bakoitzak LNPko sistema espezifiko baterako beharren araberako informazioa izango du eta.

Tradizionalki esaten da hizkuntza ulertzeko lana bi zatitan banatzen dela: sintaxia eta semantika.

Sintaxiaren arloan, nolabaiteko bi muturretako saioak aurki daitezke sintaxiari ekiterakoan: batetik, oso modelo gramatikal sinpleak erabiltzen dira, adibidez, egoera finituko gramatikak, zeinak prozesu oso erangikorren oinarrian dauden (hurbilpen horietan, gramatikak saihestuz, metodo estatistikoak baliatzen dituzte patroi linguistikoak lortzeko); eta, bestetik, linguistikoki sofistikatuak diren errepresentazio-formalismo ahalmentsuak ditugu. Sofistikatuetariko formalismo erabilienak baterakuntzan oinarritutako formalismoak ditugu. *Functional Unification Grammar* (FUG), *Head-Driven Phrase-Structure Grammar* HPSG), *Lexical Functional Grammar* (LFG), *Categorial Unification Grammar* (CUG) eta *Tree Adjunction Grammar* (TAG). Modelo gramatikal hauetarako edota antzekoetarako, bai gramatiken garapenerako bai prozesamendu linguistikorako formalismo ahalmentsuak diseinatu eta inplementatu dira, adibidez: *LFG* (Kaplan eta Bresnan, 1982), *PATR* (Shieber, 1986), *Attribute Logic Engine* (ALE) (Carpenter eta Penn, 1994), etab.

Bi muturretako formalismo horiek nabarmentzeaz gain, analisi morfologikoak ere garrantzia hartu duela esan daiteke. Azken hamar edo hamabost urteetan morfologia konputazionalak gainontzeko LNPko arloek baino aurrerapen handiagoak erdietsi ditu aplikazio errealei begira. Morfologia konputazionalaren alorrean, sakontasun teorikoetan sartu gabe, aurkitzen diren arazo nagusiak honako hauek dira: alternantzia morfologikoak eta morfotaktika.

Lehenengo arazoari aurre egiteko hurbilpen bat *cut-and-paste* metodoa izan da. Sistema horren aitzindarietako bat 1960an garatu zen: *MITeko DECOMP*. Aurrerantzean, egoera finituko teknologian oinarritutako hurbilpenak garatu dira, adibidez, bi mailatako morfologia (Koskenniemi, 1983).

Bestalde, sintaxiari ekiterakoan, analisi morfologikoaz gain desanbiguazio morfologikoa oso inportantea da. Bi hurbilpen nagusi daude: erregeletan oinarrituak eta probabilitateetan oinarrituak. 1980 urtetik metodo probabilistikoak edo estokastikoak nagusi izan dira, eta goiztiarrenetariko bat honako hau dugu: *Constituent-Likelihood Automatic Word-tagging System* (CLAWS). Sistema hori garatua izan zen ingelesezko corpus bat (*Lancaster-Oslo/Bergen Corpus of British English* ) etiketatzeko.

Sistema estokastiko gehienek eskuz etiketaturiko probarako corpus batetik eratortzen dituzte probabilitateak. Bestalde, badira eredu markoviarrean oinarritutako etiketatzaileak, etiketatu gabeko corpusetan entrenatzen direnak ere. Horrelako batzuek %96ko arrakasta erdietsi dute ingeleserako.

Ingeleserako egin diren etiketatzaile probabilistiko gehienek, nahiz eta metodo desberdinak jarraitu izan dituzten, pareko doitasuna izan dute azken hamar urteetan. Doitasun hori erregeletan oinarritutako etiketatzaile batek (*Constraint Grammar* formalismoan oinarritutako ingeleserako etiketatzaileak) hobetu du. Etiketatzaile horren doitasuna %99,7koa da, %2 eta %6 bitartean anbiguotasun arazo latzenak dituzten hitzak desanbiguatzeke utzirik. *Treebank*, hots, sintaktikoki etiketaturiko corpusa eskura izateak, desanbiguazio estatistikorako hurbilpenak gehitzea ekarri du.

Arlo honetan jarraitzen den beste korronte bat da prozedurak garatzea, etiketatzeko erregelak automatikoki inferitzeko corpusetatik. Horrela, hein batean analizatzaile morfologikoak eraikitzeko lana automatizatu egingo litzateke. Ildo honetatik, gero eta garrantzi handiagoa du corpusetan datuak egiaztatzeko metodo malguak eskura izateak.

Horrez gain, metodo estokastikoak eta erregeletan oinarritutakoak konbinatzeko hurbilpenak garatu beharra ere sumatzen da etorkizunari begira. Metodo biak beharrezkoak ikusten dira testu errealak etiketatzeko.

Semantikari dagokionean, aurretik dagoen arazoa da arlo honi dagokion gaia zein den mugatzea. Adiera terminoa modu askotara erabil daiteke eta. Modu estandar batean, konputazionalki landuriko semantikan esaten da esaldi baten esanahia ulertzea esaldi horren egiazko baldintzak ezagutzearekin pareka daitekeela. Horren ondorioz, esaldi baten egiazko baldintzak ezagutzea, esaldi horretatik zer inferentzia den zilegi egitea ezagutzea da. Tradizionalki baliozko inferentzia horiek egiteko, logikak eta logika matematikoak eskaintzen dituzte teoria semantikoen garapenerako baliabideak. Sistema logiko partikular bat, semantikan paper berezia izan duena, lehen mailako predikatuen kalkuluarena dugu. Hala ere, zenbait aplikaziotarako metodo estatistikoak edota patroi-ezkonketa metodoak erabili dituzte; esate baterako itzulpen automatikorako eta informazioaren berreskurapenerako (*Information Retrieval*, IR).

Bestalde, lengoaia naturalaren sorkuntzarako beharrezkoa da prozesamendu semantikoa. Sorkuntzak esanahiaren errepresentazio egoki bat eskatzen du, eta esaldi edota esaldi-segida bat sortu, zeinak irakurle batek ulertu ahal izateko moduko eduki bat adierazi behar duen.

Semantikaren arloan errepresentazio-lengoaiak behar dira esaldi baten esanahia edota edukia errepresentatzeko. Inferentziak egiteko baliagarriak izan daitezkeen esaldien errepresentazio ezanbiguoak erdiesteko, logika izan da erabilia. Goi mailako predikatuen logika-mota

desberdinak erabili izan dira helburu horri erantzuteko. Errepresentazio-lengoaia bereziak, hala nola, *frame* eta *script* lengoaiak, adimen artifizialaren arlotik hartu izan dira. Helburu orokorreko eta espezializatuak diren inferentzia-teknikak ere erabiliak izan dira errepresentazioaren esanahia interpretatzeko, kontuan harturik testuinguru linguistikoaren ezagutza, egoerazko testuingurua eta mundua.

Teoria semantiko gehienetan zein inplementazio konputazionaletan esaldien interpretazioa ez da zuzenean ematen. Lehendabizi, sintaktikoki desanbiguaturiko esaldi bat lengoaia logiko artifizial baten bidez adierazten da, eta adierazpen horri interpretazio bat emango zaio lehen mailako predikatu-kalkuluaren antzeko erregelak erabiliz.

Lehenbiziko modu horretako semantikaren deskribapena Montaguek garatu zuen 1973an. Montague eta honek sorturiko tradizioaren ondorengo beste batzuentzat tartehizkuntza logikoa komenientzia kontu bat zen, zeina konposizionalitatearen printzipioa betetzen zela esanez onar zitekeen. Beste hurbilpen batzuentzan, adibidez, *Discourse Representation Theory* (DRT), tarteko mailako errepresentazioa teoriaren osagai beharrezkoa da, oinarri psikologikoek justifika dezaketena, edota beharrezkoa egiten dena erreferentzia esplizitua egin behar zaienean izenordeei edota beste erreferentziazko elementuei, hauen esanahia jasoaz. Inplementazio konputazionalen kasuan, arrazoi praktikoak direla medio errepresentazioa baldintza *sine qua non* da konputaziorako.

DRTn ere tarteko errepresentazioa beharrezkotzat jotzen dute, eta horrez gain, esaldia isolaturik hartu beharrean diskurtsoa hartzen dute kontuan.

*Dynamic Semantics*-en arloan garaturiko teoriek zenbait aldaera garatu dituzte esaldi baten esanahia, testuinguruan sortzen dituen aldaketekin pareka daitekeelako ideian oinarrituz.

*Situation Theory* semantika tradizionalenaren oinarri logikoak birplanteatzera etorri da, munduaren egoera partziala edota situazioaren nozioaren formulazio egoki bat aurkitzeko, eta proposizioaren nozio egokiago bat.

Oro har, eta salbuespen gutxi batzuekin, osagai semantikoa duten sistemen inplementazio gehienak Montagueren lanean (gehiago edo gutxiago) oinarritzen dira.

Estaldura sintaktiko osoko sistemen kopurua nahiko handia izan arren, oso sistema gutxik eskain dezakete estaldura semantiko osoa.

Gaur egun, semantika teorikoaren beharrik handiena da bideak bilatzea datu errealen estaldura zabalagoa eta sendoagoa erdiesteko. Horrek honako lan hauek eskatzen ditu: (i) errepresentazio-moduak aurkitzea, nahiz eta errepresentazio semantiko osoa ez erdietsi, mailaren bateko prozesamendu semantikoa lortzeko; (ii) elkarlan estuagoa lexikoi eraikuntzako

lanarekin; (iii) esaldi mailako semantikaren eta testuaren edota dialogoaren teorien arteko integrazioa.

Aipaturiko hurbilpen semantiko horiez gain, semantika lexikala oso garrantzitsua da. Semantika lexikalak hitzen semantika biltzen du, lexikoko elementuen artean dauden erlazio lexiko-semantikoak: sinonimia, antonimia, hiperonimia/hiponimia, erlazio meronimikoak, eta beste. Hiztegi arruntetan dagoen informazioa erauzteko hainbat saio egin dira. Lan horien helburua, gehienetan, lexikoko unitateen artean erlazio lexiko-semantiko horiek esplizituki errepresentatzea izan ohi da, azkenik sare semantiko moduko bat erdiesteko. Ingelesezko sare semantiko ezagunena-edo WordNet (Miller, 1990) izenekoa genuke.
Bestalde, diskurtsoaren arloko ikerketan trataturiko arazoek bi galdera orokorri erantzun nahi diete: (1) esaldi-multzoetan zer informazio da indibidualki esaldiei dagokiena? (2) esaldi bat erabiltzen deneko testuinguruak zenbateraino eragiten du esaldi baten esanahian, edota bere zenbait partetan?

Alderdi konputazionaletik bi arlotan lan egiten da gehien bat, testu zabaletan eta dialogoetan, idatzizkoetan zein mintzatuetan. Hauen artean badira zenbait ezaugarri bat datozenak. Baina aplika daitezkeen eremuak eta dituzten beharrak desberdinak dira. Testuetan egindako lanak eragin zuzena du dokumentuen analisian eta informazio-berreskurapen aplikazioetarako, aldiz, dialogoetako lanak gizaki eta makinaren arteko interfazeen garpenerako garrantzitsuak dira. Bai testuak bai dialogoak bi ezaugarri adierazgarri konpartitzen dituzte: (1) diskurtsoa diskurtso-ataletan banatzen dela esaldia sintagmetan baino argiago (2) adierazpen erreferentzialen azterketa: izenordeak, deskribapen zehatzak, eta aditz-sintagmaren interpretaziorako gertaera-erreferentzia barne direla.

Lengoaia naturalaren sorkuntzak (*Natural Language Generation*, NLG) honako helburu hau du: kalitate handiko lengoaia naturaleko testuak nola sor daitezkeen konputagailu barneko informazioaren errepresentazioetatik abiatuta aztertzea. Bestalde, sorkuntza zein ezagutza zailagoak suertatzen dira hizketaren kasuan. Zailago izatearen arrazoien artean, honako hauek aipa daitezke: ziurgabetasun handiagoa, erredundantzia, ulertutzat ematen den informazio gehiago, errepresentazio akustiko baten proiekzioa errepresentazio testual batera, etab.

### II.1.3 Aplikazioak.

Aurrerago aipatu bezala, lengoaia sortzea eta ulertzea oso prozesu konplexuak dira, eta gaur egungo ordenagailuak urrun ikusten du giza adimenaren ahalmen linguistiko orokorra.
Dena dela, badira aplikazio lagungarriak, esate baterako testuen edizioa eta gestioarekin lotutako aplikazioak:

- Informazio-erauzketarako sistemak
- Ortografia-zuzentzaile/egiaztatzaileak
- Sintaxi-zuzentzaileak

- Kategorizazio-sistemak
- Testu-sorkuntza automatikoa

Aplikazio horiez gain, linguistika konputazionalaren garapenean oinarrizkoak izan diren hiru aplikazio-mota honako hauek dira:

- Konputagailuaren erabilera LNren bidez.
  Gizaki bakoitza bere ama-hizkuntzaz komunikatu ahal izatea konputagailuarekin izango litzateke modurik idealena konputagailuarekin komunikatzeko, hau baita modurik naturalena, malguena, eraginkorrena eta ekonomikoena giza-komunikaziorako.

  Azken hamarkadan, hizketaren ezagupenerako teknologiak (*speech recognition technology*) aurrerapen handia egin du. Hizkuntz mintzatua ulertzeak bi oinarrizko lan eskatzen ditu: *speech recognition* (SR) eta lengoaia naturala ulertzea. Hala ere, gizakiaren hizketa ulertzeak ezagutza-iturri desberdinen integrazioa eskatzen du: munduaren edota testuinguruaren ezagutza, hiztuna ezagutzea eta topikoa, maiztasun lexikala, prosodia, hitz baten aurreko erabilerak edota semantikoki erlazionatutako topikoak, etab. Integrazio hurbilpen sinpleena dagoen *speech recognition* sistema bat eta lengoaia naturaleko sistema bat lotzea litzateke.

  Dena dela, arlo honetan aurrerapenak izan diren arren, oraindik teknologia urruti dago giza adimenaren ahalmen linguistiko horren parekorik lortzetik.

  Linguista konputazionalek modelo formalak garatzen dituzte giza adimenaren ahalmen linguistikoaren ezaugarriak simulatuz, eta konputagailu-programa gisa inplementatzen dituzte. Programa horiek teoriak ebaluatzeko, eta gehiago garatzeko oinarriak dira.

  Teoria linguistikoez gain, psikologia kognitiboak paper garrantzitsuagoa jokatzen du konpetentzia linguistikoa simulatuz. Psikologiaren alorrean, batez ere psikolinguistikak aztertu du giza-hizkuntzaren erabilera prozesu kognitibo gisa.

  LK aplikatuak giza-hizkuntzaren erabileraren modelizazioaren alderdi praktikoan jartzen ditu indarrak. Arlo honetako metodoak, teknikak, tresnak eta aplikazioak, maiz, hizkuntzaren ingeniaritza edota hizkuntz teknologia gisa ezagutua da. Nahiz eta LKko sistemak urruti dauden gizakien trebetasuna lortzetik, hainbat aplikazio daude. Helburu nagusia da giza-hizkuntzaren ezagumendua duten software-produktuak sortzea. Horrelako produktuek gure bizitzak alda ditzakete. Oso beharrezkoak dira gizaki eta makinaren arteko harremana hobetzeko, kontuan izanik arazo nagusia komunikazio-problema bat dela. Gaur egun, konputagailuek ez dute gure hizkuntza ulertzen, eta konputazio-lengoaiak zailak dira ikasteko, eta ez dute bat egiten gizakien pentsamenduarekin. Nahiz eta makinek ulertzen duten hizkuntza eta bere domeinua oso mugatua izan, giza hizkuntzaren erabilerak softwarea onargarriago bilaka dezake eta erabiltzaileen produktibitatea igo.

Lengoaia naturaleko interfazeek erabiltzaileari konputagailuarekin bere hizkuntzan komunikatzeko aukera ematen diote. Interfaze horietako aplikazio batzuk, datu-baseei galdetzea, testuetatik informazioa eskuratzea, sistema adituak eta robot-kontrola dira.

Hizkuntza mintzatuaren ezagutzan eginiko aurrerapenek hizkuntza naturaleko hainbat sistemaren erabilgarritasuna handitzen dute. Konputagailuekin hizkuntza mintzatuaren bidez harremanetan jartzeak informazio-teknologien aplikaziorako hainbat eremu berri irekiko lituzke.

Dena den, hizkuntza mintzatua bestelako komunikazio-moduekin ere konbinatu beharra dago, adibidez, saguarekin edo atzamarrarekin seinalatu ahal izatea.

Komunikazio multimodal bat kooperazio-modelo orokor batean txertatuz gero, konputagailua edota makina lagun bilakatzea erdietsiko genuke.

- Itzulpen automatikoa.

  Gizaki eta makinen arteko komunikazio-arazoak baino lehenago, ama-hizkuntza desberdineko gizakien arteko komunikazioa dugu. LK aplikatuaren jatorrizko helburuetariko bat itzulpen automatikoa izan da. Oraindik urruti daude helburu handi hori lortzetik, hots, edozein testu itzuli ahal izatetik. Hala eta guztiz ere, LKak sortu ditu software-sistemak giza-itzultzaileen lana errazteko eta produktibitatea igotzeko.

  1950 urte bukaera aldera hasi ziren arlo honetako lanak, oso esperantza handiarekin eta zailtasunez jabetu gabe. Aurkituriko arazoek bai linguistika bai linguistika konputazionaleko lanak bultzatu zituen, adibidez, hasierako parserrak, 1960 urtean. Lan gogorra egin zen arlo honetan, baina arrakastarik gabekoa izan zen. Konturatu ziren testuaren ulerkuntzan lana egin ezik, ezin zela erabateko itzulpen automatikoa eta kalitate handikoa erdietsi.

  Itzulpenaren automatizazioa ez da ia inoiz erabatekoa, eta automatizazio-mailaren arabera ondoko sailkapena egiten da:

  1) Erabateko itzulpen automatikoa: errealitatea baino ametsa da gaur egun, non eta helburua ez den edukiaren ideia orokorra ateratzea.
  2) Giza laguntzaz buruturiko ordenagailu bidezko itzulpena: lanaren gidaria makina da, baina fase desberdinetan laguntzak eska ditzake; hitz baten adiera zuzena hautatzeko edo esaldi baten analisia nondik hasi behar den erakusteko adibidez.
  3) Ordenagailuz lagunduriko giza itzulpena: lanaren gidaria pertsona da, baina ordenagailuaz baliatzen da hiztegi berezituan kontsultak egiteko, testuaren formatua txukuntzeko eta zailtasunik gabeko testu-zatiak itzultzeko. Agian itzulpenaren zati handi bat ordenagailuak egingo du ia laguntzarik gabe, baina testua egokitzeko aurreprozesaketa edota emaitza zuzentzeko postedizioa ohikoak izaten dira.
  4) Datu-banku terminologikoak: hiztegi berezituak erabiltzeko aukera hutsa eskaintzen duten laguntza-sistemak.

Itzulpenean aurrera egiteko arazo nagusietariko bat da esaldiak interpretatu egin behar direla, eta interpretazioa testuinguruak baldintzatzen duela. Pentsatu izan da esanahia eta interpretazioaren arteko desberdintasuna egitea ez zela praktikoa, eta murriztutako domeinuetan edota domeinu teknikoetan desberdintasun hori egitearen interesa desagertzen dela.

Itzulpen automatikoan hasieratik hona aldaketa gutxi eman dira. Orduko arazoek jarraitzen dute: 1. *Interlingual* eta *transfer* hurbilpenen arteko bereizketa; 2. Alderdi linguistikoaren aldean garrantzi handiagoa dute zentzuak eta ezagutza orokorrak.

Ireki den ikerkuntza lerro berria da: informazio-iturri nagusi bezala lehendik itzulita dauden materialak erabiltzea, berriak sortzeko.

Kontua da itzulpen automatikoaren arazoa ez dela soilik linguistikoa, baina gizakiak duen ezagutza orokorraren eta zentzuaren parte garrantzitsu bat izatea ere lortu ezina dirudi. Etorkizunerako bidea erabat automatikoa den itzulpenaren ordez, gizaki eta makinen artean egingo den lan bat bezala proposatzea litzateke.

- Informazioaren berreskurapena (Information Retrieval, IR).

  Kontuan izanik informazio asko lengoaia naturalean agertzen dela: liburu, aldizkari, etab., testuetatik informazioa automatikoki eskuratzeko interesa oso handia da.

  IReko sistemek galdera bati erantzuterakoan, corpus bateko informazio esanguratsuena aukeratu behar dute, testu hori erantzun gisa emanez.

Bukatzeko, esan beharra dago Interneten hazkunde azkarrak, eta informazioaren gizarteak dituen beharrek LKn aldaketa berriak ekar ditzaketela. Amaraunean hizkuntzen teknologien beharra dago: bilaketarako, nabigaziorako, iragazte eta informazioaren prozesamendurako. Horrez gain, amaraunaren eleaniztasuna gero eta handiagoa izateak ere, aldaketak eskatzen ditu hizkuntzen teknologian. Amaraun eleanitzak tresna eleanitzak eskatzen ditu indexatzeko eta nabigatzeko.

LKren abantailetariko bat, ikerkuntza oso teorikoa, eta oso aplikazioari begira dagoena integratzean datza. Bestalde, giza-zientzia, zientzia esperimentalak, humanitateak, eta ingeniaritzako adituen ezagumendua konbinatzeak erakargarri egiten du arlo hau. Hurbilpen zientifikoak eta teknika praktikoak, linguistikaren, konputazio-zientzien, psikologiaren eta matematikaren arloetatik datoz. Nahiz eta LKk dituen ikerketarako helburuak mota askotakoak diren, oinarrizko motibazioa lengoaia naturala muina duten sistema espezifiko praktikoen garapena izan da.

## II.2 Lexikografia konputazionala.

Atal honetan lexikografia konputazionalari sarrera labur bat eskaini ondoren, informazio lexikalaren eskurapenaz eta errepresentazioaz mintzatuko gara.

## II.2.1 Sarrera.

Lexikografia konputazionalaren alorrean hainbat alderdi desberdin bildu ohi dira: ordenagailuz lagunduriko lexikografia (CAL, *Computer Aided Lexicography*), hiztegien edukia automatikoki analizatu eta arakatzeko metodologia eta tresnak, hiztegien azterketa hizkuntzaren egitura semantikoa ikertzeko, eta LNPrako lexiko-sistemen eraikuntza eta horretarako laguntza automatikoak. Horiexek dira, besteak beste, lexikografia konputazionalaren baitan jorratzen diren bideetariko batzuk. LNPren aldetik egindako lexikografia konputazionalaren ikerrarloaren ikuspegi orokorra ematen zaigu *Computational Lexicography for Natural Language Processing* (Boguraev eta Briscoe, 89)-n. *Longman Dictionary of Contemporary English* hiztegiaren (LDOCE) gainean egindako lanak besterik aurkitzen ez badira ere, deskribatutako teknikak nahiz lan desberdinetan ateratako ondorioak interes orokorrekotzat har genitzake.

Informatikaren aroak aldaketa sakona ekarri du lexikografiaren mundura. Informatika dela medio, astindua ere ezagutu du hiztegigintzak azken urteotan: ederki ahantzia dago paperezko fitxekin lan egiten zeneko sasoia. Gaur, ordenagailua da lexikografiaren lanabes arruntena. Eta, hori horrela, ordenagailuz irakurgarriak diren hiztegiak (MRDak) sortu eta lantzen dira gehienbat. Ia ateratzen diren hiztegi guztiak ateratzen dira euskarri elektronikoren batean (CD-ROMean, batez ere).

Ordenagailuen erabilera lexikografia klasikoan lan hauetara zuzendu da: datuen grabazioa, corpusen lanketarako konkordantzia-programak, testuetako informazio-bilketarako sistemak, hiztegientzako laguntza-inguruneak, etab. Ordenagailuz lagunduriko lexikografiaren arlo honetako lanen artean, Collins argitaletxearen COBUILD hiztegia (Sinclair, 87) edo 1984an hasitako *New Oxford English Dictionary* (OED) proiektuaren inguruko hainbat lan aipa daitezke (Simpson, 85; Weiner, 89).

Bestalde, corpusa oinarri gisa hartzen duten hiztegigintzarako hurbilpenek ez dute tresna konputazional gehiegi izan. Hurbilpen honetako ezaugarri nagusiak honako hauek dira: informazioaren eskurapena corpusetik egingo da; errepresentazio eta modelizazio formala, eta informazio lexikalaren erabilpena gizakiari zein LNPko helburuei begira egingo da. (Heid , 1994)-en jasotzen den legez, COBUILD-ekoak izan dira arlo honetan gehien lan egin dutenak, adibidez HECTOR proiektua garatu zuten. Ildo beretik, eta aipaturiko hutsune hori betetzeko DELIS proiektua garatu zela aipatzen du Heid-ek. Proiektu horretan corpus-azterketarako eta lexikoi-eraikuntzarako tresnak diseinatu, inplementatu eta integratu ziren ingurune berean.

Corpusa eta hiztegigintza lotzeko tresnen garapenaz aparte, hiztegien mantenimendurako zein sorkuntzarako, datu-baseek ere oso paper garrantzitsua dute hiztegigintzan. Esate baterako, 1992tik 1995era *Van Dale Lexicographic Information System* (VLIS) proiektua

garatu zuen Van Dale argitaletxeak datu-base eleanitza eraikiz hiztegi elebakar zein eleanitzen produkzioa eta mantenimendua ahalbideratzeko.

CD-ROMean argitaratutakoek, hiztegi datu-baseak izanik, kontsulta-aukera zabalak eskaintzen dizkiote jendeari. Gai berezituetako terminologi bankuak ere oso tresna baliagarriak dira; horien artean aipa ditzakegu TERMIUM (Kanadako terminologi bankua), UZEIk prestatutako EuskalTerm, etab.

Badira, harantzago joanez, hitza ideiatik abiatuz zein alderantziz bilatzeko aukera ematen dutenak ere. Esate baterako, MEMODATAK (Caen, Frantzia) salgai jarri zuen duela urte batzuk DICOLOGIQUE izeneko produktua hiztegi interaktibo bezala.

Bestalde, lexikoiak LNPn gero eta garrantzi handiagoa du. Ikertzailea (arlo desberdinetarikoak: linguistika, linguistika konputazionala, adimen artifiziala, psikolinguistika, etab.), hizkuntzen industria zein instituzioen interesa lexikora lerratuko da. Horrela bada, lexikoaren inguruan hainbat ekintza sustatuko dira: mintegiak, kongresuak, argitalpenak, lantalde berezituak, etab. (Walker, 1989)-n zerrenda osoa dugu. Ez da harritzekoa, beraz, lexikoi horien eraikuntzarena izatea arlo honetako gairik harrotuenetarikoa.

Beste batzuen artean, Europako Elkarteak martxan jarririko programa aipa dezakegu, *Framework Research Program* (1987tik 1991ra garatzen da) bi helburu nagusi zituela:

- LNPrako ezagutza lexikala errepresentatzeko datu-base lexikalen eraikuntza, beste batzuen artean honako zentroetan ari ziren: Pisako Unibertsitatea (Calzolari eta Zampolli ) eta *I.B.M. T.J. Watson Research Center* (Byrd eta beste, 1987).

- Dauden baliabide lexikalak berrerabiltzea, hiztegietan dagoen ezagutza arakatuz eta batez ere euskarri magnetikoan daudenez baliatuz, LNPko sistemetako ezagutza lexikala osatu. Robert Amsler-ek uste zuen lexikografia konputazionalaren hirugarren aroa hasten hasia zegoela, eta aro berri honetan berrerabilgarritasuna oinarrizko ezaugarria izango zela (Amsler, 1989).

LNPrako lexiko-sistemak eraikitzeari gagozkiola, aurrerago ikusi dugunez LNPko aplikazioen problemarik larriena lexikoi konputazional "handia" eraikitzean datza. Lexikoak ere gero eta garrantzi handiagoa hartuko du, eta hainbat ikertzailek aitortuko dio lexikoari LNPko lanetarako lehentasuna, hiztegiak osagairik behinenak bihurtuko direla. Alderdi teoriko, hiztegia hizkuntzaren egitura semantikoaren ikerketarako bide den neurrian, zein praktikotik, hiztegiak adimen artifizialaren arazo larriena den jakintzaren eskuratzean izan dezakeen baliagarritasunagatik.

Hori guztia kontuan izanik jabetu gaitezke egun lexikografia konputazionalak hartu duen indarraz, adimen artifizialeko, eta zehazkiago, ezagutzaren errepresentaziorako lengoaien

ekarpenez ere laguntza jaso duela, tresneria eta ingurune egokien aldetik batik bat. Lexikoiek testuak interpretatu eta desanbiguatu ahal izateko adina informazio beharko dute, hizkuntzaren ulermena erdiesteko ahalegin horretan aurrera egin nahi bada.

Lexikoi konputazionalek hiztegi arruntek baino askoz ere informazio linguistiko esplizitu gehiago behar dute, esate baterako, lexikoi konputazional bateko sarrerak gutxienez informazio mota hauetaz hornitu beharko dira: morfologikoa, sintaktikoa (adib. azpikategorizazioa) eta semantikoa (adib. hautapen-murriztapenak), beti ere LNPko sistema batean integratzeko moduan antolaturik. Beraz bi puntu nagusi azpimarratuko genituzke :

- Prozesu automatiko edo erdiautomatikoen beharra, ezagutza lexikalaren eskurapenak lan eskerga eta kostu handikoa baitakar eskuz eginez gero. Horrela koherentzia eta kontsistentzia ziurrago izanen dira.
- LNPrako behar diren lexikoiek gero eta errepresentazio-eredu zailagoak eskatzen dituztela.

## II.2.2 Informazio lexikalaren eskurapena.

Lexikoi konputazionalak hutsetik edota eskuz eraikitzea zailtasun handikoa izanik, datu lexikalen eskurapenaren arazoari begira "berrerabilgarritasuna" hitz gakoa da. Baliabide lexikalei buruzko eztabaidek eta honen inguruan garatzen ari diren proiektu gehienek berrerabilgarritasuna dute amesturiko jomuga nagusia. Calzolarik dioenari (Calzolari, 89) jarraituz, bi eratara uler daiteke berrerabilgarritasuna:

- Batean ideia da, jada dauden eta eskura daitezkeen baliabide lexikalak berriro erabil daitezkeela, jatorriz pentsaturik ez zeuden helburu edo aplikazioetarako.
- Etorkizunean aplikazio desberdinetarako baliagarriak izango diren baliabideak sortu beharra.

Berrerabilgarritasunari arreta jartzeko arrazoi asko daude: ekonomikoak, estrategikoak, teknikoak, linguistika teorikoarenak zein konputazionalarenak. Halaxe dio behintzat EUROTRA-7 ikerketako koordinatzaileak (Heid, 91). Aipatu ikerketan aplikazio konputazionaletarako baliabide lexikal eta terminologikoen berrerabilgarritasuna izango dute aztergai hamaika instituziok (industria eta mundu akademikokoak).

Iturri lexikalen berrerabilgarritasunaz hitz egiterakoan, euskarri magnetikoan dauden hiztegiez (MRDez) arituko gara. Batez ere, euskarri magnetikoan dauden hiztegiak erabili izan baitira oinarrizko informazio-iturri gisa LNPrako lexikoiak hornitzeko zeregin horretan (Boguraev eta Briscoe, 89; Castellón, 93), hauetan gorderik dagoen tradizio lexikografikoaz aprobetxatu nahirik. LNPko ikertalde askok jardun du MRDez baliatzen joan deneko azken

hamarkadan. Eta lehenengo kapituluan aipatu dugun bezala, —§ I.1— bide horretatik jo dugu guk ere, hiztegiaren edukia aztertuz, maila lexikaleko informazio sintaktikoa erauzte aldera.

Hala ere, MRDetatik informazioa eskuratzea ez da nolanahiko lana, teknika eta metodologia berrien garapena eskatzen baitu (LNP zein informatikaren aldetik). Euskarri magnetikoan gorderik egon arren, ez da ahaztu behar giza erabiltzaileari begira eginiko hiztegiak direla, eta paperean inprimaturik daudenen arazo berberak dituztela. Arazo horiek informazioaren parte handi bat formalizatu gabe izatetik datoz. Horregatik, giza erabiltzailea bere ezagumendu linguistikoaz eta inteligentziaz baliatu beharko da MRD nahiz hiztegi arrunt gehienetako informazioaz jabetzeko.

Bestalde, formalizazio-ezaz gain (batez ere LNPrako egokia ez den heinean), MRDan aurki daitezkeen zenbait bertsiok hainbat inkontsistentzia eta akats dituzte, hauetan informatika tresna lagungarri soila izan baita. Hau dela eta, tratamendu konputazionala bideratzeko aurreprozesua eskatuko dute. Hau guztia kontuan izanik, proiektu bati ekin aurretik iturrien azterketa enpirikoa egin beharko da. Horrela diote (Boguraev eta Briscoe, 1989:35)-n:

> "It´s clear that very careful empirical analysis of a dictionary source must be carried out prior to any serious project..."

Azterketa horretan, bereziki bi ezaugarri hauetaz arduratu beharko gara:

- MRDetan aurki daitekeen informazioak LNPrako duen aplikagarritasuna.
- Informazioa nola dagoen egituratua eta antolatua, eta, jakina, nola eskuratuko den MRDtik.

Dena den, arazoak arazo, eta batzuk corpusak aztertzearen aldekoak izan arren (beste batzuen artean, Zernik 1991, Grishman eta Sterling, 1992), MRDak hartu izan dira nagusiki iturri lexikal aberatsentzat, halaxe diote behintzat Donald Walker-ek eta Antonio Zampolli-k *Computational Lexicography for Natural Language Processing* liburuaren sarreran (Boguraev eta Briscoe, 1989:xiv):

> "The various kinds of existing dictionaries, and in particular the dictionaries available in machine-readable form, are obviously the richest and most valuable sources, based as they are on a long lexicographical tradition which encompass a treasure store of data, information and knowledge."

MRDetako edukia automatikoki analizatu eta arakatzeko hainbat metodologia eta tresna garatu izan dira. Horrela lexikografia konputazionalaren alorrean badago hiztegien analisirako parser orokorrak garatzeko xedea ere (ikus Neff eta Boguraev, 1989, 1991). MRDak erabiliz LNPko sistemetarako osagai lexikalak eraikitzeko teknikak eta metodologiak garatzea helburu

dutenen artean Europako Elkarteko ACQUILEX[1] (Esprit BRA-3030: *Acquisition of Lexical Knowledge for Natural Language Processing Systems*, (Calzolari, 90b)) azpimarratuko genuke.

Nahiz eta informazio-eskuratze hori ez den behin ere osoa izango, eta ezta ere hutsik gabekoa edo erabat automatikoa. Jakina, MRDek soilik ezin diote LNPko behar guztiei erantzun. Gaur egunean, MRDez gain corpusa bilakatu da LNPko sistemetarako lexikoa eskuratzeko iturburu nagusietarikoa. Testu-masa handietarik informazio lexikala eskuratu nahian hainbat proiektu garatzen ari dira. Esate baterako, EDR proiektu japoniarrekoek, itzulpen automatikorako hiztegi elektronikoak landu dituzte, eta hauek eraikitzeko eta aberasteko corpusa hartu dute iturri nagusitzat.

Esan gabe doa lexikografoen lana ere ezinbestekoa dugula lexikoiak osatzeko, MRDen gabeziak eta corpus-lanketarako tresnen zailtasunak kontuan izaki. Hauen lana errazteko ingurune lexikografikoen garapena ere oso inportantea dugu, gaur egun lexikografoek laguntza informatiko handia eskura dezakete: KWIC delako konkordantzi programak, informazio-eskurapenerako sistemak linean, *lexicographer's workbench* delakoa (Lenders, 90) , etab. dira, besteak beste, zenbait kasu.

Aipaturiko hiru iturriok, hala nola, MRD, corpusa eta lexikografoen lana elkarren osagarri dira LNPko baliabide lexikalen premiari erantzuteko.

Hala ere, izan dira ikerkuntzarako beste joera batzuk ere lexikoi konputazionalak eraikitzeko. Hona hemen zenbait: ezaugarri-egiturak, hautapen-murriztapenak, eta azpikategorizazioa esplizitu egiten dituztenak sintaxian oinarrituz (Gross, 1975); hitzei buruzko mota guztietako informazioa modu formalizatuan jasotzen duten hiztegiak (Mel'cuk eta beste, 1981), lexikoi konputazionalen diseinua LNPko sistema desberdinetarako (Flickinger eta beste, 1985).

Bukatzeko, euskarazko iturriei gagozkiola, hor ditugu I. Sarasolaren Euskal Hiztegia, eta Elhuyarrek, Harluxet Fundazioak eta Adorez taldeak, besteak beste, kaleratutako hiztegi-lanak euskarri elektroniko desberdinetan.

## II.2.3 Informazio lexikalaren errepresentazioa.

Datu lexikalen eskurapenaren arazoari begira "berrerabilgarritasuna" hitz gakoa bada ere, gaur egun esan genezake interesak errepresentazioaren aldera lerratu direla. Lexikoaren premia duten guztien lan bateratu eta koordinatu baten beharra nabari da. Horrela bada, estandarizazioa

[1] ACQUILEX: Esprit BRA 3030, batez ere, MRDetan oinarrituko da. ACQUILEX II: Esprit BRA 7315, bigarren honetan arreta handiagoa jarriko dute corpusean. Hala ere, biek zuten helburu nagusi bera: jada existitzen diren baliabideen berrerabilgarritasuna aztertu LNPrako sistemak eraikitzeko.

da egin beharreko ezinbesteko urratsa. Proposamenak hasiak ziren plazaratzen 1990. urtean (Sperberg-McQueen eta Burnard, 1990), hiztegien errepresentaziorako eskemek orokorrak eta aplikazioetarik independente izan behar dute. Behar horri erantzuteko asmoz, elkarlanerako bideak aurkitu eta informazio-trukea bermatuko duten ekimenak sustatuko dira; beste batzuen artean, honako hauek aipatuko genituzke: *Text Encoding Initiative (TEI)*[1] (Sperberg-Mcqueen eta Burnard, 1994), *The ACL Data Collection Initiative, Consortium for Lexical Research.*

Lexikografia konputazionalaren alorrean lexiko-sistemen azterketa, errepresentazioa eta erabilpena, gero eta garrantzi handiagoa hartzen ari dira. Azken hamarkadan lexikoigintzan aurrera egin da: erredundantziaren arazoa konponduz, datuen kontrola eta kontsistentzia gauzatuz, eta informazio-atzipena erraztuz.

Orainokoan nabarmendu dugun "benetako" lexikoien premia horrezaz gain, landurikoetan ez zegoen adostasunik ez lexikoiek jaso behar zuten informazioaz ez hau nola errepresentatu behar zen (Ingria, 1986). Gai honen haritik, informazio eskuratze-prozesuak ez ezik errepresentazioaren arazoak ernatzen ditu batik bat ikertzaileen interesak. Kontuan izan behar da —II.1 atalean— aipaturiko ikusmolde aldaketa horrek gramatikak erraztea ekarri duela, baina informazioa lexikoan pilatzeak sarrera lexikalak informazio erredundanteaz hornitzea ekar lezake. Informazioaren kopuruak eta konplexutasunak informazioa bera kontrolatzeko arazoak sor ditzake. Beraz, beharrezkoa izango da sarrera lexikalek zein motako informazioa behar duten erabakitzeaz gain, nola egituratu informazio hori guztia erredundantzia ekiditeko eta portaera bereko hitz-moten arteko pareko ezaugarriak antzemateko. Arazo hauei erantzuteko ezagutza-base lexikalak (*Lexical Knowledge Bases*, LKB) garatuko dituzte zenbait proiektutan, adibidez ACQUILEX-en. Ezagutza-base lexikalek herentzia-mekanismoak eta erregela lexikalak baliatuz informazio lexikalaren erredundantzia ekiditea eta kontsistentzia bermatzea lortzen dute. Honetaz gain, informazio lexikal egituratua errepresentatzeko orduan ahalik eta zehatzen izateko eta orokor diren tasunak jasotzeko, batez ere herentzia, balio lehenetsien espezifikazioa eta erregela lexikalak aipatu izan dira. Tresna hauen azpian dagoen ideia da hitz-moten hierarkia eta herentziaren nozioa. Hau da, hitz-mota bereko elementuek ezaugarri berak konpartituko dituzte. Esate baterako, erregela lexikalen zeregina izango litzateke bi hitz-motako elementuen arteko harremanak sistematikoki errepresentatzea (Flickinger, 1987). Ildo beretik, semantika lexikalaren ikuspegitik, item lexikalak errepresentatzeko *Qualia Structure* teoria garatzen du Pustejovsky-k (Pustejovsky, 1991). Teoria honen bidez hitzek dakarten polisemia sistematikoki adieraziko da lexikoian behar ez den anbiguotasun lexikala ekidinez. Horrez gain, autore honek dio, egitura lexikal bakanak ezagutza-base lexikal zabalago batean integra daitezkeela herentzia lexikalaren teoriari esker.

---

[1] TEIren helburua, dena den, ez da lexikoaren arlora mugatzen. Ekimen horretako gidalerroen helburua giza zientzietako ikerkuntzan datu-trukerako eta testuen kodeketarako formatu estandarra eskaintzea da.

Teoria honek lexikoiaren antolamendu orokorrerako behar diren printzipioak ditu, eta gure hizkuntza naturalaren lexikoia osotasun kontzeptual batean integratzen laguntzen digu.

Bestalde, hiztegi-informazio lexikalaren egitura konplexua ezin egoki errepresentatuarena nabarmentzen zaigu. Datu-eredu "konbentzionalak" desegokiak dira datu-base lexikaletarako. Adibidez, erlazionalean informazio lexikalaren egitura konplexua ezin da egoki errepresentatu. Egokiagoak bide dira datu lexikalak errepresentatzeko ezaugarri-egituretan oinarritutakoak, besteak beste arrazoi hauengatik:

- informazioa atzitzeko eta maneiatzeko bide anitz.
- hiztegi jakin baten antolaketa gordetzen ahal da, kontsultarako "transparente" eginez.
- oinarri teoriko sendoa.
- lexikoi konputazionalekiko bateragarritasuna.

Ezaugarri-egituretan oinarritutako errepresentazio-eredua inplementatzeko modurik egokientzat-edo, objektuei zuzenduriko datu-basea daukate (Ide, N. eta beste, 1993).

Datu-base lexikal (DBL) idealaren ezaugarriak hauek behar lukete izan:

- erabiltzaile eta datu-basearen arteko elkarrekintza oso garrantzitsua da. Hori gauzatzeko, interfaze lagunkoiak izatea komeniko litzateke.
- malgutasuna, hau da, edozein momentutan helburu berrietarako egokitzen erraza, informazio mota berriak aise onartuko dituena.
- berrerabilgarritasuna, hots, informazio lexikala berrerabilgarria izatea.
- dimentsioaniztasuna, hau da, helburu askotarakoa. LNPren arloko aplikazioetarako zein lexikoa beharrezkoa den beste aplikazio batzuetarako ere baliagarria. Horretarako informazio anitza beharko du izan: morfologikoa, sintaktikoa, semantikoa, pragmatikoa, etab.
- neutraltasuna (eskola linguistiko desberdinekikoa), hau da, bertan egindako deskribapen linguistikoak ez lituzke baldintzatu behar etorkizuneko aplikazioak.

Garbi dago, beraz, datu lexikalak, hiztegietakoak esaterako, oso datu konplexuak direla eredu konbentzionala jarraitzen duten datu-baseen bidez errepresentatzeko. Horregatik, aurrerago esan bezala zenbait autore ezaugarri-egituretan oinarritzen diren ereduen aldekoak dira. Horrez gain, nabarmena da zein garrantzitsua den datu lexikalak datu-base lexikal batean gordetzea; besteak beste, datu-baseek eskaintzen dituzten aukerei esker hiztegien eguneratzea

eta mantenimendua, hiztegien berstio desberdinen sorkuntza, datuen kontsitentzia bermatzea, etab. oso modu ziurrean egin daitezkeelako.

## II.3 Baliabide lexikalen beharra Lengoaia Naturalaren Prozesamenduan.

1986ko Grossetoko mintegia mugarri garrantzitsua dugu baliabide lexikalekiko kontzientziak esnatzea eta hauen garrantziaz jabetzea lortu baitzuen. Mintegiaren bukaeran *Manifiesto* dokumentua osatu zen. Dokumentu horretan baliabide lexikalen (corpus eta lexikoien) garrantzia azpimarratu zen, eta zenbait ekintza gomendatu ziren. Ondorengo urteetan alor honetan Europan garatutako hainbat ekimenen oinarriak ezarri ziren. Batez ere, baliabide lexikal berrerabilgarriak lortzeko aktibitate handia piztu zuen.

Egun eremu oso aberatsa, konplexua eta azkar aldatzen dena bilakatu da baliabide lexikalen alorra. Antolakuntza eta teknika mailan garatu egin da. Eta ikerkuntza mailan ere, teknologia berrien, metodologien eta tresnen beharra eskatzen du. Bestalde, ezin dugu ahaztu ezinbesteko oinarriak direla LNPrako eta honen aplikazioen garapenerako, zein hizkuntzen industrien etorkizunerako. Antonio Zampollik horrela zioen *First International Conference on Language Resources and Evaluation* (LREC) kongresuko komunikazioak jasotzen dituen liburukiaren sarreran (1998:XVI):

> "The choice of the term "Resources", coined rather recently, was intended to capture the idea that large collections of language data and descriptions play, for development of effective NLP systems and their applications, an essential infrastructural role comparable to the role that basic resources such as highways, railways, electrical networks and energy play for the industrial and economical development of a country"

Datu linguistikoak baliabide preziatuak dira ezagumenduaren gizartean. Baina, baliabide hitzak normalean beste testuinguru batzuetan izan du bere lekua, baliabide naturalak, ur-baliabideak, baliabide ekonomikoak, etab. Baliabide lexikal kontzeptua linguistika konputazionalari zor zaiola esan genezake. Hona hemen zer dioen Swanepoel-ek Gellerstam-en artikuluan (Gellerstam, 1995:58):

> "The computer systems and tools that are becoming available both to the researcher, the practical lexicographer and the human user are opening up a myriad of possibilities for the presentation and utilization of masses of lexical information."

Hala ere, kontua ez da soil-soilik zientzietako adituek lexikografoen eskuetan jartzea tresnak. Datu lexikalak lexikografo adituek bildu eta sistematizatu behar dituzte, zeren eta datu lexikalik gabe linguistika konputazionalak ez du izango zer berrerabili. Eta lexikoia *kondizio sine qua non* da linguistika konputazionalarentzat.

Calzolari-k *An Overview of Written Language Resources in Europe: a few Reflections, Facts, an a vision*, (Calzolari, 98) artikuluan baliabide lexikalek Europan duten egoeraren ikuspegia eta hizkuntzen ingeniaritzaren arloan duten oinarrizko papera azpimarratzen du. Horrez gain 1980ko hamarkadaren bigarren aldian eta 1990ko hamarkadaren hasieran lexikoien eta baita ere corpusen inguruan Europan garaturiko hogei proiektutik gorako zerrenda dakar. Autorearen iritziz, proiektu hauek ekarpenak egin dituzte, baina ez dute aurreikusitako estrategiarik edota plangintza garbirik. Arlo honetan diharduten guztien (herrialde desberdinetakoak, proiektu publiko zein pribatuetakoak) elkarlana bultzatu beharra dagoela esaten digu. Askotan, proiektu berri bati abiada ematerakoan berregiten baitira baliabide lexikalak, eskuragarri dagoena berrerabili gabe. Eta are okerrago, proiektuetako baliabide lexikalak ahantzita edota erabili gabe uzten dira maiz.

Egoera honi aurre egiteko, 90. hamarkadaren lehen erdian Europako Elkarteko batzorde batek hiru baldintza aipatzen ditu baliabide lexikalen oinarrizko paper hori sendotzeko:

- baliabide lexikalen eraikuntza zabalkiro onarturiko estandarretan egin beharra.
- Europako Elkarteko hizkuntza guztietarako baliagarri izango diren baliabide lexikal oinarrizkoen eraikuntza. Adosturiko diseinu batez eraikiz.
- sorturiko baliabide lexikalak komunitateak eskuragarri izan ditzan, distribuziorako politika baten beharra.

Egun Europan arlo honetan dauden proiektu garrantzitsuenak, hain zuzen ere, hiru alderdi horiek lantzea dute helburu batez ere. Proiektu hauek guztiak LE Programme (*Language Engineering Programme*)-ren barruan kokatzen dira. Hona hemen proiektu horiek zein diren:

- EAGLES (*Expert Advisory Group on Language Engineering Standards*), helburu nagusi bezala estandarren garapena arlo hauetarako:
    - baliabide lexikal zabaletarako. Adibidez, corpusak, lexikoi konputazionalak, etab.
    - gomendioak ezagutza lexikal hori erabiltzeko, linguistika konputazionalaren formalismoen bidez, markatze-lengoaien eta software-tresna desberdinen bidez.
    - baliabide, tresna eta produktuen ebaluaketarako eta gomendioetarako.
- PAROLE (*Preparatory Action for Linguistic Resources Organisation for language Engineering*). Proiektu honen helburua da hasierako corpus eta lexikoiak sortzea Europako Elkarteko hizkuntza guztietarako. Honen ondorik, SIMPLE proiektua aipatzen da PAROLE-ren jarraipen bezala, eta eginkizun gisa jada existitzen diren geruza morfologiko eta sintaktikoei semantika gehitzea izango du. Bestalde, Europako

Elkarteko hizkuntza guztietarako baliabide lexikalak sortzeko zeregin horretan, EuroWordnet nabarmentzen da. Bere xedea da datu-base eleanitza sortzea hitzen arteko oinarrizko erlazio semantikoak jasoaz.

PAROLE-ren pareko beste proiektu bat TELRI[1] dugu. TELRIk (*Trans-European Language Resources Infrastructure*), hizkuntza eta hizkuntzaren teknologia kontuetan puntan dabiltzan zentroen arteko azpiegitura sortu eta industria, ikertoki eta unibertsitateetako LNP komunitateari hizkuntz baliabide elebakar eta eleanitzak eskaintzea du helburuetako bat. Baliabide horien artean, corpusak, hiztegiak eta lexikoiak euskarri elektronikoan, datu-base lexikalak, eta hizkuntz datuak sortu, berrerabili, mantendu eta ustiatzeko software-tresnak aipatzen dituzte.

- ELRA (*European Language Resources Association*), baliabide lexikalak gordetzeko, hauen erabilpena bultzatzeko eta eraginkorki banatzeko lanak bultzatuko ditu. Honen parekoa Estatu Batuetan LDC (*Linguistic Data Consortium*) dugu.

ELRA-k antolatu zuen 1998an LREC (*First International Conference on Language Resources & Evaluation*) kongresua. Eta kongresu hau antolatzen eta bultzatzen parte hartu zuten elkarte (mota askotarikoak), eta errepresentazioa izan zuten herrialde eta hizkuntza anitzek erakusten dute Grosseton ernetako baliabide lexikalekiko ardurak hazkunde oparoa izan duela eta etorkizunari begira ere oinarrizko arloa dela.

Arlo honen inguruko informazio interesgarri bezain zabala jasotzen dute ELRA-k antolaturiko kongresuen berri jasotzen duten liburukiek.

## II.3.1 Datu lexikalak baliabide linguistiko gisa.

Baliabide lexikal terminoak (*language resources*, LR) erreferentzia egiten die datu lexikal multzoei (usu handiak) eta deskribapenei, MRD euskarrian erabiliak izango direnak LNPko sistemak hobetzeko edo ebaluatzeko. Baliabideen artean ditugu: idatzizko zein ahozko corpusak, datu-base lexikalak, gramatikak, terminologia, etab.. Hauen egituratzeari erreparatuz gero, bi multzotan bana ditzakegu:

- baliabide lexikal egituratuak: esate baterako, giza erabiltzaileari zuzenduriko hiztegiak, thesaurusak edota entziklopediak

- baliabide lexikal ezegituratuak: corpusa.

Hauetaz gain, baliabideen artean sar daitezke baita ere, baliabideen eskurapen, prestaketa, bilketa, kudeaketa eta erabilpenerako oinarrizko software-tresnak (Zampolli, 1998).

---

[1] http://www.ids-mannheim.de/telri/telri.html

Baliabide gisa hartzen da baita ere, introspekzioa, esate baterako, LNPko sistema baterako lexikoia eraikitzen ari den gizakiak hizkuntza eta munduari buruz duen ezagutza.

## II.3.2 Datu lexikalen gordailuak.

Baliabide lexikalez aritzerakoan, datu lexikal gisa definitzen dira bereziki euskarri magnetikoan daudenak, eta ikerkuntza lexikalean erabil daitezkeenak edota produktu komertzialen (hiztegiak, itzulpengintzarako laguntza automatikoak, zuzentzaile ortografikoak, etab.) oinarri direnak. Datu lexikalak lengoaiaren erabiltzaileari modu askotara hel dakizkioke, euskarri magnetikoan dauden modu askotariko hiztegiak eta konputazio-programak. Ondorengo puntuetan, aipatzen ditugun gordailuen ordenak hierarkia erakusten du, datu lexikalen gordailu sinpleenetatik hasi eta konplexuagoenetara:

1. **Hitzen maiztasun-zerrendak.**

   60 eta 70eko hamarkadetan garatu ziren. Adibidez, *American Brown Corpus* (Kucera & Francis 1967), ondorengo corpus ikerkuntzan batez ere corpus bildumak egiteko modelo gisa hartu dena.

2. **Inprimaturiko hiztegiak, Machine Readable Dictionary (MRD) bertsioa dutenak.**

   Euskarri magnetikoan dauden giza erabiltzaileari zuzenduriko hiztegiak ditugu batik bat. 60ko hamarkadan koka genezake lehena *Merriam Webster Seventh New Collegiate Dictionary* MRDn jarri zutenean. Hauetariko batzuek kode linguistikoak eta sailkapenak izango dituzte, adibidez *Longman Dictionary of Contemporary English* (LDOCE) hiztegiak.

3. **Makina-Lexikoiak (Machine Tractable Dictionaries, MTD).**

   MRDen antzekoak ditugu; dagoen aldea da LNPko programek errazago erabiltzeko prestatuak daudela.

4. **Datu-base lexikalak (Lexical Data-Bases, LDB).**

   Deskribapen maila desberdinetan informazio formalizatua dute. Informazio lexikal ugari prozesatzeko tresnarik nagusienetarikoa dugu gaur egun. Aplikazio linguistiko anitzetarako zein linguistikaren alorreko ikerkuntza orokorrerako erabil daitezke. Datu-baseen kudeaketarako sistemek erabiltzaileari eskaintzen diote aukera anitz datuak atzitzeko, sistemaren barne-funtzionamenduaz jakin beharrik gabe; eskura nahi dituen datu linguistikoei buruz jakitea nahikoa da.

**5. Ezagutza-base lexikalak (Lexical Knowledge Bases, LKB).**

Ezagutza-base lexikaletan informazio lexikala batez ere ikuspegi semantikotik antolatzen da. Ezaugarri garrantzitsuena heredentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen baitira (Copestake, 1990).

## II.4 Baliabide lexikalak estandarizatzeko premia.

Aurreko puntuan —§ II.3.ean— baliabide lexikalen beharra LNPn azpimarratzen genuen, hauek duten garrantzia, eta batez ere berrerabilgarriak izatearen premia (jada existitzen direnek zein etorkizunean sor daitezkeenek berrerabilgarriak izan behar dute). Datu linguistikoen berrerabilgarritasuna ziurtatzeko, edozein testuk, konputagailuen laguntzaz aztertuko bada, konputagailuak irakurri ahal izango duen moduan kodeturik egon behar du. Kodeketa estandar baten ezean, eta kodeketa diogunean software eta hardware bateragarriaz ere ari gara, ezinbestean, testuak ikergai dituzten hainbat arlotako ikertzaileek testu horien tratamendua erraztearren hamaika era desberdin asmatu eta erabili izan dituzte. Nork berea —eta bere modura— egin duela, ordea, azkenean batek egindakoaz beste bat baliatu nahi izan denean aurretik egindako lan guztia ez da suertatu izan nahi litzatekeen bezain lagungarri; maiz, erabat erabili ezina ere bai. Egoera tamalgarri horren aurrean, saio bat baino gehiago jo izan da azken hogei urte honetan testuen kodeketarako estandar baten bila, fruituak heldu ez badira ere.

Bestalde, estandarizazioaren beharra ez da unibertsalki konpartitua, honen lekuko TEI salbu gainontzeko estandarizaziorako ekimenak Europan kokatzen direla. Estatu Batuetako linguistika konputazionaleko ikertzaileek uko egin diote estandarren ezarpenari, eta beraien arabera estandarrak praktikatik sortuko dira (behetik gora, *bottom-up*). TEI ekimenekoek sorturiko gidalerroek (maiz TEI P3[1] modura ezagutuak) esfortzu aitzindaria errepresentatzen dute ordura arte ekimen isolatu eta noizbehinkakoak izan ziren arloan. Eta oinarrizko lana dugu formatu elektronikoan dauden testuen kodeketarako etorkizunean. Testu-datuen errepresentazio-arazoak formatu elektronikoan planteatzen dituzte.

1980ko hamarkadako bukaera aldean, oso testu-bilduma egoki gutxi zeuden linguistika konputazionaleko ikerkuntzarako, bereziki ingelesa ez den hizkuntzetarako. Beraz, testu-bilduma erraldoiak (elebakar zein eleanitz) biltzeko eta zabaltzeko hainbat ekimen sortu ziren, horien artean: *ACL Data Collection Initiative* (ACL/DCI), *European Corpus Inititative* (ECI), Estatu Batuetako *Linguistic Data Consortium* (LDC), *MULTEXT* Europan, etab. Ekimen hauek guztiak hasiera baino ez ziren behar zen ahaleginerako, eta baliabide testual handiak egoki eratzeko oraindik lan ikaragarria zegoen egiteko.

---

[1] Guidelines for Electronic Text Encoding and Interchange, TEI P3. Amarauneko helbide honetan aurki daiteke gidalerro hauei buruzko informazio zabalagoa: http://www-tei.uic.edu/orgs/tei/p3/elect.html.

Horrez gain, testu-bildumetako berrerabilgarritasunaren eskaerari erantzuteko testu-datuen kodeketarako estandarra garatu beharra zegoen. Datuak formatu *ad-hoc*-etan zabaltzen jarraitzea errealitateari uko egitea zen, kontuan izanik erabilera partikularretan zer-nolako ahaleginak eta baliabideak behar ziren datuak txukuntzeko eta berrantolatzeko, kasurik hoberenetan kostu handikoak eta askotan ezinezkoak.

Existitzen ziren eta baliagarri izan zitezkeen datu-bilduma gehienak inprimatzeko helburuarekin zeuden formateatuak, eta hau dela eta, kodeketan esplizituki errepresentatutako informazioa gehiago dago lotuta testuaren formatu fisikoari beronen egitura logikoari baino (interesgarriagoa dena LNPko aplikazioetarako); eta gainera maiz bien arteko harremana gauzatzea oso zaila edo ezinezkoa izango da lan ikaragarria ez baldin bada egiten.

Horrezaz gain, gero eta datu-bilduma gehiago daudenez eskuragarri, eta datu-bilduma handien erabilpena ezinbestekoa denez LNPko ikerkuntzan, orduan, software orokorra eta publikoa testuen lanketarako garatzen ari da berrerabilgarri izateko, eta horretarako kodeketa-formatu estandarra behar da.

Hona hemen Nancy Ide-ren arabera LNPko ikerkuntzarako testuak errepresentatzeko kodeketa-formatu estandar batek bete behar dituen ezaugarriak:

- kapaz izan behar da LNPko ikertzaileen komunitatearentzat interesgarri izan daitezkeen testu-mota eta hizkuntza anitzetan aurki daitezkeen informazio-mota desberdinak errepresentatzeko, bai prosan, dokumentu teknikoetan, egunkarietan, poesia, antzerkia, eskutitzak, hiztegiak, lexikoiak, etab.;
- informazio-maila diferenteak errepresentatu ahal izango ditu, ez bakarrik ezaugarri fisikoak eta egitura logikoa (baita fenomeno konplexuagoak ere, hala nola, testu barnean edo testuen arteko erreferentziak, elementu paraleloen alineamendua, etab.), baita ere, datuei gehi dakizkiekeen anotazio interpretatibo edo analitikoak (esate baterako, kategoria gramatikalen anotazioa, egitura sintaktikoa, etab.) ;
- aplikazioarekiko independente, hots, malgutasuna eta orokortasuna izan behar ditu, aldi berean testu bereko informazio-mota desberdinak esplizituki kodetzeko eta prozesamendu-mota desberdinetara egokitzeko.

Honelako kodeketa-sistema malgu eta moldagarria garatzea lan intelektuala da nagusiki, zeinak eskatzen duen testu-mota desberdinak jasotzeko modelo konplexuen garapena, testu-modelo orokor bat, eta hori gorpuzteko kodeketa-eskemaren arkitektura.

## II.4.1 TEI: testu-kodeketarako ekimena.

Testu-kodeketarako ekimena (*Text Encoding Initiative*, TEI) delakoak testu elektronikoak kodetzeko eta trukatzeko bere gidalerroak —TEI P3 — 1994ko maiatzean eman zituen argitara. Sei urtetan zehar mundu zabaleko hamaika ikertzaile eta ikertalderen lanaren fruitu diren mila eta hirurehun orrialdeko gidalerro horien helburua ondokoa bezain sinple eta, aldi berean, handinahikoa da: ezaugarri desberdineko testu-mota zabal bat era kontsistente eta hobezin batean kodetzeko bideak eskaintzea.

Gidalerroen garapenean, TEIk identifikatu zituen askotariko ikertzaileek zer-nolako kodetze premiak zituzten informazioaren elkartrukeari zegokionean, horretan oinarritu zituen orokor nahi zuen eskema batek bete beharreko kodeketa-printzipioak, eta identifikatu zituen zein ziren kodetze-arauak behar zituzten testu-klase eta -ezaugarriak. TEIk eskaintzen dituenetan arakatzen hasita, hona hemen batzuk:

- SGML (*Standard Generalized Markup Language*) markatze-lengoaia egokitzat jotzea gidalerroen garapenerako oinarri gisa.
- SGML erabiltzeko gomendioak —zenbait murriztapen—, bere orokortasuna eta malgutasunari eutsiz aldi berean.
- Testu-datuak kodetzerakoan beharrezko diren kategoria eta ezaugarrien identifikazioa eta analisia, maila askotan.
- Testu-egitura definizio orokorren multzo malgu eta hedagarria.
- Testu elektronikoak dokumentatzeko metodo bat, biblioteketan erabiltzen den katalogatze-arauekin bateragarria.
- Kodetze-arauak testu-mota eta -ezaugarri desberdinetarako: karaktere-multzoak, hizkuntz corpusak, linguistika orokorra, hiztegiak, datu terminologikoak, ahozko testuak, hipermedia, literatur prosa, olerkia, antzerkia, iturburu historikoak, eta testu-kritikarako aparatua.

Hasiera-hasieratik TEI eskema diseinatu zen hardwarea, softwarea eta aplikazioetatik independente izateko helburuarekin. Aplikazioetatik independente-nahi horrek izugarrizko garrantzia du, gure ustez, eta testu baten ikuspegi desberdinak kodetu ahal izateko aukera ematen digu. Izan ere, testu bat ikus baitaiteke objektu fisikoen bilduma bezala (liburukiak edo paper-orri solteak), edo objektu tipografikoen segida bezala (karaktere-sekuentziak, letra-molde eta marjina-eskema desberdinen arabera antolatuak), edo objektu linguistikoen sekuentzia bezala (grafema edo fonemak, morfemak, unitate lexikalak, sintagmak, ...), edo objektu formalez osatutako egitura bezala (ahapaldiak, lerroak, kapituluak, atalak, ...), eta abar eta abar. TEI gidalerroek helburu orokorreko kodetze-eskema bat definitzen dute, ikuspegi

desberdin horiek guztiak era desberdinetan kodetzeko aukera ematen duena, eta, nahi izanez gero, aldi berean gainera.

Gidalerroek ikerketarako beharrezko diren testu-ezaugarriak errepresentatzeko aukera asko eta asko ematen dituzte, eta oro har aitortzen zaie egun diren premia gehienetarako baliagarritasuna. Bestalde, gidalerroen diseinuak berak beren hedagarritasuna bermatzen du, eta, horretara, estandarrean definituriko elementuekin-eta aski ez duen erabiltzaileak aukera guztiak ditu dokumentu-motak (zeinak DTD, *Document Type Definition* delako fitxategian definitzen diren) bere premietara egokitzeko (gidalerroetan, gainera, adierazten zaio nola egin gauzak nahasten ibili gabe). Software berezirik gabe erabili ahal izateaz den bezainbatean, berriz, esan beharra dago posible dela editore estandarrez-eta baliatuz lan egitea, baina aitortu behar da askoz ere emaitza hobeak eta eraginkortasun handiagoa lortuko dela inondik ere editore bereziak eta, oro har, software berezitua erabiliz gero.

Bestalde, diseinatutako eskemak erantzun bat ematen die jada kodetze-proiektu gehienen oinarrizko premiei; hala ere, gidalerroak zabaldu egin behar dira eta, batik bat, erabiltzaileak behar-beharrezko diren laguntza-tresnez hornitu behar dira. Izan ere, estandar bat "saltzeko" modurik hoberena estandar horren erabilera erraztuko duten software-tresna eta baliabideak garatzea baita.

Erabiltzaile askoren premiak betetzera datoz TEI P3 gidalerro hauek: zientzia eta giza arloko ikertzaileak, argitaratzaileak, bibliotekariak, eta, oro har, dokumentuen bilaketa eta biltegitzearekin zerikusia duten guztiak. Erantzun bat ematen dio, orobat, "hizkuntzaren teknologiaren" arloko jendeari, orotariko testu-corpus eta lexikoen pilatzeari emanak baitaude azken aldi honetan, hizkuntzaren ulerkuntza, sorkuntza eta itzulpenari dagokion ikerkuntzan sartuak.

## II.4.2 SGML: testuak markatzeko lengoaia estandar eta orokorra.

Testuak markatzeko lengoaia estandar eta orokorra, hots, *Standard Generalized Markup Language* (SGML) marka-multzo bat baino marka-multzoak espezifikatzeko metalengoaia bat da. Eta metalengoaia horretaz baliatuz diseinatu dira TEI gidalerroak.

SGML lengoaian, testu barnean markatze-kodeak txertatzen dira eredu bati jarraituz eta modu deskriptiboan egituratzen da. Programazio-lengoaien edo aplikazioen mende sortzen diren egitura berezituak saihesten dira.

Markatze-sistema deskriptiboa da, eta honelako markatze-sistemek markatze-kodeak erabiltzen dituzte dokumentuaren zatiak izendatzeko. Horrela bada, <esaldia> edo </esaldia> bezalako kodeak dokumentu batean txertatzen ditugunean, testu horren zati baten hasiera edo bukaera adieraziko genuke.

SGMLk dokumentu-motaren nozioa ere ezartzen du, hau da, dokumentu bakoitza mota batekoa izango da eta dokumentu-mota batek dokumentu-multzo bat definituko du. Dokumentu-mota hau DTD (*Document Type Definition*) delako fitxategian definitzen da eta dokumentu guztiek DTD bati esleituak egon behar dute.

DTDan dokumentuaren mota formalki definitzen da, hots, bere osagaiak eta egitura esplizituki adierazten dira. Adibidez, txosten motako dokumentu bat definitzerakoan, hasieran egilearen izena etorriko dela esan beharko dugu, ondoren laburpen bat eta, azkenik, txostenaren muina daukaten hainbat paragrafo. Esaten da SGML lengoaia meta-lengoaia dela, DTDen bidez (edo, SGML terminologia erabilita, "aplikazioen" bidez) azpilengoaiak definitzen baitira. Azpilengoaia horien adibide behinena Interneten-eta hain erabilia den HTML (*Hyper Text Markup Language)* dugu.

Adibide gisa, demagun poemak gordetzeko barne-egitura definitzen dugula. *Antologia* bat hainbat *poemaz* osaturik egongo da. *Poemek, izenburu* bana eta hainbat *ahapaldi* izango dituzte. *Estrofak* hainbat *lerroz* osaturik egongo dira.

SGML lengoaian kodetuta, egitura horrekin bat datorren testu batek horrelako itxura izango luke:

```
<antologia>
  <poema>
    <izenburua>Potaren galdatzia</izenburua>
      <ahapaldia>
        <lerroa>Andria, Ieinkoak drugatzula; orai berdi girade</lerroa>
        <lerroa>ni errege balin baninz, erregina zinate;</lerroa>
        <lerroa>pot bat, othoi, egidazu; etzaitzula herabe;</lerroa>
        <lerroa>nik zugatik dudan penek hura merexi dute.</lerroa>
      </ahapaldia>
      <ahapaldia>
        <lerroa>Eia horrat, apart´adi; nor uste duk nizala?</lerroa>
        <lerroa>Horlako bat eztuk uste nik ikusi dudala;</lerroa>
        <lerroa>horrelako hitz gaixtorik niri eztarradala;</lerroa>
        <lerroa>berzer erran albaitzita; enuk uste duiana.</lerroa>
      </ahapaldia>
  </poema>
  <poema>
    <izenburua>...
    ...
  </poema>
    ...
</antologia>
```

## II.4.3 TEI hiztegigintzan.

TEIk trataturiko testu-mota konplexuenetarikoak ditugu hiztegiak. Artikulu bakoitza egituratze-maila handiko objektua da, zeinetan laburtzapen eta egitura gordailuak baliatzen diren informazioaren errepresentaziorako. Areago, hiztegien egitura oso aldakorra da, hiztegi beraren

barnean zein hiztegitik hiztegira. Hain aldakorra da egitura non, informazio-mota oro hiztegiren baten edozein tokitan ager daitekeen.

Hala ere, giza erabiltzailea gai da hiztegiko artikuluak interpretatzeko, askotan sarrerako azalpenetara jo gabe ere. Garbi dago badirela zenbait egiturazko printzipio sendo eta kontsistenteak, kodeketa-eredu batek jaso behar lituzkeenak. TEI *Dictionary Working Group* lantaldearen lehenbiziko eginkizuna kodeketa-eredu orokorra (hiztegi desberdinentzat baliagarria) eta era berean egiturazko printzipio orokor horiek jasoko dituena garatzea zen. Orokortasunaren eta deskribapen-ahalmenaren arteko gatazka testu-mota askotan gertatzen da, baina are gehiago hiztegietan.

Bestalde, hiztegiek duten arazoa da (beste testu-motek ez dutena), testu eta datu-base direla aldi berean. Esan gabe doa, hiztegiak testu diren heinean gainontzeko testu-moten ezaugarriekin bat datozela. Baina, erabiltzaileek ez dute hiztegia *Atik Zra* irakurtzen beste testuekin egin ohi duten bezala, artikulu batera jotzen dute sarrera-buruaz baliatuz (letra lodiz), artikulu horretan sarrerarekin (informazio-eremu) erlazioa duten informazio-eremuak (idazkera, kategoria, azpikategoria, definizioa, etab.) kontsultatuz.

Hiztegi elektronikoek are garbiago erakusten digute hiztegien alderdi hau: erabiltzaileak jaso ditzake hitza agertzen deneko definizioa duten sarrera guztiak, edota *da-du* azpikategoriadun aditz guztiak. Beraz, hiztegiek bikoiztasun sendoa erakusten digute beren azaleko egitura (testua) eta sakoneko egituraren artean (informazioaren edukia). Sakoneko egiturako informazio ugari ez da esplizituki ageri azaleko egituran, baina beharrezkoa da ezagutzea laburtzapenak eta hiztegietako konbentzio formalak.

Hiztegien bikoiztasun honek arazoak sor ditzake kodeketari heltzerakoan, erabiltzaileek hiztegia bi ikuspegi diferenteetatik kodetzea nahi baitezakete. Azaleko egituratik sakoneko egiturako informaziora heltzeko inferentzia-mekanismoak garatzea zaila da oso. Dena den, bi ikuspegi hauetatik kodetzeko aukera eman behar du TEIk, independenteki zein aldi berean. Bi arazo hauei heltzen diete TEIkoek, batetik orokortasunaren beharra eta ahalmen deskriptiboaren arteko gatazka, eta bestetik datu-base ikuspegiaren eta testu gisa ikusten duenaren arteko oposizioa.

Hiztegien kodeketa-eskemaren deskribapen osoa nahi duenak jo beza TEI P3ko hamabigarren kapitulura (Sperberg-Mcqueen and Burnard, 1944), *Print Dictionaries* (pp. 321-70).

Hiztegietako lantaldearen zeregina zen artikuluaren deskribapena, hasi goi-mailako elementuetatik eta behekoetaraino. Lan honi ekiteko hiztegi hauetara mugatu ziren: mendebaleko hizkuntzetako hiztegiak, eta bereziki modernoak, tamainakoak eta egitura eta eduki aldetik aniztasun nahikoa erakusten dutenak.

Hiztegi-sarreren egitura aldakorra da, bai hiztegi berean bai hiztegitik hiztegira. Kodeketa-eskemak egitura horiek guztiak jaso ditzan, elementu orori hiztegi-sarreraren edozein posiziotan azaltzeko aukera emango zaio. Dena den, hiztegi gehienak egiturazko printzipio berdintsuei jarraitzen zaizkie, ideialki etiketatzeko gomendioek isla ditzakete printzipio horiek. TEIko ekimenekoek bi elementu-mota definitzen dituzte hiztegi-sarrerak adierazteko: *<entry>* zeinak hiztegi konbentzionalen ezaugarriak biltzen dituen, eta *<entryFree>* zeinak oinarrian elementu berdinak erabiltzen dituen, baina konbinazio askeagoak baimenduz. Hiztegietako lantaldeak DTD orokor bat definitu zuen hiztegi ororentzako aplikagarria izango zena, eta aldi berean hiztegietako egituraren berri emango zuena. DTD hori garatzerakoan kontuan harturiko puntuak eta aurkituriko arazo nagusiak, hauexek ditugu:

- **Oinarrizko osagaiak.**

Artikuluetan zehar ager daitezkeen zenbait informazio-mota, hala nola, sarrera-buruaren formari dagokiona (ortografia, ebakera, etab.), informazio gramatikala (kategoria, azpikategoria, etab.), definizioak edo itzulpenak helburu-hizkuntzan, etimologia, erlazioak, erabilerari dagokion informazioa, eta adibideak.

Lehenengo urratsa hiztegientzako DTD bat definitzerakoan, elementu atomikoen tipologia eta hauentzako nomenklatura egokia zehaztean datza. Atomikoak dira, artikulu batean deskonposatu ezin diren informazio-eremu garrantzitsuak. Ez dute beste hiztegi osagairik beraien barnean. Hiztegien oinarrizko informazio-eremuei buruzko eztabaida eta azterketa, TEIren aurretikoa da (Danlex, 1987; eta batez ere, Amsler eta Tompa, 1988).

Hiztegietako lantaldekoek artikuluko oinarrizko osagarritzat harturikoak, TEI P3ko hamabigarren kapituluan jasotzen dira (333. orr.).

- **Egitura hierarkikoa eta informazioaren hedadura.**

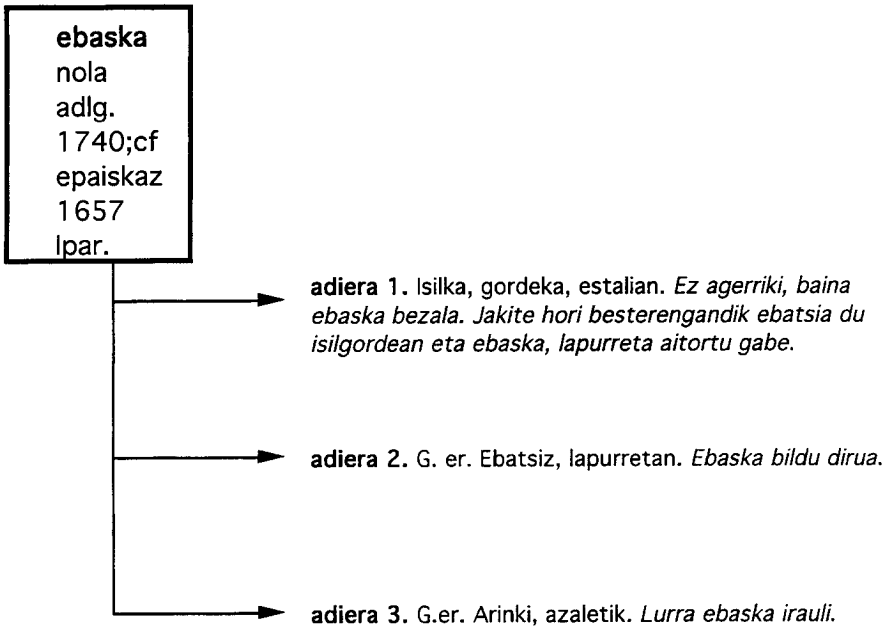
Hiztegi-sarreren egiturazko propietaterik kontsistenteena antolamendu hierarkikoa da. Hemen aztertzen da artikuluak izan ditzakeen banaketak eta azpibanaketak. Kasu batzuetan artikuluak azpiatal bi edo gehiago izan ditzake. Adibidez, bakoitza sarrera-buruaren kategoria-homografo diferenteei dagokio (halaxe banatzen bada artikulua). Beste kasu batzuetan, sarrera banatan etorriko dira kategoria diferenteko homografoak. Bestalde, artikuluak adiera desberdinak izan ditzake, eta era berean adiera horiek azpiadiera desberdinez osa daitezke. Hierarkia oso habiratuak aurki daitezke hiztegietan, zenbait elementu elkarren artean lotuagoak daudela erakusteko, edota adiera xehetasunak oso fin bereizteko.

Hiztegien antolamendu hierarkikoari esker informazioa faktoriza daiteke. Hau da, hierarkiako maila batean espezifikaturiko informazioaren hedadura, maila horri eta

habiratuta dauden guztiei dagokie. Esate baterako, EHn informazio gramatikala (kategoria, azpikategoria, etab.), data, lema-aldaera eta laburdura, oro har, informazio horiek guztiak ondorengo adierei aplikatzen zaizkie. Ikus dezagun EHko sarrera bat faktorizazioaren nozioa ilustratzeko:

**ebaska**. nola adlg. (~1740; cf; *epaiskaz* 1657). *Ipar.*. **1**.Isilka, gordeka, estalian. *Ez ageriki, baina ebaska bezla. Jakite hori besterengandik ebatsia du isilgordean eta ebaska, lapurreta aitortu gabe.* **2**. G.er. Ebatsiz, lapurretan. *Ebaska bildu dirua.* **3**. G.er. Arinki, azaletik. *Lurra ebaska irauli.*

**II.1 Irudia.**-*ebaska* sarrera EHko MRDn azaltzen den bezala.




**II.2 Irudia.**-*ebaska* sarrera, faktorizazioaren nozioa aplikatu ondoren.

Hau da, goiko laukitxo horretan sartu dugun informazioa sarrera-buruarekin batera jarri dugu; esan bezala, informazio hori guztia dituen adiera desberdinei aplikatzen zaiela irudikatuz.

- **Tratamendu aldaera.**

Aurrerago aipatu ditugun egiturazko printzipioen arabera, oso erraz idatzi behar litzateke hiztegiaren egitura deskribatzen duen DTDa. DTD horrek habiratuko luke homografoa sarreren barnean, adierak homografoaren barnean, azpiadierak adieren barnean, etab.

Horrez gain, faktorizaturiko osagai espezifikoak hierarkiako maila egokian azaltzeko baimenduak egon behar dute. Adibidez, <*entry*> definituko litzateke bere barnean sarrera (<*form*>) bat duela eta sarrera horrek homografo bat edo gehiago izan ditzake. Homografodun sarrera bakoitzaren azpian dagokion informazio gramatikala azalduko da (<*gramGrp*>) eta horrez gain adiera bakar bat edo gehiago izan ditzake sarrera bakoitzak (<*sense*>). Jakina, kasu honetan artikulu berean konsideratzen ditugu homografo-zenbaki desberdineko sarrerak. Baina, hiztegi-egitura konplexuagoa eta aldakorragoa dela ezin da ahaztu. Ondorengo puntuetan, TEI lantaldearen hausnarbidearekin jarraituz, hiztegien egituraren berri ematerakoan aurkituriko zenbait arazo jasoko ditugu.

• **Hiztegitik hiztegira dauden desberdintasunak.**

Nahiz eta antolakuntza hierarkikoko eta informazioa faktorizatzeko printzipioak mendebaldeko hiztegi moderno gehienen egituran betetzen diren, hala eta guztiz ere, egitura desberdinak aurki ditzakegu hiztegi diferenteen artean. Adib. hiztegitik hiztegira toki diferenteetan aurki dezakegu etimologiari dagokion informazioa.

Hori dela eta, DTDak ager daitezkeen bariante desberdinak jasotzeko bezain malgua eta orokorra izan behar du, osagai orori edozein posiziotan, eta nahi adina aldiz agertzea baimenduz.

• **Desberdintasunak hiztegi beraren barnean.**

Orokortasunak hiztegi berean ematen diren salbuespenak saihestu behar ditu. Zehatzago izanik, ikusten da nola goi-mailako osagaiak hierarkiako edozein tokitan ager daitezkeen, segun eta zein eremuren gainean aplika daitezkeen goi-mailako informazio horiek. Adibidez, ebakera jasotzen duten hiztegietan, normalean hierarkian posiziorik gorenean azaltzen da eta sarrera guztiari dagokio, hala ere, hierarkian beherago ere azal daiteke homografoaren mailan.

Horrez gain, hiztegi-sarreretan oso hedatua dago *override* sistema, esate baterako, oso arrunta da salbuespenak azaltzea adiera espezifiko baterako faktorizaturiko informazioa aplikagarria ez denean.

Egituran ematen diren aldaketak ez dira bakarrik artikuluaren eduki konplexuak sortuak, izan daiteke ere argitaletxeen estrategia aldaketagatik. Hori horrelaxe gertatzen da hiztegi handiekin, hala nola, *Oxford English Dictionary* (OED) edota *Trésor de la Langue Francaise* (TLF), zeinak hamarkadetan zehar lexikografo-talde desberdinen pean egon izan diren.

Hiztegi beraren barnean aurki ditzakegun egiturazko aldaerek, hiztegietarako DTD orokorragoa eskatzen dute. Hierarkiako maila bakoitzean goi-mailako edozein osagai

(sarrera, homografoa, adiera, azpiadiera, etab.) ager baitaiteke. Horrela bada, osagai hauek biltzen dituen DTDa definitzerakoan, kontuan hartu behar da hierarkian mailak adierazten dituzten etiketek (adib. *<entry>*, *<homograph>*, *<sense>*, etab.) osagai berak bil ditzaketela.

- **Salbuespenak.**

Nahiz eta hiztegietako DTD orokorra egin, eta horrekin edozein tipotako artikuluen egiturak aztertzeko gai garela pentsatu, salbuespenek hauts ditzakete aurreikusitako egitura horiek. OED bezalako hiztegi konplexuetan maiz gerta daitezke horrelakoak.

Hau guztia kontuan izanik, ezinezkoa da definitzea egitura finko bat zeinak argitaraturiko hiztegi ororen jatorrizko egitura jaso behar duen fideltasun osoz. Horrela bada, kasu horietarako irtenbidea DTD askea da, zeinak esaten duen artikulu bateko osagai oro (adib., sarrera, kategoria gramatikala, definizio-testua, erabilera-oharra) artikuluko edozein posiziotan azal daitekeela. Beraz, hiztegien azken DTDak hiztegi-sarrerentzako bi elementu desberdin eskaintzen ditu:

*<entry>*, zeinak hiztegi konbentzionalenen erregulartasunak jasotzen dituen.

*<entryFree>*, zeinak aurrekoan erabiltzen diren oinarrizko osagai berberak baliatzen dituen, baina osagai horien konbinaketa edota antolamendu askea baimenduz.

Hala eta guztiz ere, bi aukera hauen eskaintza ere ez da erabat arrakastatsua, kasu askotan aukera horiekin bat ez datozen sarrerak azaltzen baitira.

- **Orokortasuna vs. botere deskribatzailea.**

DTD oso estu batek emaniko hiztegi baten artikuluen egitura oso ondo deskriba dezake, baina delako DTD horrek agian ez du aplikagarritasunik hainbat hiztegitan, edota hiztegi bereko artikulu guztientzako.

*<entryFree>* delakoaren bidez, aldiz, ez dago murriztapenik egiturari dagokionean, baina horrek berak arazoak sor ditzake zenbait aplikaziotan: balidazioa, informazio-eskurapena eta hiztegi-sarreren aurkezpen tipografiko konplexua.

*<entry>* elementuaren edukia definitu zen sarreren egiturako oinarrizko osagaiak jasotzeko, nahiz eta edozein hiztegi partikularrekiko orokorregia den.

Zenbait aplikaziotarako DTDa moldatu egin beharko da. DTDak eskaintzen du hiztegiak kodetzeko gidalerro multzoa (markoa), egiturazko printzipio orokorretan oinarrituz, eta erabiltzaileek beren beharren arabera moldatu ahal izateko elementuak.

Hiztegietako informazio ugari inplizitua da edota oso konprimituta dago. Hau dela eta, garbi izan behar da azaleko testu-egitura edo aurkezten den informazioaren barne-egitura jaso nahi den. Ikuspuntu biak interesgarriak izan daitezkeenez, bien berri emango diguten metodoak garatuko dira. Testu-egitura eta barne-egitura harremanetan jartzeko adierazpen-moduak ere landuko dituzte.

- **Testu-ikuspegitik kodetuz: berreskurapena.**

Kodeketa-eskemaren arabera kodeturiko jatorrizko testua berreskuratzeko aukera ziurtatzen da. Testu-ikuspegi hertsiz kodetuz gero, kodeketako etiketak ezabatuz gero, jatorrizko bertsioko karaktere-segida berbera izango genuke (komak, parentesiak, sarrera-buruak aukerakoak direneko lokailuak, etab. barne direla). Ikuspegi erlaxatuago batez, komak, parentesiak, etab. ez genituzke kodeturiko artikuluan ikusiko, baina nahi izanez gero hauek ere berreskuratu ahal izango genituzke. Adibidez, *antena* artikuluaren adibidea jarriko dugu testu ikuspegi hertsitik nola jasotzen den erakusteko:

```
<entry>
   <form><orth>antena. </orth></form>
   <GramGrp><pos>iz. </pos></GramGrp>
   <usg type=time>(XVII ea.). </usg>
   <sense n = '1.'>
      <def> Bela-ontzietan, hiru angelutako haize-oihalaren masta mehe eta luzea. </def>
   </sense>
   <<sense n = '2.'>
      <def> Hainbat zomorrok buruan dituzten luzakin mehetako bakoitza. </def>
   </sense>
   <sense n = '3. '>
      <usg type=time>(*1973, 1977).</usg>
      <def> Uhin elektromagnetikoak igortzeko eta hartzeko eroale airetarra. </def>
      <eg><q>Telebista-antena.</q></eg>
   </sense>
</entry>
```

Adibide bera, testu ikuspegi erlaxatutik emanez gero atal bakoitzari dagokion informazioa azalduko litzateke baina bestelako ikurrik gabe. Esate baterako, data goian parentesi artean dugu, baina ikuspegi honetatik jasotzerakoan parentesi horiek desagertu egingo lirateke.

- **Datu-base ikuspegitik kodetuz.**

Ikuspegi honetatik kodetzeak jatorrizko datuak zeukaten itxura originala aldatzera garamatza. Datu-base batean lantzeko orduan izango duten egituratik hurbilago izango dugu.

- **Bi ikuspegiak kodetuz.**

Azken ikuspegiari jarraituz kodeturiko hiztegi baten jatorrizko testua berreskuratzea ezinezkoa suertatzen da, datu-baseetan errepresentatzeko egin behar izaten diren aldaketengatik. Horrela bada, hiztegietako lantaldeak kodeketa berean bi ikuspegiak jasotzeko aukera eskaintzen du, bi ikuspegi hauen arteko harremanak mapaketaren bidez gauzatuz. Kodeketa desberdinen arteko mapaketarako metodo orokorrak, TEI P3ko 14. Kapituluan —*Linking, Segmentation and Alignment*, 393. orrialdean— deskribatzen dira.

- **Norentzat pentsatua dagoen TEI.**

1. Argitaratzaile eta lexikografoentzat, zeinak informazio lexikaleko datu-baseak garatzen diharduten. Estandarizazioak informazio hori modu askotara baliagarria izateko aukera eman dezake, eta hiztegi desberdinak, datu berberetatik abiatuz, egin ahal izateko oinarri bilaka dezake (adibidez, bertsio osoa, bertsio zehatza eta poltsiko-bertsioa). Bestalde, hiztegiko datuak formatu bateratuan izateak, tresneria automatikoa dela medio, harremanak dituzten hiztegien arteko koherentzia ziurtatzea, eta hiztegi diferenteen arteko datu lexikalen trukatzea errazten ditu.

2. Linguista konputazionalentzat, zeinak lengoaiaren prozesamendurako lexikoiak eraikitzeko hiztegietako informazio lexikalaz baliatu nahi diren. Kodeketa formatu estandar eta bateratu bat izateak, datu linguistikoen trukea, eta iturri desberdinetarik informazioa biltzea erraztuko lituzke.

3. Filologoentzat. Alderdi guztietaz daude interesaturik, bai hiztegien formatu fisikoaz (hiztegien konposizio tipografikoaz, orrialde banaketarako etenak, hitz banaketa, etab.), bai edukiaz, eta edukiaren eta formatu fisikoaren arteko erlazioez ere badute interesik. Ikertzaileen artean datu-trukea ahalbideratzen duen kodeketa-formatua eta hiztegiak prozesatzeko bat datorren softwarea erabili nahi dute. Edukiaz ere nahi dute jakin, eta edukiaren eta inprimaturiko erreprodukzioen arteko erlazioez ere badute interesik.

4. Hiztegi-erabiltzaileentzako, zeinak informazio lexikala eskuratu nahi duten datu-base lexikal batean egiten den moduan, baina emaitzak inprimaturiko liburu baten gisara nahi dituztenak.

Hiztegiak formatu bateragarrian izateak hiztegi erabiltzaileentzako ekarriko lukeen abantailetariko bat, formatu elektronikoan dauden hiztegiak prozesatu ahal izateko sor daitekeen software bateragarria izango litzateke.

Hiztegietako lantaldearen lanaren emaitza argienetariko bat, irtenbide orokor baten bila eginiko hiztegietako artikuluen egituraren azterketa sakona, eta ordura arte egin gabe zegoen saioa dugu.

## II.5 MRDen ikerketaren inguruan.

Lexikoaren eraikuntza erabat automatikoa lortzeko asmoa izan arren, garaturiko metodologia gehienak erdiautomatikoak dira. Ezagutza lexikala automatikoki eskuratu ahal izateko, hiztegi konbentzionalak, corpus analizatuak eta thesaurusak erabiliko dira. Bi modutara ekingo zaio lexikoen eraikuntzari :

- Kontzeptuetatik abiatu ondorik sorturiko mundu kontzeptuala lexikoarekin lotzeko.
- Lexikotik atera mundu kontzeptuala deskribatu ahal izateko beharrezkoa den informazioa.

Bigarrenari gagozkiola, iturri lexikal nagusiak, hiztegiak, entziklopediak eta corpusak ditugu. 80.eneko hamarkadaren hasieran giza erabiltzaileari zuzenduriko hiztegiak helburu konputazionalekin erabiltzen hasiko dira. Hiztegiek hainbat datu egituratu eta antolatu dutenaren hipotesiaz abiatzen dira lehenbiziko lanak. Laster konturatuko ziren datu horiek lexikoi konputazional batean integratzeko, zenbait formalizazio eta egokitzapen beharko direla. Bestalde, zenbait informazio azaleratzeko lanak bideratu beharko dira.

Denak ez zetozen bat hiztegien berrerabilgarritasunaz, errorez, zirkularitatez eta inkoherentziaz beteak daudela argudiatuz. Halaxe, dio Fontenellek (Thierry Fontenelle, 1990:91), beste zenbait lexikograforen iritziarekin bat eginik:

> "A huge problem remains when tries to exploit the regularities found in a dictionary: hand-crafted lexicons are not always reliable insofar as they are quite often marred by inconsistencies, errors of omission and errors of commission (on this crucial issue, see Boguraev and Briscoe, 1989; and Michiels, 1982)."

Honezaz gain, informazio ugari inplizituki gordea izatearen traba aipatuko dute (Boguraev eta Briscoe, 1989). Esate baterako, EDRkoek uko egin diote hiztegi konbentzionalen berrerabilgarritasunari, corpusen alde jokatuz. M. Gross ere corpusak erabiltzearen aldekoa da.

MRDen iturria nagusia den arren, badira testu gordinetatik abiatu diren lanak lexikoiak automatikoki eraikitzeko orduan. Lehenengoetarikoa Granger-en FOUL-UP system (Granger 77). Beste batzuk: (Jacobs & Zernik 88), EBL (Asker eta beste, 92). Arlo honetan kokatzen da baita ere ARIOSTO sistema (Basili eta beste, 93a), tresna honek corpusetatik informazio lexikala eskuratzea du helburu. Horretarako metodologia hibridoa erabiliko du, LNPko teknika tipikoak, hala nola, azaleko sintaxia eta etiketa semantikoak, prozesamendu numerikoarekin konbinatuz.

MRDen aldekoek inpliziturik dagoen informazioa esplizitu egiteko metodologiak landuko dituzte. Lehenengo lanetan, MRDen gainean jardungo dute zuzen-zuzenean informazio eskuraketan (Amsler, R.A. 1981; Chodorow, M. eta beste, 1985). Oraintsuagoko lanetan MRDtik datu-base lexikaletara pasatuko da lehenik, datu-base lexikaletatik abiatuz informazio-eskuraketa azkarragoa eta malguagoa izan dadin. Laburbilduz MRDak erabiltzeak dituen alde onak eta ez hain onak hauexek lirateke :

- **Abantailak :**

  a) esfortzu gutxirekin eskura daitekeen informazio kopuru handia (morfologia, kategoria, forma...).

  b) sarreren antolaketa semantikoa sailkapen kontzeptuala egiteko baliagarri gerta daiteke.

- **Desabantailak :**

  a) informazio asko eta asko inplizituki gordea egoteak, esplizitu egiteko metodoen beharra dakar. Esate baterako, definizioaren eremuan gorderik dagoen informazioa eskuratzeko aukeretariko bat taxonomien eraikuntza izango litzateke.

  b) datu kopurua handia izanik, usu aurki genitzake koherentzia-erroreak, bai sailkapenari dagokionean bai zentzuren bati dagokion deskribapena dela eta.

Beraz, iturri lexikal "tradizionalak" hartuz gero, ezinbestekoa izango da prestatze-lana, eta analisiari ekin aurretik hiztegiaren egitura aztertu behar da sakonki.

MRDez eginiko lanez aritzean, iturri honek iturri lexikal gisa izan dituen erabilpenak irudikatu nahi ditugu ondorengo lerroetan. Hala ere, lan horien berri ematen hasi aurretik, interesgarri deritzogu Ide eta Veronis lexikografoek azken hamabost urteetan MRDen inguruan eginiko ikerketei buruz emaniko iritziak.

Lehenago jaso izan dugun bezala, bat datoz beste zenbait autorerekin, esaterakoan MRDez baliatuz proiektu bati abiada eman aurretik ezinbestekoa dela MRDetako informazioaren aplikagarritasuna eta egokitasuna neurtzea. Hiztegia hainbat lexikograforen lanaren ondorio baita, eta egin, berregin, errebisatu eta eguneratua izaten da urteetan zehar. Hori dela eta, informazioak akatsak izango ditu.

MRDen inguruan azken hamabost urteetan eginiko ikerkuntzak erabili izan dituen postulatuak hauexek ziren: MRDetako informazioa baliagarria dela eta erraz eskura zitekeela. Ide eta Veronis-ek (Ide & Veronis 94) postulatu horiek baliozkoak diren ala ez eztabaidatzen dute, eta baita ere ebaluatu urte horietan zehar eginiko lana.

Galdera nagusia da ea eginiko ahalegin hori guztia alferrik izan den? Eta galdera honen erantzuna eman aurretik, ikus ditzagun autore hauei jarraituz MRDen ikerkuntzari buruz eginiko zenbait kontsiderazio.

MRDak ikerkuntzarako erabilgarriak izan daitezen, maiz lan ikaragarria egin beharko da, eta are gehiago, MRDetatik formatu egituratuagoetara egokitze-lana diziplina beregaina bilakatu dela esan daiteke. Inkontsistentziak direla eta, ezin izango da erabateko prozesu automatikorik garatu MRDetako informazioaz aprobetxatzeko. Hala ere, garbi dago MRDek informazio baliagarria dutela, baina eskuzko moldaketen, eta gizakien iritziaren premia izango dutela sistema konputazional batean integratu nahi badira.

MRDen inguruko ikerkuntzak zenbait arlotan izan du eragina, bereziki, hiztegi eta bestelako testu-datuen kodekera eta datu-base modeloen garapenean; eta LNPrako behar den ezagutzaren zehaztapenean. Dena dela, LNPko aplikazio batzuetarako informazioa nolakoa izan behar den ez da oraindik behar adinako sakontasunarekin ikertu.

Agian MRDen inguruko ikerkuntzaren ekarpenik garrantzitsuena, LNP, lexikografia eta argitaratzaile elektronikoen interesak batzearena izan da. Bat etortze honek hiru arloetarako onurak ekarriko ditu. Elkartze honi esker hiztegi hobeak sortuko dira, eta aukera berri ernagarriak ekarriko ditu, dela hiztegi elektronikoetarako, dela hipertestuz osaturiko hiztegi datu-baseetarako, dela LNPrako material erabilgarrien sorkuntzarako.

Arlo honetako ikerkuntzan, lehenago aipatu bezala, ahalegin handiak egin dira hiztegiak informazioaren eskurapena burutu ahal izateko moduko formatu batera egokitzeko. MRDentzako kodekera formala garatu dute TEI ekimenekoek. Kodekera formatu estandarrak software beraren aplikagarritasuna eta MRDen berrerabilgarritasuna ahalbideratzen ditu. Lehenago esan bezala, aukera hauek oso probetxagarriak dira argitaletxeentzako, formatu hauetatik inprimatzeko lana zuzenean bideratu daitekeelako, eta mota diferenteetako hiztegiak ere sor baititzakete (adib. laburra, ikasleentzako, etab.).

MRDentzako egokia den kodekera-formatu egoki bat garatzeak hiztegi-sarreren osagaien identifikazioa eta hiztegien barneko egiturazko printzipioak sakonki ulertzea eskatzen du. Honekin loturik dagoen arazoa MRDen informazioa egoki errepresentatzeko datu-base modeloren bat zehaztearena dugu. Hiztegietako egitura hain da konplexua non datu-base konbentzionalen bidez ezin diren hiztegiak egoki errepresentatu.

MRDen gainean eginiko lanak, eta beste antzeko batzuek bultzatzen dute datu lexikalen eta testu-datuen errepresentaziorako datu-base modelo egokien azterketa. Datu-baseen diseinuari dagokion ikerrarloa gero eta garrantzitsuago bilakatzen ari da.

Bistan da eskurapenaren eta errepresentazioaren gaiak erlazionaturik daudela. Arlo biak ikertuak izan dira eta dira gaur egun ere, proiektu nazional eta internazionaletan. Jakina,

hizkuntza nagusiekin erkatuz gero euskararen alorrean badugu bai zer eginik. Alemanian badago, esate baterako, proiektu bat bi arloak jorratzen dituena 1992. urtetik, ELWIS[1] proiektua.

MRDek izan dute eragina ezagutza-baseen sorkuntzan ere. Eta bereziki lengoaia naturalen prozesamenduan informazio semantikoak dituen motak, natura eta papera ulertzen lagundu digute.

Bukatzeko, autore hauen iritziz, esan daiteke MRDen gainean eginiko lana ez dela alferrikakoa izan. Ondorioa oso bestelakoa da, hau da, lan hori baliagarria gerta daitekeela LNPrako zein lexikografia eta argitaraldi elektronikoetarako. Besteak beste, LNP eta beste komunitate hauetako ikertzaileen interesak bat datozelako.

Bestalde, iturri lexikal desberdinen konbinazioa ezinbestekotzat jotzen da.

## II.5.1 MRDez baliatuz eginiko lanak.

Esan beharra dago informazio lexikalaren eskurapenerako, batez ere MRD elebakarretatik abiatu direla proiektu gehienak, eta arreta txikiagoa jarri dela hiztegi elebidunetan; azken hauetan oinarriturikoen artean, honako hauek aipa ditzakegu: (Ageno eta beste, 94); (Knight & Luk 94).

Laburki bada ere ikus ditzagun ondorik MRDez baliatuz eginiko zenbait lan:

- **Hitz-zerrendak.**

Euskarri magnetikoan gorderiko hiztegietarik hitz-zerrendak sortu izan dira behar desberdinetarako: zuzentzaileak garatzeko (hauetarako algoritmo eraginkor eta sendoak lortzeko) hitz-zerrenda erraldoien beharra izaten da. Zerrenda hauek oso helburu zehatzetarako erabili izan dira eta informazio mugatua ematen dute zerrendaturiko hitzei buruz. Ondoren aipatzen ditugun lanak (Boguraev eta Briscoe, 1989)-tik hartuak dira:

- Yannakoudakis (1983) 93.000 hitz, *The Shorter Oxford Dictionary*; Fawtrop eta Yannakoudakis (1983) 57.000 hitz, *The Teacher's Word Book* 80.000 hitz, (Thorndike eta Lorge, 1944).

- Itzulpen automatikorako (Tucker eta Niremburg 1984). Indexazio automatikorako (Klingbicl, 1985).

- Hitz-maiztasuna egiaztatzeko (Coltheart, 1981; Kucera eta Francis 1967).

---

[1] Alemaneraz ELWIS akronimoa Korpusgestütze Entwicklung lexikalischer Wissensbasen, hots, korpusetan oinarritutako ezagutza-base lexikalen garapena. Proiektu hau Baden-Württemberg-eko Zientzia eta Ikerkuntza Ministerioak babestua da.

- **Sintesirako.**

Sintesian sortzen diren arazoei aurre egiteko MRDetara joko dute batez ere Britainia Handian; eremu honetako hainbat proiektu aipatzen ditu Briscoek (1985).

- **Analizatzaileetarako.**

CRITIQUE (EPISTLE zena; Heidorn eta beste, 1982), IBMko proiektu hau azpimarratzen dute *Computational Lexicography for Natural Language Processing* liburuan (Boguraev eta Briscoe, 1989).

- **Taxonomiak eta semantika lexikala (hiztegien azterketa hizkuntzaren egitura semantikoa ikertu nahirik).**

Amsler izan zen lehena hiztegien berrerabilgarritasunaren aldeko apustua egiten. *Merriam Webster Pocket Dictionary* (MPD) hiztegiaz baliatuko da taxonomia semantikoa eraikitzeko. Definizioetatik kontzeptuen hierarkia semantikoak erauz daitezkeela dio, inplizituki gorderik dagoen informazioa azalaraziz. Hiztegiko definizioez arituko dira hainbat ikertzaile : Amsler eta White 1979; Chodorov eta beste 1985; Alshawi 87; Walker/Zampolli/Calzolari 88; Nakamura/Nagao 88; Veronis eta Ide 1990; Klavans eta beste 1990; Briscoe /Copestake/Boguraev 1990; Vossen 1990; Vossen 1991; Castellón 1993; Wilks eta beste 93; Dolan eta beste 93; Artola 1993; etab., definizioetan gorderiko kontzeptuarteko erlazioak landuko dituzte. Erlazio hauek erauzteko beharrezkoa den metodologia eta programeriaz ere arituko dira hainbat lanetan, eta jakina, behin informazio semantikoaren erauzketa bideraturik, errepresentazioari dagozkion arazoez arituko dira.

Aurreko lanen artean (Artola, 93), Donostiako Informatika Fakultatean garaturiko tesi-proiektua dugu. Lan horretan LPPLren (*Le Plus Petit Larousse*) MRD bertsioa abiapuntutzat hartu eta, giza erabiltzailea gogoan izanik, honi zuzenduriko hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza burutu zuen egileak.

Bestalde, proiektuetan ere batez ere alderdi semantikoa jorratu izan da. Beste batzuen artean elkarrekin erlazionaturik[1] dauden honako hauek aipa ditzakegu: LINKS, LEXALIZA eta ACQUILEX.

Azkenari dagokionez, ACQUILEX proiektuaren helburu nagusia MRDak erabiliz LNPko sistemetarako osagai lexikalak eraikitzeko teknikak eta metodologiak garatzea izan da. Informazio lexikalaren erauzketan —sintaktiko-semantikoa, hainbat MRDtatik eta ingurune

[1] LINKS ('LINKS in the Lexicon'-en laburdura) proiektua (300-169-007) hiru urtez garatu zen Herbehereetako ikerkuntzarako erakunde baten laguntzaz. LEXALIZA proiektua (202.121) Amsterdam-go unibertsitate batek sortu zuen. ACQUILEX (Acquisition of Lexical Kowledge for Natural Language Processing Systems) ESPRIT proiektu bat da (BRA 3030, bi urte eta erdiz garatu zena), eta bertan Amsterdam-go, Bartzelonako, Cambridge-ko, Dublin-go eta Pisako unibertsitateek parte hartu zuten.

eleanitzean— datza proiektuaren funtsa, ezagutza lexikal eleanitzeko base (LKB, *Lexical Knowledge Base*) bakarra sortzea izaki helburu orokorra. Paperezko hiztegien MRD bertsioen gainean eginiko lanak: birformateatzea, berrinterpretatzea eta informazioaren azalaraztea; ez dira hiztegien ezagutza lexikalaren erauzketatik bereiziriko jarduerak. Lan horiek guztiak hiztegietako ezagutzaren errepresentazio estandar eta esplizitu baten bila eginiko urratsak baitira. Proiektu hau orain dela urte batzuk bukatu zen.

MICROSOFT laborategikoek ere zenbait ikerketatarako MRDak hartu dituzte oinarrizko iturri lexikal bezala. Horietako batean, erlazio morfologikoak automatikoki identifikatu nahi dituzte. Hau da, elkarren artean harreman morfologikoa duten hitzak identifikatuko dituzte, eta eratorriaren adierak lotuko dituzte oinarrizko formaren adiera bat edo gehiagorekin erlazio morfologikoen atributuei esker (Pentheroudakis & Vanderwende, 1993). Horrez gain, ezagutza-base lexikalak (Dolan eta beste, 1993) eta hitzen adierak desanbiguatzeko (Dolan, 1994) ere baliatuko dituzte MRDak.

CRL (*Computing Research Laboratory*) New Mexico-ko laborategikoak ere MRDz baliatzen dira maiz beren proiektuak garatzeko. Horrela, hainbat hiztegi elebakar zein elebidun eraiki dituzte, horretarako baliabide desberdinak erabiliz, horien artean MRDak. Hiztegi hauek LNPko aplikazio desberdinetarako erabiliko dituzte, beste batzuen artean itzulpen automatikorako eta bestelako itzulpen-tresnetarako (Zajac, 1998).

- **Datu-base lexikalak eraikitzeko.**

Datu-baseak eraikitzen laguntzeko MRDez baliatu direnen artean, (Pin-Ngern eta beste, 93)ko artikulua aipatuko dugu. Bertan erakusten dute datu-base lexikal zabala eta LNPko hainbat aplikaziotarako eraiki daitekeela MRD desberdinetako materialak konbinatuz.

- **Ezagutza lexikalaren eskurapenerako.**

Ezagutza lexikalaren eskurapenari, oro har, dagokionez oso ikuspegi zabala jasotzen da (Rigau, 98)-n. Bertan, beste batzuen artean, honako eremu hauetako lanak aipatzen dira: hitzen adieren desanbiguaziorako (*Word Sense Disambiguation*), informazioaren eskurapenerako (*Information Retrieval*, IR), desanbiguazio sintaktikorako, itzulpen automatikorako lexikoi elebidunak eraikitzeko, ezagutza-base lexikalak eraikitzeko, MRDak semantikoki aberasteko eta prozesamendu semantikorako.

Oro har, MRDetatik abiatuz, batez ere erlazio semantikoei buruzko informazioa eskuratzeko saioak egin dituzte, metodo erdiautomatiko edota partzialak proposatuz informazio semantikoa eskuratzeko.

Hala ere, ikus dezagun zer-nolako informazioa eskura daitekeen MRDetatik Boguraev eta Pustejovsky-ren arabera: argumentu-egitura, paper tematikoen egitura (*Aktionsart*), qualia egitura (Pustejovsky, 1989); herentzia lexikalaren egitura.

Bukatzeko, azpimarratu nahi nuke, bigarren kapituluko azken atal honetan MRDez baliatuz erdiets daitezkeen informazio-mota desberdinak deskribatzen jardun dugula. Eta gure tesiaren ikergaiari dagokionez, erabilera-adibideak izango ditugu informazio-iturri nagusi. Hona hemen, (Sinclair, 1987:137)n hiztegietako adibideei buruz dioena:

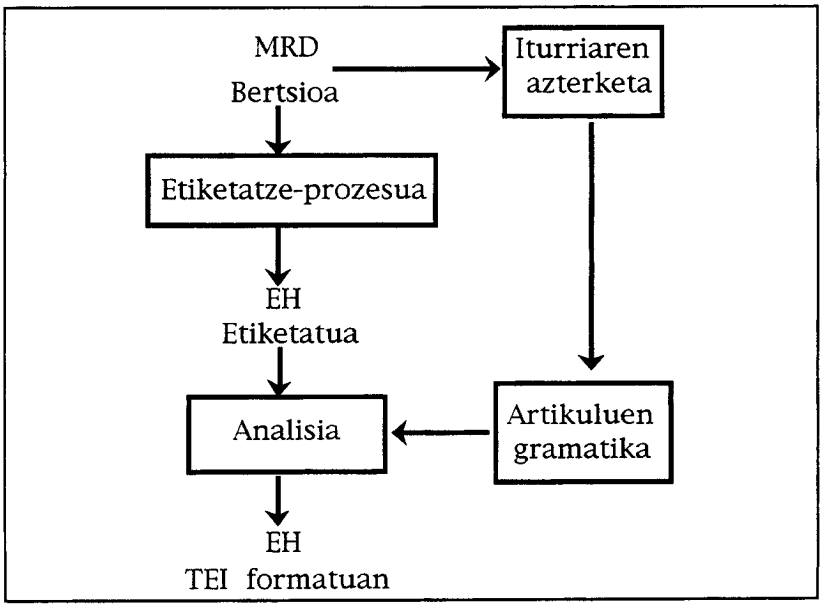
> "This should help to reinforce the meaning -not by acting as a reformulation of the definition, but showing how the word is actually used in appropriate context, a typical grammatical structure, and with words that are normally associated with it."

Baina, adibideen azterketari ekin ahal izateko EHren prestatze-lanari heldu behar zaio lehendabizi. Lehenago aipatu dugun bezala, MRDen ikerketaren inguruko II.5 puntuan, MRDak ikerkuntzarako erabilgarriak izan daitezen formatu egituratuagoetara egokitu behar dira. Egokitze-lan hori burutzeko ezinbesteko lehen urratsa EHren prestatze-lana dugu. Horrezaz gain, aurrerago ere aipatua genuen —§ II.4, Baliabide lexikalak estandarizatzeko premia—, oso garrantzitsua dela baliabide lexikalak berrerabilgarriak izatea. Horrela bada, prestatze-lan horren baitan burutzen dugu iturri lexikal hau TEIko gidalerroen arabera egokitzeko lana. Beraz, TEIra egokitzerakoan estandarizazioaren behar horri erantzuten diogu. Modu horretara, EH TEI formatuan izaterakoan iturri lexikal honetaz balia nahi duen orok eskuragarri izango du. Hain zuzen ere, EHren prestatze-lanaren azalpenari ondorengo kapituluan ekingo diogu.

# III. *Euskal Hiztegia*-ren prestatze-lana.

Hirugarren kapitulu honen helburua EHren azterketa automatikoa ahalbideratzeko jarraitu ditugun oinarrizko urratsak deskribatzea da. Lehen kapituluan, motibazioak azaltzerakoan esan bezala —I.1 atalean— LNPrako "ezusteko" baliabide lexikala dugu EH, MRDn izan arren giza erabiltzaileari zuzenduriko hiztegia baitugu, eta ez baitu gure helburuetarako nahi genukeen egitura, hots, ordenagailu bidez zuzenean aprobetxatzeko modukoa. Hori dela eta, aditzen adibideen prozesamendu automatikoari heldu aurretik hiztegi-testuaren egituratzea burutu behar izan dugu. Egituratze-lan honetan ataza nagusi hauek garatu ditugu: iturriaren ezaugarrien azterketa, etiketatzea, analisirako gramatikaren idazketa, analisia bera eta analisiaren emaitza TEIko gidalerroen proposamenetara egokitzea. Lan honen emaitza, azken urratsari esker, gure lanerako ez ezik euskara aztergai edo lantresna duen edonorentzat baliagarria izango da. Beraz, aipaturiko oinarrizko urratsak ezinbestekoak dira ondorengo azterketei ekin ahal izateko. Horrela bada, lehendabizi EHren MRDtik abiaturik TEIko formatu estandarrera egokitzeko, hiztegi honen gainean egindako azterketak eta prestatze-lanak azalduko ditugu.

Ondorengo irudian, —III.1 irudian— ikus daitezke iturburu-hiztegiaren gainean egindako eraldaketak, MRDtik hasita TEIko errepresentazioraino.




**III.1 Irudia.**-Hiztegiaren prestatze-lanaren urrats nagusiak.

Prestatze-lanaren azken urratsari dagokionean, EHren analisiaren emaitza TEIra egokitzeko lanak deskribatuko ditugu, eta zenbait adibide emango ditugu TEIko formatuan.

Ondoren, prestatze-lanari esker, EDBL aberasteko zuzenean baliatu dugun informazioa azalduko dugu.

Bukatzeko, EHren prestatze-lan honen ondorioak azpimarratuko ditugu.

Bestalde, hirugarren kapitulu honen sarrera modukoa bukatzeko, esan beharra dago oinarrizko urrats hauek HLEHren[1] (Hauta-Laneko Euskal Hiztegiaren) eta EHren gainean burutu direla. Hala ere, ez dira bi hiztegi desberdin, eta halaxe dio hiztegiaren egileak ere (Sarasola, 1997:XV) :

> "Hirugarren idaztaldi hau ez da, alde batetik, bigarrenaren bateratze eta eguneratze bat besterik"

Beraz, urrats hauetan EHri egingo diogu erreferentzia nahiz eta azaltzen diren ezaugarriak bientzat balio duten, eta hala gertatzen ez den kasuetan desberdintasunak nabarmenduko ditugu. Bestalde, sarreran esan bezala, aditzen adibideen azterketarako EHko adibideez baliatuko gara, azken idaztaldi hau, egindako aldaketen ondorioz, egokiagoa iruditzen baitzaigu.

## III.1 EHren ezaugarrien azterketa.

Lexikoaren eskurapenak garrantzi handia du linguistika konputazionalean, eta ikertzaile ugari konbentzituak daude ezin dela sarrera lexikalik huts-hutsetik kodetzen hasi, eskulana izugarrizkoa izango litzatekeelako. Haatik, lehen parteko lehenengo kapituluan zehar aipatu dugun bezala, jada existitzen diren baliabide lexikalen (hiztegiak edota corpus testual handiak) berrerabilgarritasunak, eta hauetatik informazioa eskuratzeko metodologien garapenak berebiziko balioa dute.

Horrez gain, LNPrako osagai lexikalak euskarri magnetikoan dauden hiztegietatik (MRD) abiatuz elikatzeak, hiztegiaren egituraren, eta lexikografoek darabilten kodeen azterketa sakona eskatzen du (Boguraev 1991, Calzolari 1989, Fontenelle 1992).

Beraz, gure lehenbiziko lana iturriaren azterketa izan da, ondorengo ataletan azalduko ditugun honako hiru puntu hauei erreparatuz: EHko informazioaren aplikagarritasuna EDBLn, informazioaren egituraketa eta irizpide lexikografikoak, eta informazioa eskuratzeko metodologia eta nola errepresentatu.

---

[1] HLEHren gainean egindako azterketari buruz informazio gehiago nahi izanez gero, ikus honako artikulu hauek: (Arriola eta beste, 1995;1996; Arriola & Soroa, 1996).

## III.1.1 EHko informazioaren aplikagarritasuna EDBLn.

Proiektuaren helburuak aurkezterakoan —ikus § I.2— esan bezala, EH hiztegi arrunta izanik EDBL aberasteko eta osatzeko baliatu nahi dugu. Baina, baliagarria ote da euskararen prozesamendu automatikoan oinarri lexikala den informazio-biltegi erraldoi horren premiei erantzuteko? Berez, LNPko beharrei erantzuteko pentsatuta ez izan arren, gure ustez baliagarri izan zitekeelako hipotesia genuen. Gainera, arlo honetan bazeuden hainbat proiektu eta ikerlan ildo beretik zihoazenak eta, beraz, esparru honetan bagenuen bide hau jorratu zutenengandik zer ikasi. Lehen urrats bezala, halako "ezusteko" baliabide lexikalaren ustiaketari ekin aurretik, iturri horretan zegoen informazio jori horretatik zein izan zitekeen EDBLrako interesgarri aztertzeari ekin genion. Azterketa hau behar-beharrezkotzat jotzen dute lexikografia konputazionalaren arloan maisu diren Boguraev eta Briscoe ikertzaileek. Ikertzaile hauek, honetaz gain informazioa nola dagoen egituratuta eta zer-nolako tresnen bidez erauziko den aztertu beharra azpimarratzen dute.

Aplikagarria den informazioari dagokionez bi mota nagusi bereiz daitezke: informazio esplizitua eta inplizitua.

Esplizituari dagokionez honako eremu hauetaz aprobetxa gaitezke:

- Artikuluetako buruak eta azpi-sarrerak erabili dira sarrera berriak eskuratzeko.
- Kategoria gramatikala: hiztegi-sarrerek sarrera buruaren ondorik dagokien kategoria azaltzen dute, eta EDBLko kategoria eremua osatzeko baliagarria da (osatzeko zein konprobatzeko jada landurik daudenak eta bat datozenak EH hiztegikoarekin).
- Aditzen artikuluetan azaltzen diren laguntzaile-mota etiketak: *da, du, da-du*, etab., aditz bakoitzak hartzen duen laguntzaile-mota adierazteko egokiak dira.
- Adberbioen artikuluetan, *non, nola*, ... informazioa dugu eta adberbio-mota eremua osatzeko balia dezakegu.
- Erabilerari dagozkion laburdurak jasotzeari ere interesgarri deritzogu, desanbiguatze-prozesuetan lagungarriak izango direlakoan. Batez ere, *g. er., gaur g. er.*, ... moduko laburduren bidez dakigu artikulu horretako sarrera gutxi erabilia dela. Honelako informazioa, desanbiguatze-prozesuetan erabilgarria da, bereziki metodo estatistikoen ikuspegitik, hauetan maiztasuna garrantzitsua baita.
- EHn "+" ikurrarekin azaltzen diren artikuluak, estandarrarekin bat ez datozenak, edota jada zaharkituak daudenak dira (sarrera hilak deitu izan ditugunak gramatikan). Forma hauek EDBLn errore tipikoentzat egokitutako atalean gordetzen dira, eta euskararen tratamendu automatikoaren bidean eginiko lehen tresna orokorra den zuzentzaile ortografikorako erabilgarriak dira (XUXENerako, Agirre et al., 1992).

Informazio-mota hauek guztiak analisia burutu ondoren zuzenean integra daitezke EDBLn.

Informazio inplizituari dagokionez, batez ere definizioak eta adibideak dira interesgarrienak. Eta horrez gain, esan beharra dago inpliziturik dauden informazio-mota hauek eskuratzeko analisia bera erdietsi ondoren, eremu horiek aztertzeko lan-prozedura sakonak garatu behar direla.

Definizioen analisitik informazio lexiko-semantikoa erauzi nahi da EDBLn integratzeko. Erlazio lexiko-semantikoekin EDBL poliki-poliki aberasten joango da; sinonimia, erlazio taxonomikoak, hautapen-murriztapenak, eta abar. Nahiz eta EDBLn oraindik adiera mailako bereizketarik egin ez, horretarako prestutasuna adierazten duten ezaugarrien artean sarreren bereizketarako homografo-zenbakiaren lanketa dugu. Homografo-zenbaki hauek esleitzeko HLEHko homografoak hartu izan dira kontuan.

Lehen kapituluan esan bezala —I.1 puntuan—, adibideetara mugatuko da gure proiektua. Gure kasuan, EHko adibideak corpus "berezi" gisa hartu, eta aditzen adibideen eremutik oinarrizko azpikategorizazio-ereduak eskuratzeko (bai EDBLn txertatzeko, bai ondoren egin daitezkeen azterketa sakonagoetarako) metodologia landu dugu. Adibideak aztertzeko arrazoiak —I.1 atalean— azalduak ditugu, hala ere, puntu honetan gogoratu nahi genuke ezagutza gramatikal edota lexikaleko osagarririk garrantzitsuenetarikoa dugula aditzen azpikategorizazioa. LNPko ikertzaileek gero eta behar handiagoa sumatu dute lexikoi konputazionalak informazio honekin hornitzeko, eta horrela hauetan oinarritzen diren aplikazioak hobetzeko.

Bestalde, badira hiztegian LNPren ikuspegitik oso baliagarriak diren informazioak, baina ezin daitezkeenak jaso, ez dagoelako horiek bereizteko modurik. Honek erakusten du, ondorengo puntuan ikusiko dugun legez, informazioa kodetzeko eredu sendo baten premia. Ikus ditzagun zenbait adibide:

*hertsatu, hertsa, hertsatzen*
*II ("-(r)a") Joatera hertsatu zituzten.*
*II ("-(r)i") Hona amoros batek bere maitearen aitari hertsatzen dion kobla.*
*hertsi, herts, hersten*
*(Era burutua izenondo gisa; 1571)*
*(Era burutua adizlagun gisa....)*
*herren*
*II (Adizlagun gisa)*
*II (Izenondo gisa)*
*(Gauzei buruz) Oinen gainean orekan ez dagoena.*
*(Izen gisa) Herrentasuna.*

Parentesi artean azaltzen diren datu linguistikoak nahiko genituzke jaso, baina hain dira erabiliak parentesiak denetariko datuak eta "edonola" adierazteko, non ezin dugun horien artean dagoen informazioa bereizi (automatikoki edo erdiautomatikoki, jakina).

Ondorengo puntuari lotu aurretik, esan beharra dago, zenbait autorek zalantzan jartzen dutela zenbateraino den LNPren mesederako MRDetatik ateratako informazioa. Autore horien aburuz, aplikazio zehatzetarako izan dezakeen erabilgarritasunaren arabera neur daiteke soilik hiztegi baten erabilgarritasuna.

Hala eta guztiz ere, hiztegietako altxorra nola aprobetxa daitekeen, hortxe dago gakoa. Batez ere, kontuan izanik MRD gehienetan informazioa oso era "gordinean" dagoela antolaturik. Salbuespen gisa, lexikografia konputazionaleko hainbat lanetarako oinarri izan den LDOCE (Longman, 1987) aipa dezakegu.

## III.1.2 Informazioaren egituraketa eta irizpide lexikografikoak.

Levin autorearen hitzetan (Boguraev eta Levin, 1988:14), hona hemen non dauden arazoaren sustraiak:

> "The root of the problem: lexicographer lack a sophisticated enough theory of lexical organization that they can use as guide in order to ensure that the entries are both consistent and complete."

EHren egituraketari gagozkiola, oro har, honako edukia aurkituko dugu artikuluetan: sarrera-burua, data, aldaerak, kategoria gramatikalak, laburdurak, definizioa, adibideak, erlazioak (sinonimia eta antonimia), azpisarrerak eta azalpen gramatikalak. HLEHn azpisarrerak (Sarasolak hitz-forma eta esaldi bereziak deitzen dituenak) artikuluan zehar azaltzen ziren bitartean, EHn sarreraren bukaeran eta hurrenkera alfabetikoan azalduko dira.

Informazio hau guztia antolatzeko, erabiltzen diren kode nagusiak hiztegigileak azaltzen dizkigu hiztegiaren sarreran:

- Adierak: hauek ordenatzeko kronologiari erreparatzen diote. Honako kode hauek darabiltza:
  - A,B, ...adierak multzo berean biltzeko.
  - 1, 2, ... adiera-zenbakiak, adiera desberdinak bereizteko.
  - a), b), ... edota #1, #2, II (azken hau EHn) ... azpiadierak adierazteko.
- Erlazioak:
  - Ik. sinonimiako erlazioa adierazteko.

- Ant. antonimiako erlazioa adierazteko.

- Kode tipografikoak: zenbait kode tipografiko baliatuko dira informazio-motak bereizteko:

  - Letra-molde lodia erabiliko da, honelako atalak bereizteko: sarrera-burua, homografo-zenbakia, adiera-zenbakia, azpiadierak eta erlazioak.

  - Letra-molde etzana erabiliko da: kategoria gramatikalak, aditz-izenen partizipioak, laburdurak eta adibideak adierazteko.

Lexikografoa testu-prozesatzaile batez baliatu da artikuluak egituratzeko, kode tipografikoak erabiltzen ditu nagusiki informazio-mota desberdinak bereizteko, eta artikuluen atal desberdinak antolatzeko. Kode hauek kontuan hartuko ditugu hiztegiaren gramatika idazterakoan.

### III.1.3 Informazioa eskuratzeko metodologia eta nola errepresentatu.

MRDei buruz aipaturiko ezaugarriak direla eta, erraz jabetu gaitezke hauetatik informazioa eskuratzea ez dela nolanahiko lana. Horrexegatik, lexikografia konputazionalaren esparruan hiztegi-sarreren parsing-a garrantzi handiko arloa bilakatu da. Garrantzi honen arrazoi nagusien artetik, lehenago aipatu dugun LNPko aplikazioek duten ezagutza lexikalaren premia larriaz gain, (Bläsi & Koch, 92)-en artikuluan aipatzen direnetatik honako hau jaso dugu:

- Hiztegi zabalagoak eta kalitate handiagokoen eskaera handia dela eta, eta, honekin batera, kontuan izanik hiztegi-sarrerei dagozkien egiturazko ezaugarriak, lexikografoei laguntzeko tresneria behar-beharrezkoa da.

Horrela bada, ezagutza horren garantiazko eskuratzea erdietsi nahi bada, ezinbestekoa da hiztegi-sarreren parsing-a garatzea. Hiztegi-sarreren parsing-erako sistemei eskatzen zaien ezaugarri nagusia, ahal bezain orokorrak izatearena da, eta ahal bada jada existitzen diren osagaiez baliatzea.

Helburu gisa, tresna orokorrak eraiki nahi dituzten proiektuak izan arren, esan daiteke ez dagoela edozein hiztegitik informazioa eskuratu ahal izateko moduko tresna orokorrik.

Gure kasuan, landuriko metodologiak HLEHren eta EHren MRDak analizatzeko balio izan du. Lehenbizi, HLEH analizatu genuen eta honen gainean egindako lanak EHren analisia burutzeko ere balio izan du, eta aurreko lan hartan izandako esperientziak bigarren MRD honen analisia erraztu eta hobetu egin du.

Errepresentazioari dagokionean, informazio lexikala konplexua da, edozein hiztegi-sarrera ikustea baino ez dago. Datu-base klasikoak (erlazionala, a.b.) ez dira ongi egokitzen. Hiztegietan, informazio mota bereko diruditen datuak (ortografia, ebakera, kategoria, definizioak,...) estrukturalki desberdinak diren bitartekoak erabiliz errepresentatzen dira.

Gure kasuan, problemarik zailenetarikoa adiera-habiraketarena dugu, eta irtenbide desberdinak aztertu ondoren, horien artean ACQUILEX taldekoek VOX hiztegiko MRD bertsiorako zerabilten LDB (*Lexical Database*) eredua, azkenean TEIko gidalerroez baliatzea erabaki dugu (ikus § III.2.5).

## III.1.4 EHren ezaugarri nagusiak.

EH hiztegia euskara batuaren biltegi gaurkotua da, hastapenetan arau-emaile izateko bokazioa zuena baina gerora hauta-lanetan erabiltzeko moldatu dena. Euskarazko hiztegi elebakarra da, orokorra, tradizio literarioan oinarritua, 33.111 sarrera (HLEHren bertsioak 30.000 ditu) inguru dituena eta gaur egun lexikoa zertan den jakiteko erreferentzia nagusia.

EHren bertsioaren gainean egindako neurketen arabera, honako datu hauek jaso ditugu: 36.461 adiera ditugu orotara, horietatik 5.586 adiera aditzenak dira, 23.898 izenen adierak, 4.792 adjektiboenak eta 2.359 adizlagunenak.

Hiztegia artikuluz osaturiko bilduma dugu. Eta artikulua sarrera-buruak eta honi dagokion informazioak osatuko dute.

Hona hemen segidan hiztegiko unitate nagusia den artikuluen osaera:

- Sarrera: letra-molde lodiz bereizten dira. Eratorriek ere badute beren sarrera.
- Kategoria: eremu honetan kategoria gramatikalak nahiz ezaugarri morfologikoak ager daitezke. Sarreraren ondoren dator normalean kategoria.
- Data: sarreraren agerpen-data hiztegi edo testu literario batean.
- Definizioa: adiera bakarreko sarreretan definizio-testuak osatzen du eremu hau. Adiera bat baino gehiago dagoenean bereizgarri desberdinak erabili dira definizioak banatzeko, edota ñabardurak markatzeko.
- Erlazioak: eremu honetan sinonimia edo antonimia bezalako adiera-erlazioak adierazteko kode lexikografiko ohikoak erabiltzen dira: Ik., Ant., ...
- Adibidea: testu etzanez adierazten da adibidea, normalean definizioaren ondoren. EHn erabilera okerrei buruzko argibideak ematen dira: *e.* (erabil), *h.* (hobe). Zenbait artikuluren bukaeran eta makoen artean, erabilera okerrak eta horien ordez erabili

behar liratekeenak jasotzen ditu. Adibidez: [*Aldizkaria gaurkotu asmoz: e. aldizkaria gaurkotzeko asmoz].

- Laburdurak: erabilerakoak, tematikoak, hizkuntz mailak, euskalki-argibideak, etab. Oro har, HLEHn azaltzen zirenak ditugu EHn ere; hala ere, badira aldaketa batzuk.
- Azpisarrerak: maiuskulaz ageri dira eta sarreratzat har daitezke, baina beste sarrera baten baitakoak. HLEHn azpisarrerak sarreran zehar azaltzen badira ere, EHn sarrera bukaeran kokatzen dira.
- Azalpen gramatikalak: eremu honetan sarrerari buruzko informazio gramatikalak aurki daitezke.

EHk CD-ROM bertsioa ere badu (HLEHk ez du), baina gure lanerako euskarri magnetikoko bertsioaz (MRD) baliatuko gara. EH irakurgarria da ordenagailuz: testu-fitxategi baten gisara, hiztegiko artikuluak sekuentzialki editatu dira, eta lehenago aipatu dugun bezala, informazio-motak banatzeko bereizgarri tipografikoak erabili dira. Hona hemen MRDko artikulu bat:

**ahalketu, ahalke** edo **ahalketu, ahalketzen**. da-du ad. (~1870; ahalketu 1571; alketu 1621; ahalkatu 1635). Ipar. edo goi. **1**. da ad. Lotsatu. Ahalke zaitez, zerbitzari laxoa. Erori orduko ahalketu ziren, eta estali zituzten beren gorputzak piku hostoz. II Zerbaitez edo norbaitez ahalketu. Ahalketzen baita erokeria haiek ikusteaz eta entzuteaz. **2**. du ad. (1571). Lotsatu. Ik. **ahalkearazi**. Neure bihotzeko hasperenek ahalketzen ninduten. II (XIX b.). Zah. (G. er.). Beldurrarazi. Mamu horrek txoriak ahalketzen ditu.

**III.2 Irudia.**-*Ahalketu* sarrera EHko MRDn azaltzen den bezala.

## III.2 EHren prestatze-lanez.

Esan bezala, EH informazio-iturri aberatsa da, baina informazio hori era gordinean dago eta, giza erabiltzaileari zuzenduta dagoenez, inprimatutako bertsioen akatsak ditu, bereziki informazioaren egituraketa desegokia LNPko premietarako egin nahi dugun erauzketaren ikuspegitik. Beraz, aurrerago aipatu dugun bezala, premiazkoa da edozein tratamendu automatikori heldu aurretik hiztegi-testua egituratzea. Ondorengo puntuetan azalduko ditugu egituratze-lana burutzeko egindako lanak.

## III.2.1 Etiketatzea.

Ataza honen helburua da iturburu-testua —hiztegia— unitateetan segmentatzea, unitate bakoitzari etiketa bat esleituz; etiketatzean artikuluak identifikatzeko behar diren item-ak eta eremuak identifikatzen dira. Badira osagai horien artean informazio lexikografikoari buruzko kodeak eta estiloari edo letra-moldeari dagozkion kode tipografikoak. Gainontzeko testu-atalak *TX*, hots, testu bezala kodeturik azalduko dira. MRDko sarreran ikus daitekeenez —ikus III.2 Irudia—, nahiz eta sarreraren zatiak esplizituki adierazita ez egon, lagungarri izan daitezkeen kode tipografiko eta lexikografikoak egon badaude, eta hauetaz baliatuko gara gure lanari aurre egiterakoan. Kode horiek guztiak —ikus III.3 Irudia— lagungarriak dira giza erabiltzaileak hiztegia kontsultatzerakoan. Baina, ordenagailuak uler ditzan azalarazi egin behar dira, eta horretarako sortu ditugun markak edota etiketak —ikus III.4 Irudian— azalduko ditugu.

| **Kode tipografikoak** | |
|---|---|
| | letra lodia |
| | letra etzana |
| | letra maiuskula |
| | azpimarra |
| **Kode lexikografikoak** | |
| Adiera-multzoa | **"A."**, **"B."**, **"C."** |
| Adiera-zenbakiak | **"1."**, **"2."** ... |
| Adiera-xeheak | "a)", "b)", "#1", "#2", II |
| Erreferentziak | "Ik." |
| Antonimoa | "Ant." |

**III.3. Irudia.**-Hiztegian erabilitako kode tipografiko eta lexikografikoak.

Ondoko taulan ikus daitekeen bezala, kode tipografiko eta lexikografikoez gain baditugu bestelako batzuk ere, hala nola : *data hasiera/bukaera, laburduraren hasiera/bukaera, kategoriaren hasiera/bukaera eta maiuskula hasiera/bukaera.*. Etiketa hauek sortzearen arrazoia analisi xehea egin nahi izatea da. Jakina, zenbat eta etiketatze finagoa orduan eta informazio xeheagoa jaso ahal izango dugu.

| Markak | Deskripzioa | Oharrak |
|---|---|---|
| [SH] /[SB] | Hiztegiko artikulu bakoitzaren hasiera/bukaera | |
| [LH] / [LB] | Lodia hasiera/bukaera | Kode tipografikoa. WORD fitxategian letra-molde lodiz markatutako zati guztiak. |
| [EH] / [EB] | Etzana hasiera/bukaera | Kode tipografikoa.WORD fitxategian letra-molde etzanaz markatutako zati guztiak. |
| [MH] / [MB] | Maiuskula hasiera/bukaera | Kode tipografikoa. |
| [AMH] / [AMB] | Adiera multzo hasiera/bukaera | Kode lexikografikoa. |
| [AZH] / [AZB] | Adiera-zenbaki hasiera/bukaera | Kode lexikografikoa. |
| [HZH] / [HZB] | Homografo-zenbaki hasiera/bukaera | Kode lexikografikoa. |
| [IKH] / [IKB] | Erlazioaren hasiera/bukaera | Kode lexikografikoa. |
| [ANH] / [ANB] | Antonimoaren hasiera/bukaera | Kode lexikografikoa. |
| [AXH] / [AXB] | Adiera xehearen hasiera/bukaera | Kode lexikografikoa. |
| [DH] / [DB] | Dataren hasiera/bukaera | Datak bereizteko etiketa. |
| [LAH] / [LAB] | Laburduraren hasiera/bukaera | Laburdurak bereizteko etiketa |
| [KAH] / [KAB] | Kategoria hasiera/bukaera | Kategoriak bereizteko etiketa. |

**III.4 Irudia.**- Etiketatzean erabilitako markak edota etiketak.

Etiketatzearen urratsa gauzatzeko erabilitako markak edota etiketak azaldu ondoren, hona hemen MRDko *ahalketu* sarrera etiketaturik:

[SH][LH]ahalketu, ahalke edo ahalketu, ahalketzen.[LB] da-du [KAH] ad. [KAB] [DH](~1870; ahalketu 1571; alketu 1621; ahalkatu 1635).[DB][LAH] Ipar. edo goi. [LAB][AZH]1.[AZB] da [KAH] ad.[KAB] Lotsatu. [EH]Ahalke zaitez, zerbitzari laxoa. Erori orduko ahalketu ziren, eta estali zituzten beren gorputzak piku hostoz.[EB] [AZH] II [AZB] [EH]Zerbaitez edo norbaitez ahalketu. Ahalketzen baita erokeria haiek ikusteaz eta entzuteaz.[EB] [AZH] 2.[AZB] du [KAH]ad. [KAB] [DH](1571).[DB] Lotsatu. [IKH] Ik. [IKB] ahalkearazi. [EH]Neure bihotzeko hasperenek ahalketzen ninduten.[EB] [AZH] II [AZB] [DH] (XIX b.).[DB][DH]Zah. [DB] (G. er.). Beldurrarazi. [EH]Mamu horrek txoriak ahalketzen ditu.[EB] [SB]

**III.5 Irudia.**-*Ahalketu* sarrera etiketatzea burutu ondorik.

## III.2.2 Gramatikaren idazketa.

Gramatika idazteak hiztegiaren egitura orokorra islatzea eskatzen du, hau da, artikulu bakoitzean zer osagai eta nolako hurrenkeran azalduko diren jasoko dugu. Horregatik, gramatika idatzi aurretik artikuluen egitura aztertu zen: osagaiak, kodeak, etab., eta hauen guztien kokapena eta agertzeko modua. Oso garrantzitsua deritzogu gramatika honen idazketari, zeren eta gramatika honek adierazten du lexikografoak, bere artikuluak idazterakoan, erabili duen egitura, gramatika.

Osagaiak bereiztu ahal izateko etiketatze-urratsari esker badakigu artikulu bakoitza non hasi eta non bukatzen den, eta hauetako atalak ere non hasi eta non bukatzen diren. Esate baterako *adibideak*, etiketa edota marka hauen artean dauden testuak ditugu: [EH] / [EB]. Baina, horrez gain posizioa ere kontuan hartu behar dugu atal horiek taxuz bereizi nahi badira. Horretarako, eskuz aztertu ditugu etiketatzean bereizten ditugun atal horiek zein posiziotan ager daitezkeen. Adibidez, posizioari dagokionez *adibideak*, oro har, definizioaren ondorik azalduko direla dakigu.

Gramatikaren bidez, atal horiek bereizita ditugula aprobetxatuz eta posizioa kontuan harturik, artikulu orok bere barnean dituen atalak ezagutzen ditugu. Gramatika adierazteko formalismoa DCGa (*Definite Clause Grammar*) dugu, baina formalismo horretako erregelen ordez, metalengoaia batez baliatuko gara sarreren egitura orokorra deskribatzeko:

- <A>/<B> A edo B, aukerako sinboloak.
- <A> <B> A eta B, Aren jarraian B.
- [<A>] aukerako sinboloa.

- <*NULL> sinbolo hutsa.

- <*A> bukaerako sinboloa.

Erregelen sintaxia:

<ELEMENTUA>=<EL1> <EL2>... <ELN>.

Aurrerago esan dugun moduan, gramatika orokorra irudikatzen dugu —III.6 Irudian— aurkezten ditugun erregela horien bidez. Analisirako erabili dugun gramatika osoa ikusi nahi duenak, jo beza A eranskinean aurkezten den DCG gramatikara.

Hona hemen sarreren egitura orokorra jasotzen duen gramatika:

```
<SARRERA> =<LEMA> [<ERLAZIOAK>] <KATEGORIA>
                [<*DATA>] [<DEFINIZIO_ADIBIDEAK>]
                [<ADIERA_XEHETASUNAK>]
                [<AZPI_SARRERAK>].
<LEMA> = [<HOMOGRAFO_ZENBAKIA>] [<LEMA_HILA> / <LEMA_ESTANDARRA>].
<HOMOGRAFO_ZENBAKIA> = <*HH> <*ZENBAKIA> <*HB>.
<LEMA_HILA> = <*GUR> <*LH> <*LEMA> <*LB>.
<LEMA_ESTANDARRA>= <*LH> <*LEMA> <*LB>.
<KATEGORIAK> = [<*AZPIKATEGORIA>] <KATEGORIA>.
<KATEGORIA> = <*EH> <*KAT> <*EB>.
<DEFINIZIO_ADIBIDEAK> = <DEFINIZIOA> [<ADIBIDEAK>]
                        <DEFINIZIO_ADIBIDEAK> / <*NULL>.
<DEFINIZIOA>=[<ADIERA_ZENBAKIA>][<ADIERA_MULTZOA>]
            <*DEFINIZIOA> [ERLAZIOAK].
<ADIERA_ZENBAKIA> = <*LH> <*ZENBAKIA> <*LB>.
<ADIERA_MULTZOA> = <*AMH> <*ADIERA_MULTZOA> <*AMB>.
<ERLAZIOAK> = [<SINONIMOA> / <ANTONIMOA>]
              <ERLAZIOAK> [ADIBIDEAK]. /<*NULL>.
<SINONIMOA> = <*IKH> <*SINONIMOA> <*IKB>.
<ANTONIMOA> = <*ANH> <*ANTONIMOA> <*ANB>.
<ADIBIDEAK> = <*EH> <*ADIBIDEA> <*EB>.
<ADIERA_XEHETASUNAK> = <ADIERA_XEHETASUN_IKURRA>
                       [<KATEGORIAK>] [<DEFINIZIO_ADIBIDEAK>] <ADIERA_XEHETASUNAK>
                       /<*NULL>.
<ADIERA_XEHETASUN_IKURRA>=<*AXH> <ADIERA_XEHETASUN_IKURRA> <*AXB>.
<AZPI_SARRERAK> = <*AZPI_SARRERA> [<ERLAZIOAK>] <KATEGORIA>
                  [<*DATA>] [<DEFINIZIO_ADIBIDEAK>]
                  <AZPI_SARRERAK> /<*NULL>.
```

**III.6 Irudia.**-Sarreren egitura orokorra jasotzen duen gramatika.

Goiko irudian —III.6 Irudian— ikus daitekeenez, edozein informazio edozein ordenatan agertzeko aukera ematen da. Hori nabarmenago ikus daiteke DCGari begiratu bat emanez gero (ikus A eranskina). Hala ere, elementu desberdinak konbinatzeko aukera hain da handia non ezin izan ditugun zenbait sarrera ezagutu gramatikaren bidez. Hots, gure gramatikak ez ditu egitura horiek aurreikusi. Nolabait, esan daiteke sarrera horiek duten egitura konplexuagatik gure gramatikak ezin izan dituela ezagutu. Adibidez, honako sarrera hauek aipa ditzakegu: *2 ondo, oro, 1 izan, izaten* (ikus EH hiztegia).

Bestalde, esan beharra dago ezagutzen diren sarreretan zenbait errore ager daitezkeela. Horietako gehienak zenbait informazio-mota ez dagokion formatuarekin azaltzeagatik gertatuko dira. Esate baterako, kategoriari dagokion informazioa adibide gisa ezagututa ager dakiguke. Adibidez, halaxe agertzen zaigu EHko zenbait aditzen adibideetan: ageri, aisatu, lasterkatu, ohartu, piztu, etab.

Beste kasu batzuetan, ez dago modurik laburdurak eta definizio-testua bereizteko. Adibidez *baiadun* artikuluan egiten dugu topo arazo horrekin:

*<entry>*
*<form><orth>baiadun.</orth></form>*
*<GramGrp><pos> izond. </pos></GramGrp>*
*<usg type=time>1749 </usg>*
*<usg type=geo.>Ipar.</usg>*
*<def>G.er.. Erruduna, hobenduna.><</def>*
*<eg><q>Ez dira hain baiadun.</q> </eg>*
*</entry>*

*Baiadun* artikuluaren definizio eremuan ikus daitekeenez, definizioarekin batera *G. er.*. laburdura dugu definizioaren eremuan. Laburdura gisa *G. er.* espero genuen azaltzea, baina beste puntu bat gehiagorekin agertzen zaigunez (*G. er.*.), ezin dugu laburdura gisa ezagutu. Hori dela eta, azkenean definizio-testu moduan agertzen zaigu.

Bestalde ikur berak erabiltzeagatik informazio desberdinak adierazteko, zenbait informazio ezin izango ditugu bereiztu gure analisi automatikoaren bidez. Horixe gertatzen da, esate baterako, makoen artean agertzen diren erabilera okerrekin. Adibidez: *1 abendu* sarreran : [*Abenduak 20 egin zuten bilera: e. Abenduaren 20an egin zuten bilera]. Ikur berak, hots, makoak erabiltzen dira EHn sarreren lehen agerraldiak adierazteko ere. Hori dela eta, esan bezala, analisi automatikoaren bidez ezagutu ahal izateko gramatika hobetu beharra dago.

Bestalde, horrek guztiak erakusten du formalizazio zorrotzago baten premia nabaria dela. Beraz, lexikografoari eska dakioke adierazpide desberdinak erabiltzea informazio bakoitzerako. Adibidez, SGML erabil zitekeen informazioaren formalizaziorako. Hau da, informazio bakoitza nola adieraziko den aurretik definitua egon behar du, eta horrez gain, modu horretara adierazia agertuko dela bermatzen duten lan-prozedurak jarraitu behar lirateke. Esate baterako, lan-prozedura ezinbestekotzat jotzen dugu hiztegiaren gramatika idatzia izatea hiztegiaren idazketari ekin aurretik.

## III.2.3 Analisia.

Analisi-prozesuaren emaitza hiztegiko testua izango da, baina gramatikaren bidez definitutako egituraren arabera antolatuta. Hau da, emaitza hiztegi-informazio egituratua izango dugu, artikuluak eta hauetako atalak bereizten direla: definitutako hitza, kategoria, definizioa, adibidea, etab. Horrez gain, atal bakoitza osatzen duten osagaiak ere bereizten ditugu. Esate baterako, definizioaren atalean, adiera-zenbakia edota adiera-multzoa azaltzen diren bereizten ditugu, eta ondoren definizio-testua eta erlazioak baldin badaude.

Egitura honetan kode tipografikoak ez zaizkigu azaltzen, hemendik aurrera eman behar diren urratsetarako ez baitute interesik. Hala eta guztiz ere, artikuluaren jatorrizko formatua berreraiki nahi balitz, ez litzateke inongo arazorik izango kode tipografiko horien berreskurapenerako.

Analisiaren emaitzei dagokionez, hona hemen HLEHren eta EHren emaitzak:

- Eskuartean dugun HLEHren bertsioak 22.767 artikulu ditu, eta horietarik 19.290 artikulu analizatu ditugu. Beraz, %85 ezagutzen du gure analisiak. Ezagutu diren horietarik %90 ondo analizatua dago (neurri hori 300 artikuluz osaturiko lagin baten azterketatik ateratzen dugu). HLEHko analisiari buruzko informazio gehiago nahi izanez gero ikus (Arriola eta beste, 1995; 1996; Arriola & Soroa, 1996).

- EHren bertsioak 33.111 artikulu ditu, eta horietatik 30.042 guztiz analizaturik daude, hau da, gure analisia[1] gai izan da %98.49 ezagutzeko. Ikus daitekeenez aurreko HLEHaren bertsioarekin baino estaldura handiagoa lortzen dugu. Eta, aurreko kasuan ez bezala, analisi partzialak ere erdietsi ditugu. Hau da, erabat ezagutuak izan diren artikuluez gain, 3.074 artikuluetatik atal batzuk berreskuratu[2] ahal izan ditugu, hau da, artikulu hauetan ez ditugu atal guztiak ezagutzen, baina ezagutzen direnak jaso egiten ditugu. Eta azkenik, 5 sarrera ezin izan ditugu ezagutu ez osorik ez zatika. Ezagutzen dena zein doitasunarekin analizatu dugun ikusteko, 100 artikuluko lagina aztertu dugu. Azterketa horretan ikusi dugu %97 artikulu zuzenak direla, honako atal hauei erreparatuz gero: sarrera, kategoria, adiera, definizioa eta adibideak. Zeunden erroreak data edota kode gramatikalen ataletan aurkitu dira gehienetan; gainontzeko ataletan erroreak gutxiagotan agertu zaizkigu.

Konparazio modura, analisien arrakastari dagokionez %85 / %90 bitartekoa baino gehiago ezin dela lortu aipatzen dute (Heid et al. 92)n.

---

1 EHren gainean bi analisi egin ditugu: bata, DCGko gramatika erabiliaz, eta bestea, patroi-ezkonketaren (*pattern-matching* ) bidez. Bigarren honen zeregina da gramatikak ezagutzeke utzi duenetik ahalik eta informazio gehien biltzea, batez ere, honako eremu hauei erreparatuz: lema, data, kategoria, definizioa eta adibideak.

2 Gramatikaren bidez ezagutu ez direnak, berreskuratze-urratsean ezagutzen dira patroi-ezkonketa teknika jarraituz.

Aipatu beharra dago, EHren analisiaren emaitzak hobeak direla analisi-estrategian egindako aldaketa bati esker. Hau da, HLEH analizatzerakoan ez zen emaitza partzialik onartzen eta artikulu oro osorik ezagutzea edo batere ez ezagutzea erdiesten zen emaitza gisa. EHren analisian, aldiz, analisi partzialak onartzen dira, eta, horrela, nahiz eta artikulu batzuetan atal guztiak ez ezagutu, ezaguturikoak jasotzeko aukera dago.

Bestalde, kontuan hartzekoa da egin dugun analisiak artikuluen atal guztiak ezagutzea izan duela xede, eta, jakina, horrek gerora begira iturri lexikal hauek egituratze xehe batez antolatzen ditu. Baina, analisia zaildu egiten du, hau da, egin zitekeen, analisi hain sakona egin beharrean, analisi "errazago" bat, esate baterako, sarrera-buruak, adibideak eta definizioak soilik bereizten dituena. Modu horretara eginez gero, emaitza hobeak eta probabilitate handiagoa izango genuke informazio bakoitza dagokion eremuan ezagutua izateko, baina, bestalde, orain bereizita ditugun zenbait informazio ezin izango genituzke aztertu.

## III.2.4 Akatsen tratamendua.

Automatikoki eginiko analisiaren ondorik, ezagutu gabe geratu diren artikuluen (edota artikulu atalen) zein ezagutu ditugunen lagin batzuk aztertu dira akatsen zuzenketa bideratzeko, eta ikusteko zein ziren ez ezagutzearen arrazoi nagusiak. Horrela bada, akatsak sailkatu ditugu aurrera begira hauen tratamendu erdiautomatikoa bideratzeko:

- Errore tipografikoak edota sakatze-akatsak, hau da, WORD testu-prozesadorean aritzerakoan lexikografoak eginikoak, askotan hiztegiaren zenbait atal ez dagokien letra-moldeaz azaltzean ezin izango ditugu analizatu edota gaizki analizatuko ditugu. Adibidez, kategoria letra etzanez adierazia azaltzea. Adibidez (HLEHtik hartua):

> **isastu, isats, isasten.** du ad. (1930) Erratzatu, eskobatu. Sukaldea isastu. #1 Haize garbiak hodeiak isastu ditu eta zerua argitu.

- Gramatikan aurreikusi ez ziren elementuen agerpena. Esate baterako, hiztegiaren sarreran ematen diren laburduren zerrendatik at dauden laburdurak ager daitezke. (Arazo hau, HLEHren kasuan aurkitu dugu). Horrela bada, artikulu batetan zerrenda horretan azaltzen ez den laburdura bat agertu zaigu *v.*. Ikur hori *Ik.* laburduraren baliokidea da betetzen duen zereginaren arabera. Jakina, espero ez dituzun laburduren agerpenak analisia zaildu egiten du. Adibidez:

> **mehaka.** *iz.* (*1745, 1929). Aldaka, gorputz-alboa. *Mehakako mina.* #1 *Berezk.* Gorputz-alboetan enborraren behealdean gertatzen den irtenune bakoitza. v. **aldaka (2), hanka (2).** *Mehaketatik heldu zion.* #2 MEHAKA-HEZUR. v. **mehakezur.** *Baratzatik zetorren, esku bat mehaka-hezurrean, saski bete barazkirekin.*

- Sarreren egiturari dagozkion arazoak. Esaterako, adiera bat baino gehiago dituzten sarreretan adiera bakoitzaren aurrretik adiera-ikurra agertu behar litzateke, baina zenbait sarreratan adieraren aurreko adiera-ikurrik ez zaigu azalduko. Adibidez, EHtik jaso dugun ondorengo adibidean, lehengo adierari dagokion adiera-ikurra falta da.

> **harrotasun**. iz. (*1745, 1747). Harroaren, hantustetsuaren nolakotasuna edo izaera. Ik. **hantuste**; **harrokeria**. Harrotasunak galtzen ditu, azkenean, handiki guztiak. Beren buruari ederretsiz, harrotasunean igeri dabiltza zerriak lokatzetan bezala. **2.** (1899). Nor berak, goresgarritzat edo aipagarritzat hartzen duen zerbaitez sentitzen duen gogo betetasun bidezkoa. Bere seme nagusiaz agertzen duen harrotasuna. Gay harrotasunaren eguna. **3.** (*1745). Harroa edo trinkotasunik gabea denaren nolakotasuna. Harrotasunik gabeko soinekoekin. Belakia eta, bere harrotasunagatik, gutxi pisatzen duten beste gauzak.

- Egileek zehazten dituzten zenbait kode lexikografiko, beraiek adieraziriko eginkizunez gain beste erabilera batzuetan ere aurki daitezke. Hau ere HLEHn aurkitu dugu. Hiztegiaren aurkibidean esaten da lehen agerraldiei dagozkien ikurra dela izarñoa. Hala ere, gurutze-ikurraren betekizunetarako erabilia da. Hona hemen zenbait adibide:
  *mezpera. Ik. bezpera.
  *mezprezatu. Ik. mesprezatu.
  *mermelada. Ik. marmelada.

- Lehenago aipatu dugun moduan, badira zenbait informazio prozesu automatikoaren bidez ezin izan ditugunak bereizi. Adibidez, EHren kasuan erabilera okerrei dagokien informazioa. Hori horrela gertatzen da, makoen ikurra zeregin gehiagotarako erabilia delako. Kode edota ikur berdinak informazio desberdinetarako erabiltzeak sortzen duen "anbiguotasuna" dugu arazo nagusietariko bat gramatikaren arrakasta eta ezagutzen denaren zuzentasuna erdiesteko.

Hiztegia erabat ongi ezagutua edukitzeko modu bakarra eskuzko prozesu baten mende dago: ezagutu diren sarrera akastunak eskuz zuzenduz, eta analizatu ez direnak ingurune egoki baten bidez osatuz. Zuzenketa zein osatze-prozesu horretan, hiztegiko informazioa SGMLz kodetuko genuke TEIko gidalerroak jarraituz.

Azterturiko hiztegi hauek (HLEH zein EH) oso-osorik datu-base lexikal batean gorde nahi badira, edota TEIko gidalerroak jarraituz kodetuta adierazi nahi badira, akats hauei aurre egiteko eta ezagutzeke geratzen diren sarrera (edota sarrera-atalak) osatzeko ingurune egoki bat prestatu behar da lexikografoaren lana arintzeko. Beraz, analisi automatikoaren bidez ezin izan dena ezagutu, edota ez dena erabat zuzen analizatu, erdiautomatikoki landu beharko da.

## III.2.5 EH TEI formatuan.

Badugu, orain arte egindako lanaren ondorioz, EH euskarri magnetikoan, non bere sarreren sakoneko egitura ezagututa eta etiketatua dagoen —ikus § III.2—. Lehenago esan bezala, konputagailuz trata daitekeenez, hiztegi honek euskararako baliabide lexikal paregabekoa eskaintzen digu. Hala ere, baliabide lexikal anitzi gertatzen zaion bezala, gureak ere arazo nagusi bat dauka, hots, baliabide honetan oinarritzen den edozein tratamenduk gure hiztegiko errepresentazioari lotuta egon behar baitu. Arazo hau ez da berria eta azken hamarkadan komunitate zientifikoak hainbat saio egin du sarrera lexikalen errepresentazio orokor bat lortzeko. Hau dela eta, EHren azterketaren emaitza TEIk testu elektronikoak kodetzeko eta trukatzeko bere gidalerroak —TEI P3-n, (Sperberg-McQueen & Burnard, 1994)— proposatzen dituen etiketen bidez adieraztea erabaki dugu —ikus § II.4.1—.

Ikus ditzagun zehatz-mehatz EHren MRDtik abiatuta lortutako baliabide lexikala adierazteko eredutzat TEIren aukeraketaren arrazoiak:

- Lexikografoen lanerako onuragarria delako, informazioaren egituraketari dagokionean kodeketa-sistema finkoa eskaintzen baitu. Ez da inola ere gauza bera, esate baterako, testu-prozesadore batez editatuko hiztegia, testu huts dena, edo honelako adierazpide aberatsago eta egituratuago bat erabiliz osatutakoa.

- Honen inguruan garatzen ari diren tresnak kontuan harturik, informazioaren atzipena, kontsulta-aukeren ugaritasuna eta aberastasuna, etab. ahalbidera daitezkeelako.

- Estandarra denez, informazioaren berrerabilgarritasuna eta elkartrukatzea gauzatu baitaitezke.

Abantaila hauek kontuan harturik, EHren azterketaren emaitza TEIn proposatzen diren etiketen bidez adieraztea erabaki dugu. Hau da, egituratze-lanaren emaitza gisa hiztegiaren analisia erdietsi dugu, eta lorturiko emaitza egituratu hau formatu estandarrera egokitzeko TEIn proposatzen diren etiketen bidez adieraziko dugu analisiaren irteera. Beraz, estandarizatze-lana DCG gramatikaren emaitza egituratua kodetze-arau horiek jarraituz moldatzean datza. Moldatze-lan hori da, hain zuzen ere, TEIko kodeketaren arabera errepresentatzeko egin beharreko lana.

Ondoren, analisiaren emaitza TEIko kodeketaren arabera errepresentatzeko, gramatikaren eta TEIko etiketen artean definituriko kidetasunak deskribatuko ditugu. Horri ekin aurretik, esan beharra dago hiru multzo nagusitan sailkatzen direla TEIko etiketa horiek beren eginkizunaren arabera: egitura hierarkikoa antolatzeko etiketak, goi-mailako osagaiak adierazteko etiketak eta esaldi-mailako osagaiak adierazten dituztenak. Azken etiketa horiek, goi-mailako osagaiak adierazteko erabiltzen diren etiketen baitan azalduko dira. Eta goi-mailako osagaiak adierazteko erabiltzen direnak zuzenean azalduko dira hiztegi-sarrera batean.

Egitura hierarkikoa adierazteko dauden etiketetarik, <entry> eta <sense> baliatzen ditugu. Multzo honetan agertzen da <hom> etiketa ere, hots, homografoa adierazten duena. Baina, HLEHn zein EHn homografoak ez doaz artikulu berean, baizik eta aparteko artikuluetan azaltzen dira, hori dela eta, bi homografoak sarrera gisa tratatuko dira. Hala ere, nahi izanez gero, <superentry> batean biltzeko aukera legoke.

Goi-mailako osagaiak adierazteko etiketei dagokienean, honako hauek baliatu ditugu: <form>, <gramGrp>, <def>, <eg>, <usg>, <re>. Hauen baitan azal daitezkeen osagaiak adierazteko etiketak gramatika eta TEIko etiketen arteko parekotasunak zehazterakoan azalduko ditugu.

Horrela bada, EHko (zein HLEHko) analizaturiko edozein sarrera TEIko erara errepresentatu ahal dugu automatikoki. TEIko erara ikusi aurretik, laburki bada ere, urrats hau egin ahal izateko jarraituriko eremuen egokitzapena azalduko dugu. Egokitzapen hauen bidez adierazi nahi dena da, gramatikako atal bakoitzari TEIko kodekeraren arabera zer etiketa dagokion. Ikus daitekeenez geziaren ezkerretara dauden atalak gramatikari dagozkionak dira eta gainontzekoak TEIko eremuak dira —eremu horiek guztiak hasiera (*<form>*) eta bukaera (*</form>* ) esplizituki dute adierazia—:

**•lema** ⟶ **<form>** type: <orth> </orth> **</form>**
Adibidez: *<form><orth>edale. </orth></form>*

**•lema_hila** ⟶ **<form>** type: <orth> <variant> </orth> **</form>**
Adibidez: *<form><orth> bakotx.<variant> Ik.bakoitz </variant></orth></form>*

**•data_zer**[1] ⟶ **<usg type=time> </usg>**
Adibidez: *<usg type=time>1643. </usg>*

**•kategoriak** ⟶ **<gramGrp> </gramGrp>**
numeroa ⟶ <number>
mota/azpikategoria ⟶ <subc>
kategoria/kategoria_ald ⟶ <pos>
Adibidez: *<GramGrp><pos>izond eta iz. </pos></GramGrp>*

**•laburdurak** ⟶ **<usg> </usg>**
Honako laburdura hauek bereiz daitezke:
<usg type = geo> ⟶ adib. *batez ere bizk.*
<usg type = dom> ⟶ adib. *hizkl.*

[1] Hasieran <notes> eremuan sartu genuen, ez baitzuten aurreikusia beraiek azterturiko hiztegietan dataren beharra egongo zenik. Azkenean erabilera-eremuan sartu dugu data, TEIko zerrendan gure arazoa plazaratu ondoren jasoriko erantzunak kontuan harturik.

<usg type = reg> ⟶ adib. *herr.*
<usg type = style> ⟶ adib. *irud.* , *adkor.*
<usg type = time> ⟶ adib. *zah.*
<usg type = gram> ⟶ adib. *batez ere pl.*

Badira zenbait laburdura anbiguotzat har daitezkeenak, hots, mota bat bakarra ez dutenak. Horrelakoak sailkatu gabe emango ditugu <usg> jarriaz. Adibidez:

*gaur ipar edo herr, zah edo herr*

**•adiera_ikur ⟶ <sense n= 1> </sense>**
**•adiera_multzoa ⟶ <sense n= A> </sense>**

**•adiera_xehetasunak ⟶ <sense> </sense>**
Adibidez:

*<sense>*
*<eg><q>Talai etxea. </q></eg>*
*</sense>*

**•definizioa ⟶ <def> </def>**
Adibidez: <def>Edozein isurkari irentsi.</def>

**•adibidea ⟶ <eg> <q> </q></eg>**
Adibidez: <eg><q>Arto musker, mendi, baserri zaharrak; ale gorriz abailduta sagarrak.</q></eg

**•erlazioa ⟶ <xr type = syn >**
<lbl> Ik. </lbl>
<ref target = erlazio-hitza>
</ref>
**</xr>**
Adibidez: *<xr type = syn> Ik. bahitura, berme. </xr>*

**antonimoa ⟶ <xr type = antonym>**
<lbl> Ant. </lbl>
<ref target = erlazio-hitza>
</ref>
**</xr>**
Adibidez: *<xr type = ant> Ant. egoskor </xr>*

**azpi-sarrerak ⟶ <re> </re>**
Adibidez: *<re>*

<form><orth>TALAIAN </orth></form>
<GramGrp><pos>adlag. </pos></GramGrp>
<usg type=time>1967. </usg><usg type=geo>Bizk. </usg>
<def>Zelatan. </def>
<eg><q>Baina dukesa zaldunaren talaian zebilen.</q></eg>
</re>

Ondoren, TEI formatuaren arabera kodeturiko zenbait sarreraren adibideak emango ditugu:

<entry>
<form><orth>ahalketu, ahalke edo ahalketu, ahalketzen. </orth></form>
<GramGrp><subc>da-du </subc><pos>ad. </pos></GramGrp>
<usg type=time>~1870; </usg> <form type=variant><orth>ahalketu </orth><usg type=time>1571; </usg></form>
<form type=variant><orth>alketu </orth><usg type=time>1621; </usg></form>
<form type=variant><orth>ahalkatu </orth> <usg type=time>1635</usg></form>
<usg type=geo>. Ipar. edo goi. </usg>
<sense n='1. '>
<GramGrp><subc>da </subc><pos>ad. </pos></GramGrp>
<def>Lotsatu. </def>
<eg><q>Ahalke zaitez, zerbitzari laxoa. Erori orduko ahalketu ziren, eta estali zituzten beren gorputzak piku hostoz. </q></eg>
<sense>
<eg><q> Zerbaitez edo norbaitez ahalketu. Ahalketzen baita erokeria haiek ikusteaz eta entzuteaz. </q></eg>
</sense>
</sense>
<sense n='2. '>
<GramGrp><subc>du </subc><pos>ad. </pos></GramGrp>
<usg type=time>1571. </usg><def>Lotsatu.</def>
<xr type = syn> Ik. ahalkearazi. </xr>
<eg><q>Neure bihotzeko hasperenek ahalketzen ninduten. </q></eg>
<sense><usg type=time>XIX b.. </usg><usg type=time>Zah. </usg>
<def>G. er.. Beldurrarazi. </def>
<eg><q>Mamu horrek txoriak ahalketzen ditu.</q></eg>
</sense>
</sense>
</entry>

**III.4.1 Adibidea.** *ahalketu* sarrera TEI formatuaren arabera.

Aditz-sarrera honetan azpimarratu nahi genuke nola sarrera-mailan laguntzaile-mota *da-du* dela, baina lehenengo adieran *da* laguntzaile-mota duela eta bigarren adieran *du* laguntzaile-mota duela. Beraz, adibideak aztertzerakoan sarrera-mailako laguntzaile-mota hori kontuan hartu arren, lehentasuna emango zaie adiera bakoitzerako zehazten diren laguntzaile-motei.

```
<entry>
<form><orth>maitale. </orth></form>
<GramGrp><pos>izond. eta iz. </pos></GramGrp>
<usg type=time>1897. </usg>
<sense n='1. '>
  <def>Maite duena, norbait edo zerbaitenganako maitasuna duena.</def>
  <xr type = syn> Ik. maitatzaile; maitari. </xr>
  <eg><q>Jainkoen jauresle eta gurasoen maitale. Zure Egilea, zure Eroslea, zure Ongile
  eta maitalea. Ni beti izango naiz abarkadunen maitale. Herri maitale sutsua. Maitale
  maitemindua. Bihotz maitalea. </q></eg>
  <sense>
    <def>Pl. Elkar maite dutenak. </def>
    <eg><q>Maitaleek, ordea, ez dute nahi izaten ikusleen begiraturik. Maitaleen
    bakarkako hizketak. Maitale bikoteak. </q></eg>
  </sense>
</sense>
<sense n='2. '>
  <GramGrp><pos>iz. </pos></GramGrp>
  <def>Norbaiti buruz, berarekin ezkontzatik at sexu harremanak dituen pertsona.</def>
  <xr type = syn> Ik. amorante; ohaide. </xr>
  <eg><q>Bere senarrak maitale batekin ihes egin zuela.</q></eg>
</sense>
</entry>
```

**III.4.2 Adibidea.** *maitale* sarrera TEI formatuaren arabera.

Aurreko adibide horretan (III.4.2 Adibidea), lehenengo adieraren osaera komentatu nahi dugu. Ikus daitekeen bezala, lehenengo adiera horretan, lehendabizi definizioa dugu, ondoren sinonimo bi eta zenbait adibide. Adibide horien ondorik, definizio gisa ezagutu dugun atala azalpen gramatikal gisa-edo har genezake. Azalpen horretan, *Pl.* laburdura dugu (plurala adieraziz) eta informazio gramatikal gisa sailkatua egon behar zuen arren, gure analisiak ez du bereizketarik lortu. Azkenik, azalpen gramatikal moduko horri dagokion adibidea ezagutu dugu.

```
<entry>
```

<form><orth>gar.</orth></form>
<GramGrp><pos> iz. </pos></GramGrp>
<usg type=time>1545. </usg>
<sense n='1. '>
<def>Erretzen ari den zerbaitetik altxatzen den eta surtan dagoen gas masa argitsua. </def>
<eg><q>Arbazten gar bizia. Mila kandelaren gar dardarakorrek argitua. Gizonen hitzak, gora-goraka garraren irudiko dabiltzanak. Gar distiratsuak. Garrik gabeko sua. Kea eta garra dariola. Infernuko betiko garretan. Zerua garretan zegoen, airea txoriz josia. Egur hezeak egiten du gar eta egiten du negar, sua datxekio eta ura dario. Haur besoetakoaren soak haragia erretzen zion, gar bat balitz bezala. Su garra; su eta gar. </q></eg>
<xr type = syn>Ik. su. </xr>
<eg><q>Su eta garrezko ibai handi bat. </q></eg>
<sense>
<eg><q>Maitasunaren garretan erretzen. </q></eg>
</sense>
</sense>
<sense n='2. '>
<def>Lehia bizia.</def>
<xr type = syn> Ik. kar.</xr>
<eg><q> Gaztetan, sasoiaren kemena eta garra gainezka zerionean. Bere gorrotozko garra hoztu zenean. Bero daukat nik ere bihotzean garra. Bihotz sutsu haren garra ez da oraindik ilaundu. Ez duzula gar bera otoitzetan. Ume txikiak ipuinak entzuten dituzten garraz. Honelako lehiaketek gure baserritarrei euren abereak sendoago hazteko garra ematen zieten. Non dugu, oi, lehengo gar eta kar hura?</q></eg>
</sense>
</entry>

**III.4.3 Adibidea.** *gar* sarrera TEI formatuaren arabera.

Aurreko adibide honetan ere (III.4. 3 Adibidea), adiera-eremuaz jardungo dugu. Lehenengo adiera honela osatzen da: definizioa, adibideak, sinonimoa, adibidea eta adiera-xehetasun moduko bat adibide batez adierazia.

Aurreko bi adibideetan, adiera-eremua izan dugu hizpide. Horren arrazoia izan da erakustea adiera-eremua dugula konplexuenetarikoa ongi analizatzeko.

Atal honekin bukatzeko, aipatu beharra dago analisiaren emaitzaren eta TEIko kodeketaren artean parekotasunak ezartzerakoan ez dugula arazorik izan. Esan dezakegu TEIko hiztegigintzako gidalerroak bere horretan "eroso" egokitu zaizkiola EHri. Hori dela eta, ez dugu aldaketarik edota atributu berririk sortu beharrik izan. Hor aurkeztu ditugunak dira baliatu

ditugun ezaugarriak. Hala ere, badira zenbait ezaugarri baliagarri suertatu ez zaizkigunak, esate baterako, goi-mailako osagaien adierazpenerako erabili daitezkeenetatik ez ditugu ondorengo hauek erabili: <trans>, <xr>, <etym>. Eta, goi-mailako osagaien adierazpenerako erabili ditugun zenbait ezaugarritatik, ez ditugu hauetan eskaintzen zaizkigun atributu guztiak baliatu. Adibidez, erabili dugun <form> ezaugarriaren baitan azal daitezkeen atributu hauek ez ditugu erabili: *compound*, *derivative*, *inflected*, *phrase*, *orth*, *pron*, *hyph*, *syll*, *stress*, *lbl*..

TEIz kodetzearen zenbait onura aipatu ditugu lehenago ere. Hala ere, puntu honen bukaeran gogora ekarri nahi genuke berriro ere, halako kodetzea hiztegigintzan erabiltzea oso lagungarria dela hiztegi kontsistente bat osatzeko. Hiztegi-egitura modu esplizitu eta formal batez adieraztea lexikografoen zein baliabide lexikal horretaz aprobetxatu nahi diren guztien mesederako dela uste dugu. Azalduriko egituratze-lan honi esker aprobetxatze-lana tresna informatikoez baliatuz gauzatu daiteke, esate baterako EHko definizioen azterketari ere heldu ahal izan zaio lan honi esker (ikus Agirre eta beste, 1997). Gure kasuan, ondorengo kapituluan azalduko dugun legez aditzen adibideez aprobetxatzeko lanari lotuko gatzaizkio.

## III.3 EDBLren aberasketa.

Puntu honetan HLEHtik jasoriko informazio esplizituaz arituko gara. Behin iturria prestatua izan ondorik, informazio hauek guztiak eskuratzeko urratsak automatikoki burutu dira. Hona hemen datu-basean jaso ahal izan diren datuak:

- Sarrera berriak: 8.009.
- Homografoen lanketa: EDBLko sarrerei homografoa esleitu zaie HLEHko irizpideak jarraituz.
- Hitz baten arrarotasunaren berri ematen duen eremua gehitu da EDBLn, hain zuzen ere HLEHko *G.er.* laburdura duten sarrerak dira horietako batzuk.
- HLEHko (edota EHko) azpikategoria-etiketa landu nahi da EDBLko laguntzaile-mota (Lag_Mota) eremuan, aditzei dagokien informazioa osatzen joateko. Azken informazio-mota hau, oraindik integratzeke dago EDBLn, zailtasunak baitaude horretarako. Hau da, laguntzaile-mota aditzen adierari lotua dago, beraz EDBLko aditzen sarreretan hiztegiko adiera-bereizketa berak egin bitartean zail dugu informazio hau datu-basean jasotzen.

EDBLren aberasketarako, HLEHaz baliatu gara aipaturiko informazioak jasoaz, hau izan zelako lehendabizi aztertu genuen hiztegia. Dena den, kontuan izanik sarrera aldetik EHk sarrera berriak dituela, sarreren aberasketa burutu daiteke berriro ere EHko analisiaren emaitza aprobetxatuz.

# III.4 EHren prestatze-lanaren ondorioak.

Azterketa honek ematen du aukera hiztegia modu honetara egiteak dakartzan gabeziak eta akatsak nabarmenarazteko. Horrez gain, nabarmendu beharra dago sarreren kontsistentzia eta osotasuna bermatzeko behar-beharrezkoa dela hiztegiaren egitura modu formal eta esplizitu batez taxutzea, artikulu ororen atalak eta atal bakoitzeko informazio-motak zer moldez adieraziko diren zehatz-mehatz definituz. Hain zuzen ere, analisi automatikoa burutzeko prestatze-lana dela medio, hiztegiaren egitura azaleratu dugu, batez ere analisirako gramatikari esker ikus daitekeen legez. Hain zuzen ere, gramatika hori izan du buruan lexikografoak hiztegia egiterakoan.

Ildo honetatik, egungo hiztegigintzan ezin daiteke ahatz informatika lexikografoen lanerako euskarri ezinbestekoa dela, eta aldaketa sakona ekarri duela lexikografiaren mundura. Esate baterako, datu-base lexikalak oso tresna lagungarriak dira lexikografoen hainbat eginkizunetarako: artikuluen idazketarako, kontsistentzia bermatzeko, bertsio desberdinak lantzeko, eta, normalizazio bidean dagoen hizkuntza baten kasuan, baita gertatzen diren aldaketa edo hartzen diren erabakien berri eguneratua gordetzeko ere.

Bestalde, testu huts izatetik, formatu eta adierazpide aberatsago batera egokitzeko oinarrizko urratsa ere bada egindakoa. Adierazpide berrituak prozesamendu informatikorako ateak zabaltzen ditu, bai informazioa bildu eta gordetzeko —datu-base erlazionalen moldean adibidez—, bai informazioa erabiltzeko, hitz eta adieren arteko erlazioak ustiatuz. Hiztegi elektronikoei eskatzen zaizkien ezaugarriak, informazioa atzitzeko arintasuna, etab., lortu ezinezkoak dira ez bada aldez aurretik hiztegien azterketa sakonik egiten. Paperezko formatuan argitaratu ondoren, hiztegi horren bertsio elektronikoa[2] sortzea jomuga interesgarria iruditzen zaigu kontuan izanik gaur egun maneiatzen ditugun teknika eta metodoak.

Iritzi hori zenbait arrazoi desberdinek justifikatzen du:

- Erabilgarria: zinez uste dugu euskarazko hiztegien erabiltzaileek eskertuko dutela, eta ohitura hartu ahala gehiago eskertu ere, hiztegiak bertsio elektronikoetan ere argitaratzea. Gure iritziz, badago nahikoa arrazoi hiztegi elektronikoen erabilera paperezkoena baino aberatsago gertatuko dela uste izateko. Onura horiez ohartu eta hiztegi elektronikoak erabiltzen ikastea denbora eta ohitura da batez ere. Esan behar

[2] Izen bereko proiektu bi onartu ziren; bata EHUn (erref. UPV 141.226-TA073/96) eta bestea Gipuzkoako Diputazioan (erref. 779/1996): "Hauta-Lanerako Euskal Hiztegiaren Bertsio Elektronikorantz". Proiektu hauen xedea EHren bertsio elektroniko osoago bat egitea zen. Horretarako, bertsio hori sortu aurretik, jada kalean zegoen EHren CD-ROM-ak eskaintzen zituen funtzionalitateak aztertu genituen. Hauek ezagutu ondoren, EHren bertsio elektroniko osoagoa garatzeari ekin genion. Bestalde, aurten bertan, Europako Elkartearen laguntzaz (FEDER proiektua, 2FD97-2000-2001), EH oinarri gisa hartzen duen beste proiektu bat garatuko da. Proiektu honetan EHren TEIko bertsio automatikoa eskuz orraztuko da TEIko gidalerroak jarraituz.

da, gainera, badirela dagoeneko euskarazko hiztegiak bertsio elektronikoan merkaturatu direnak.

- Euskararen modernizaziorako aurrerapausoa: teknologia berrien aroan euskarak ez luke atzera geratu behar eta, alde horretatik begiratuta ere, oso mesedegarri litzateke lan hau.

- Aurreko puntuarekin lotuz, hiztegi elektronikoak oinarrizko baliabidearen rola jokatuko luke beste hainbat aplikazio automatikorako.

- Corpusetako lexikografiaren osagarri da lan hau. Uste dugu honelako lanak corpusetan oinarritutako lexikografiaren osagarri izan daitezkeela, zeren eta hiztegi komertzialek erabilgarria den informazio oparoa dute, are erabilgarriago izango dena lexikografoentzako ingurune batean jarriz gero (*lexicographer's workbench*).

Ondorio bezala ere, azpimarratu nahi genuke azterketaren emaitza eredu estandar batera moldatzeak dakartzan onurak. Estandar internazional gisa ezagutzen den markatze-lengoaia batez, hots, SGMLz kodetua dugu EH TEIko gidalerroak jarraituz. Eta markatze-lengoaia horren abantaila nagusien artean honako hauek ditugu: argitasuna, sinpletasuna eta zorroztasun formala. Bestalde, lehenago aipatu bezala, TEIko gidalerroak ere "eroso" egokitu zaizkio hiztegiari.

Beraz, estandarizatze horren onuren artean, honako hauek azpimarratuko genituzke: iturri lexikal hauek gure azterketaz haraindi berrerabilgarriak izatea eta estandarren inguruan sortzen ari den tresneriaz baliatzeko aukera izatea.

Bukatzeko, esan beharra dago, kapitulu honetan deskribatu dugun prestatze-lanari esker, ja adibideen azterketa ahalbideratzen duen urratsa egina dugula. Hurrengo kapituluan, lehendabizi, adibideen azterketa burutzeko metodologia garatzerakoan kontuan izandako aurrekariak deskribatuko dira, eta ondoren, bosgarren kapituluan, adibideen azterketarako garatutako metodologiaren deskribapenari ekingo diogu, adibideen azterketan jarraituriko urratsak azalduz.

# BIGARREN PARTEA: EHko ADITZEN ADIBIDEEN AZTERKETA

## IV. EHko aditzen adibideen azterketarako aurrekariak.

Laugarren kapitulu honetan, bi ardura nagusi izango ditugu: azpikategorizazioa lantzeko zenbait hurbilpen jasotzea, ondoren gure lana zehazteko, eta metodologiaren baitan garrantzi handia hartzen duen Euskararako Murriztapen Gramatika (EUSMG) deskribatzea.

Lehenengoari dagokionean, jarraian, gure tesi-proiektutik hurbilago dauden lanen berri emango dugu azpikategorizazioa automatikoki eskuratzeko zenbait lan azalduz —ikus § IV.1— . Hurrena, lan honetan oinarrizko argumentu-egituraz —ikus § IV.2— zer ulertzen dugun azaltzeaz arduratuko gara.

Bigarrengo ardurari dagokionean, EUSMG deskribatu baino lehen, CG formalismoa aurkeztuko dugu —ikus § IV.4—. Kontuan izan behar da CG formalismoaz baliatu garela Euskararako Murriztapen-gramatika garatzeko. Horrez gain, gogora dezagun CGz baliatuz burututako lanak lehenengo kapituluan aipatzen genuen azaleko sintaxiaren gaia —ikus § I.4.2— lantzera eraman gaituela, hein batean bederen. Lehendabizi, labur bada ere, formalismo hau aurkeztu aurretik, analisi sintaktikorako joera nagusiak deskribatuko ditugu —ikus § IV.3— formalismo hau kokatzearren. Ondoren, lehenago aipatu bezala CGren aurkezpen orokorra egingo dugu. Eta, hurrengo atalean —ikus § IV.7— formalismo hau aukeratzeko arrazoiak aipatuko ditugu. Jarraian, EUSMGko atal nagusiak deskribatuko ditugu —ikus § IV.8—. Horrez gain, hautatu dugun formalismoaren jarraipen gisa har daitekeen Dependentzia-Gramatikan oinarritutako parserraren (*Dependency Grammar Parser*) ezaugarri nagusiak izango ditugu mintzagai —ikus § IV.5—.

Bukatzeko, sintagmak atzemateko zenbait tresna deskribatuko ditugu. Batez ere, hurrengo kapituluan azalduko dugun kateen osaketatik hurbilen daudenak.

## IV.1 Azpikategorizazioa automatikoki eskuratzeko lanak.

Azpikategorizazioaren nozioak hainbat alderdi dituen legez, informazio hori eskuratzeko ere ikuspegi desberdinetatik heldu zaio gai honi. Gu batez ere Linguistika Konputazionalaren esparruan egin diren zenbait saioz mintzatuko gara.

Argumentu-egituraz informazio zehatza izatea eskatzen zaie lexikoi edota hiztegiei Linguistika Konputazionaleko aplikazio errealak garatzeko. Behar honi erantzuterakoan hiru bide nagusi hartu izan dira: lehenengoa, lexikoiak eskuz egitearena (adib. Fitzpatrich eta Sager, 1974, 1981); bigarrena, MRDetatik abiatuz hiztegia automatikoki eraikitzea oso kodeketa gramatikal zehatzarekin (adib. Boguraev eta Briscoe, 1987); eta hirugarrena, lexikoiak automatikoki eraikitzea corpusetatik eskuraturiko informazioarekin.

Badira estrategia hauek konbinatzen dituztenak ere, (Boguraev eta beste, 1987), eta (Carroll eta Grover 1989) Alveyko *Natural Language Tools* (ANLT) hiztegia eraikitzeko hurbilpena deskribatzen dute. Hurbilpen horretan *Longman Dictionary of Contemporary English* (LDOCE, Procter, 1978) MRDa erabiltzen dute tresna gisa hitzen sailkapenerako ANLTko klase lexikalen arabera. Prozesua erdiautomatikoki gauzatu zen; sailkapena ezin zenean zehaztu LDOCEko MRDan oinarrituz, sistemak lexikografoari galdetzen zion klase egokia hautatzeko prozesua burutu ahal izateko.

Antzerako proiektua garatu dute (Grishman eta beste, 1994)-k, COMLEX hiztegia osatzeko eskuzko lana errebisatuz eta finduz domeinu publikoko MRD baten gainean, eta corpusetatik eskuraturiko informazioa konbinatuz.

Lexikoiak, eskuz landuak zein MRDetatik eratorriak lexikografo eta linguisten lanean oinarrituta daude, eta gehienek akats berberak izango dituzte: ezosoak eta inkontsistentziaz beteak izatea. Hiztegiak eskuz eraikitzeak daukan arazorik larriena horrek dakarren kostu handia da. Horrekin batera kontuan hartu behar da ez dela erraza lexikografoarentzat aditza erabili ahal deneko testuinguru guztiak zehaztea. Landu beharreko sarrera kopurua hain da handia non ezin den ziurtatu denen egokitasuna (zeinak are eskuzko lan gehiago eskatzen duen). Hiztegiak lexikografo taldeek garatuak direnez, MRDetatik informazioa eskuratzeak berriz ere arazoa lehenengo mailan jartzen digu, hau da, eskuratze-prozesua abian jarri ahal izateko formatu egokiago batera moldatu behar baitira. Askotan egokitze-lan horri ekiterakoan, automatikoki ebatzi ezinezko akatsekin egingo da topo, horrelaxe diote (Ide & Véronis, 1994:280)-n:

> "To make the MRD usable for research, considerable effort was often required, and in fact the translation of MRDs in typesetter format to something more usable has become an area of study in itself."

> "Because of the inconsistency, fully automated procedures cannot determine the appropriate interpretations."

Honi gaineratzen zaion arazoa da, MRDko errepresentaziotik gramatikak edota parserrak behar duen eredura egokitze-prozesuarena. Azkenik, esan behar da ez dela hiztegietan konplementazio-patroien maiztasunari buruzko informaziorik ematen, eta hau beharrezkoa dela ikuspegi estatistikotik lantzen diren analizatzaileentzat.

Aipatu ditugun irtenbideetarik, corpusetatik argumentu-egiturari buruzko informazioa jasotzearena hartzen da egokientzat eta esperantzagarrientzat.

Ondoren, arestian esan bezala, azpikategorizazioa automatikoki lantzeko zenbait saio aurkeztuko ditugu gain-gainetik; azpimarratzekoa da ikusmolde desberdinetatik jorratu arren gai hau, gehienak datozela bat corpusa oinarritzat hartzerakoan.

Brent-en hurbilpena oinarritzen da esaldi ezanbiguoetan eta corpuseko oso informazio zehatzean. Corpus etiketatu (Brent eta beste, 1991a) zein etiketa gabekoetatik (Brent, 1991b) aditzen azpikategorizazio-ereduak eskuratzen ahaleginduko da. Estrategia nagusia, ahalik eta informazio sintaktiko gehien ateratzea corpusetik, eskuratze-prozesuan oso erraz formaliza daitekeen informazio gramatikala erabiliz. Informazio gramatikalaren oinarriak *The Case Filter of Rouvert and Vergnaud* (1980)-tik jasotzen ditu. Zenbait patroi lexikal definitzen ditu (gehienetan klase itxietakoak, adib. izenordeak) pista bezala erabiliko dituenak ingeleserako konplementazio-patroiak zehazteko. Manning-ek, (Manning, 1993), oso ikuspegi murriztua duela irizten dio Brent-en bideari, ez baitu informazio sintaktikorik baliatzen, eta beraz oso corpus handia behar du oso konplementazio-patroi gutxi ateratzeko.

Brent-en ildotik, hau da, edozein testutatik azpikategorizazio-ereduak erdiautomatikoki erdietsi nahi dituzten lanen artean dugu Dimitrios Kokkinakis-ena ere (Kokkinakis, 1996). Sistema honen helburu nagusia datu-basean dagoen balentziari[1] buruzko informazioa corpusean egiaztatzea da. Informazio hori egiaztatu ondoren, hedadura zabaleko suedieraren

---

[1] Kontzeptu hori erabilia da gorago dagoen egitura baten menpe dauden osagaien kopurua eta ezaugarriak adierazteko, esate baterako, konplementuak Fink (1977), Allerton (1982). Zenbait autoreren arabera *syntactic valence* eta *semantic valence* bereiztu behar dira, adibidez Vater (1975), Allerton (1982). Aditz baten balentzia sintaktikoa zuzenean ikus daiteke perpausean. Balentzia sintaktikoa aditzaren mende dauden kopuru zehatz bateko unitate sintaktikoez osatzen da. Unitate sintaktiko hauek aditz horren konplementu gisa jokatzen dute eta nahitaezkoak dira aditzaren esanahia osatzeko. Elementu hauek aditzari azaleko egituran laguntzen diotenak dira.

Bestalde, predikatu baten balentzia semantikoa bere ezaugarri semantikoen batuketa da. Predikatuaren eta bere argumentuen konbinaketa posibleen bitartez adierazten da. Mota honetako balentziaren bidez, rol semantikoen kopurua eta ezaugarriak zehazten dira. Beraz, balentzia-mota hau esaldiaren sakoneko egiturarekin dago lotuta.

gramatika eraikitzeko erabiliko da. Horretarako, bi iturri nagusi baliatuko dituzte: aditzen balentziak dituen datu-basea (*Verb Valency MRD*) eta datu-base lexikala (*Gothenburg Lexical Database*, GLDB). Bi iturri horiek eskuz eraikitakoak dira, eta maila batean bederen corpus erreal batean egiaztatu da bietako informazioa. Egiaztapen hori burutzeko corpus hauek baliatu dituzte: *Swedish Language Bank* (SLB) eta *Stockholm-Umea Corpus* (SUC). Sistemak iturri horiez gain, etiketatzailea eta aditz-lematizatzailea erabiliko ditu. Sistema honek aditzen argumentuak atzematen ditu corpusean. Horretarako azpikategorizazio-balentziari buruzko informazioa duen datu-base batez baliatzen da. Datu-base honetan aditzen lemak, eta hauei dagozkien argumentu-egiturak eta aztarnak gordetzen dira. Horiek guztiek lagunduko dute heuristikoetan oinarritzen den algoritmoari argumentu-egiturak identifikatzen. Argumentuen etiketak modu bat dira esaldiko osagaien arteko azaleko harremanak azalerazteko. Algoritmoaren bidez datu-baseko balentziari buruzko informazioa corpusean aztertzen da, eta horrela ikusiko da bat datorren, baztertzekoa den edo aldatu egin behar den jasoriko emaitza enpirikoen arabera.

Manning-ek (1993), lan honetan ere kategoria gramatikalez etiketaturiko corpusa eta egoera finituko izen-sintagma analizatzaile bat baliatuko ditu, 19 konplementazio-patroi ezagutzeko xedearekin. Sistemak proposatzen dituen sarrera lexikaletan analisi sintaktikotik etor daitezkeen erroreak ekiditeko, teknika estatistiko batez iragazten ditu proposamenak. Horrela bada, filtro horren arabera maiztasun handiarekin patroi okertzat hartzen direnak zigortuak izango dira, eta aditz baten patroi legez onartu aurretik aditz horren adibide-proportzio handiago batekin aurkitu beharko dira.

(Monedero eta beste, 1995)-n SOAMAS izeneko tresna aurkezten zaigu. Tresna hau garatzen dute espainierako aditzen azpikategorizazio-ereduak automatikoki jasotzeko morfosintaktikoki etiketaturiko corpusetatik. Esaldi sinpleekin lan egiten dute, eta esaldi horietako aditz-sintagma ezagutzeaz eta analizatzeaz arduratzen dira. Horrela, azaltzen diren konplementuak, forma inpertsonalak, erreflexibotasuna, etab. atzeman ahal izango dituzte. Hauen lana Brent-en eta Manning-en lanetan oinarrituta dago.

(Ushioda eta beste, 1993)-n azpikategorizazio-ereduak eta hauen maiztasuna ateratzeko metodoa landuko dute. Hauen ustez, parsing-erako erabilgarria izan dadin maiztasunaren berri eman behar baita azpikategorizazio-eredu bakoitzerako. Horretarako aurrez etiketaturiko corpus baten gainean egingo dute lan anbiguotasun arazoak ekiditeko. Kategoria gramatikalez etiketaturiko corpusa, eta egoera finituko izen-sintagma analizatzaile baten bidez sei konplementazio-patroiren ezagupena eta maiztasuna lortuko dituzte. Corpus honen gainean parsing partzialeko teknika erabiltzen dute, hau da, ez dute analisi osorik egiten, baizik eta izen-sintagma ezagutzaile batean oinarritzen dira. %83ko doitasuna dute konplementazio-patroi hauek testu batean zehazterakoan, eta gertatzen diren errore gehienak izen-sintagma mugak oker markatzetik eratortzen direla diote.

(Briscoe & Carroll 1994)-k, argumentu-egitura corpusetik eskuratu nahi dute sarrera lexikalak garatu ahal izateko. Nagusiki balentziari erreparatzen diote. Hurbilpen honetan, corpuseko esaldien analisi azalekoa baina osoa erdiesten dute. Corpus horretan kategoria gramatikalei dagozkien etiketez gain, puntuazio-markak ere izango dituzte etiketatzaile baten bidez desanbiguatuak. Orduan, predikatu baten inguruan dauden azaleko analisi guztien artetik adierazgarri direnak eskuratuko ditu sistema honek. Kasuan kasuko predikaturako eskuraturiko patroiak ebaluatu egiten dituzte, patroi horren zuzentasuna neurtzeko garaturiko heuristiko eta teknika estatistikoen bidez. Horien ondoren predikatu horri dagokion sarrera lexikala eraiki ahal izango da. Sistema hau 160 azpikategorizazio-mota bereizteko gai dela diote. Azpikategorizazio-mota horiek ANLT eta COMLEX sintaxia hiztegietan azaltzen diren azpikategorizazio-motak biltzen dituzte.

(Eckle eta Heid, 96)-k azpikategorizazio sintaktikoa eskuratzen dute egunkariez osaturiko corpusetik. Proiektu honen helburua da laguntza eskaintzea eta ahal duen heinean azpikategorizazioari buruzko informazioa jasotzen duen lexikoiaren eraikuntza automatizatzea LNPrako. Aditz-zerrendak eta hauen erabilerari buruzko adibide-sortak ateratzen dituzte corpusetik. Zerrenda horiek azpikategorizazio-moten arabera sailkatu, eta eskuz konprobatzen dituzte egon daitezkeen erroreak zuzentzeko. Hau egin ondoren, lexikoiko proto-sarrerak sortuko dituzte automatikoki. Proiektu honekin alemanerarako dagoen hutsunea bete nahi dute; COMLEX-ek ingeleserako jasotzen du azpikategorizazioari buruzko informazioa, eta Eckle eta Heid-ek ere horren pareko hiztegi elektronikoa eraikitzeko asmoa dute.

Semantika lexikalak ere izan du eragina argumentu-egitura corpusetatik eskuratzeko orduan. (Pugeault eta beste, 94)-n helburu orokor gisa ezagutza lexikala testuetatik eskuratzeko sistemen hobekuntza planteatzen dute datu lexiko-semantiko xeheak erabiliz. Primitibo semantiko batzuk definitzen dituzte, euren ustez, aditzen semantikak baldintzatzen baitu aditzen portaera sintaktikoa, hau da, argumentu-egituran argumentuak nola gauzatzen diren sintaktikoki. Primitibo horiek *Lexical Conceptual Structure*-n (LCS) ere erabiliak dira. Zenbait proiektutako testuetan (*Research Projects Descriptions,* RPD) aurkituriko aditzak aztertuko dituzte. Klase jakin bateko aditzen rol tematikoen distribuzioa aurreikus daiteke beraien semantika kontuan hartuz gero. Thesaurus batez baliatuz, testutik interesgarri diren predikatuak eta argumentuak eskuratuko dituzte. Eta predikatu eta argumentuen arteko erlazioa rol tematiko baten bidez errepresentatuko dute. Sintaxia eta errepresentazio kontzeptualaren zubia dela diote. Rol tematikoen esleipenerako erregelak abian jarri aurretik, parserrak aditz-egitura oinarrizkoak ezagutzen ditu.

(Utsuro & Matsumoto, 97)-n japonierarako azpikategorizazio-lehentasunak ikasten dituzte estatistikoki EDRko corpus etiketatutik, eta ikasteaz gain doitasuna ere neurtzen dute. Aditzetarako eskuratzen duten informazio lexiko-semantikoa analisi sintaktikorako parserren hobekuntzan aplikatzea dute helburu. Aditzen kokakidetza lexiko-semantikoa ikasterakoan bi

anbiguotasun hauei aurre egin behar zaiela diote: 1) kasu-dependentziena eta 2) izen-mota orokortzearena. Bi gai hauetan ikertzaile askok dihardute aditzen agerkidetza lexiko-semantikoa ikasten, eta emaitzen neurketa desanbiguatze sintaktikoari erreparatuz egingo dute. (Resnik, 1993) eta (Li eta Abe, 1995)-n aztertzen dute nola aurkitu argumentu den izen baten abstrakzio maila egokia zuhaitz-egiturako thesaurus batean.

(Alonge, 1992)-n, definizioen azterketa egingo dute aditz-moten semantika eskuratzeko. Informazio semantiko hau baliatuz aditzen propietate sintaktikoak hiztegietatik erauztera joko dute. Definizioetatik ateratako osagaien arabera aditzen propietate sintaktikoak erauzten dituzte.

(Langer eta beste, 96)-n hiztegi elektroniko bateko informazio semantikoa eta izen-sintagmen gramatika bat erabiltzen dute, corpusetatik azpikategorizazioari buruzko informazioa erdiautomatikoki eskuratzeko alemanerarako.

(Delisle eta Szpakowicz, 97)-n, testuetatik eskuratu nahi dute argumentu-egitura. Horretarako TANKA deituriko sistema garatzen dute. Sistema honen osagaiak honako hauek dira:

- DIPETT (*Domain-Independent Parser of English Technical Texts*) analizatzaile sintaktikoa. Analisiaren emaitzaren zuhaitza bere azaleko heuristikoetan oinarrituz.
- HAIKU moduluak sintaktikoki markaturiko erlazio semantikoak atzematen ditu hiru mailatan: klausula mailako erlazioak (Barker & Szpakowicz 95); kasuak (Delisle 94; Delisle eta beste 96) eta izenen eta modifikadoreen arteko harremanak (Barker, 97).

Kasu-erlazioak funtsezkoak dira predikatuen argumentu-egitura eraikitzeko, hauek klausulako aditz nagusiaren eta honen argumentu sintaktikoen (*subj.*, *obj.*, *PPs*, *adverbials*) harreman semantikoa errepresentatzen dute eta. Kasuek adierazten dituzten harremanak maiz aditzaren rol tematikoekin bat datoz.

Testuetan kasuak predikatuen argumentu-egitura gisa azaltzen dira, kasu bakoitza markatzaile sintaktiko baten bidez adierazten da, eta partikula zehatz batek betetzen du. Esate baterako, *The thief broke the window with the stone* adibidean, *with* markatzailea da eta *stone* ekintzaren instrumentua (*instrument*).

DIPETT eta HAIKU (kasu-analizatzailea) moduluei esker TANKA sistemak predikatuen argumentu-egitura lortzen du zuzenean testuetatik. Hori bai, azpimarratu behar da, hau guztia erabiltzailearen begiradapean egiten dela. HAIKU-k DIPETT-ek sorturiko egitura sintaktikoa hartuko du, eta esaldi bakoitzerako aditzaren esanahia ongien errepresentatzen duen kasu-patroia aukeratzen du. Kasuak adierazteko kasu-markadoreak erabiliko dira bi modutara. Markadore lexikala eta posizioari dagokion markadorea (*positional marker*) zeina azaleko maila sintaktikoko rola den.

Esaldi bakoitzeko analisi sintaktiko bat erdiesten dute DIPETT-i esker, eta anbiguotasunak erabiltzaileak ere konpon ditzake. Ondoren, kasu-markadore patroi bat (adib. *PSUBJ-POBJ-at*) sortuko du HAIKU-k eta, era berean, azkenik kasu-patroi bat (adib. *agt-obj-lto* (*location to*)).

Sistema erdiautomatikoa da, urrats desberdinetan erabiltzaileak hartzen baitu parte. Adibidez, semantikoki anbiguoa den esaldi batek kasu-patroi bat baino gehiago izan ditzake, eta erabiltzaileak horietako bat aukeratuko du. Kasu-markadore patroiak eta kasu-patroiak hiztegian joango dira gordetzen. Kasu-markadore patroien hiztegian, kasu-markadore patroien sarrerak izango ditugu. Sarrera hauek kasu-patroi zerrenda izango dute dagokion kasu-markadore patroiarekin loturik. Eta kasu-patroien kopurua ere adieraziko da. Kasu-patroi bakoitzeko adibide esaldi bat dago, erabiltzaileak modu elkarreragilean erabiltzeko.

Kasu-patroien hiztegian aditzak eta hauei dagozkien kasu-patroiak jasoko dira. Adibidez, *agt-obj* kasu-patroia hiru aditzi lotua azaltzen da: *delete*, *know* eta *print*.

Hiztegietan gordetako informazioari esker, HAIKU-k automatikoki eraikitzen ditu aditzentzako argumentu-egiturak. Aditz bakoitzerako zerrendako elementu orok kasu-markadore patroia, kasu-patroia eta esaldi-patroia errepresentatuko ditu.

Bestelako ikuspegiak ere badira, adibidez (Gomez, 1995)-n deskribatzen da nola ikasketarako algoritmoak ezagutza sintaktikoa eskura dezakeen parser bati erabiltzaileek, zeinek ez duten ingeleserako parser sintaktikoaren berririk, emaniko adibideen bidez. Horrela bada, erabiltzaile hauek emandako adibideen bidez erator daiteke aditz, izen eta adjektiboen azpikategorizazioa.

(Chanod eta beste, 1991) lanean argumentu-egitura jaso ahal izateko analisi sintaktikoa burutu ondoren, posprozesatzaile bat garatzen dute. Lehen urrats batean esaldien osagai oinarrizkoak ezagutzen dira eta ondorengo urratserako beharrezkoak izan daitezkeen datuak gordeko dira. Analisirako gramatikek egitura sintaktiko kontsistenteak ezagutuko dituzte, nahiz eta semantikoki zuzenak ez izan. Horrez gain, esaldi bakoitzerako ahalik eta analisi gutxien jasoko dituzte. Hau lortzeko *relaxed parsing* (Jensen, 1988) estrategia darabilte. *Attachment* eta asignazio anbiguotasunak kodetu egingo dira, eta tarteko modulu batek hartuko du ardura horiek konpontzeko. Horretarako, informazio semantikoa baliatuko du *online* iturrietatik, adib. hiztegietatik. Tarteko modulu honek lehenbiziko forma logikoa sortzen du, eta, kontrol funtzionala, urruneko mendekotasunak, aktibo/pasibo erlazioak, adjuntu eta konplementuen identifikazioa, etab. tratatuko ditu. Argumentu-egiturak osoak bezala hartuko dira sakoneko informazio funtzionala esleitzen zaienean. Sakoneko informazio funtzional hori, honelako etiketen bidez adierazten dute: sakoneko subjektua (*DSUBJ*), sakoneko objektua (*DOBJ*), etab.

Eta gainera, azaleko etiketak esleituko zaizkie adjuntu eta modifikadoreei orokorrean. Posprozesatzaileak analisi sintaktikoan gorderiko informazioa aztertzen du, eta, nodoak duen motaren arabera prozedura batzuk jarriko ditu martxan. Posprozesatzaileak esleituko ditu predikatuak, argumentuak, eta modifikadoreak. Eta beharrezkoa denean, zenbait egitura sintaktiko berresleituko ditu berriki kalkulaturiko funtzio sintaktiko sakonen arabera.

Proiektuen artetik, duen garrantziagatik eta Europako hizkuntza desberdinetarako lantzeko helburua dutela kontuan harturik, SPARKLE[1] aipatuko dugu. Proiektu honen helburuetariko bat edozein testutik ezagutza lexikala eskuratzeko sistemak garatzea da. Sistema horrek gai izan behar du azpikategorizazioa, argumentu-egitura eta hautapen-murriztapenak eskuratzeko testuetatik. Ezagutza lexikalaren eskurapenean aurrera egiteko azaleko sintaxiaz baliatzen dira testutik eskuratzen den informazioaren egokitasuna eta doitasuna hobetzeko. Zehatzago izanik, proiektu honen lehen urratsean garatutako azaleko parserraren irteeraren gainean lan egiten dute, predikatuen klase semantikoak, azpikategorizazioa, argumentu-egitura, hautapen-murriztapenak eta alternantziak eskuratu ahal izateko. Azaleko sintaxiaren alorra, ingeleserako, alemanerarako, frantseserako eta italierarako landu dutenez, intereseko hizkuntzaren arabera parametrizagarria izango da eskuratze lexikalerako garatzen duten sistema.

Bukatzeko, IXA taldeko beste partaide batzuen lana (Aldezabal eta beste, 1998) aipatuko dugu. Lan honek Briscoe eta Carroll-en ildotik jarraitzen du, hau da, zenbait tresna konbinatzen ditu sistema eraikitzeko: Euslem lematizatzailea (Aduriz eta beste, 1996); euskararako Murriztapen-gramatika (Aduriz eta beste, 1997), baterakuntzan oinarritutako euskararako gramatika PATR II formalismoaz baliatuz. Tresna hauen integrazioak osatzen du etiketatu gabeko corpusetatik azpikategorizazio-ereduak landu ahal izateko muina. Horrezaz gain, oso interesgarria deritzot integrazio horri bezainbeste, edota are gehiago, azpiesaldiak ezagutzeko duten ahalmenari eta hauek atzemateko heuristikoei. Azpiesaldien ezagutzaile horren zeregina esaldi konplexuetan (koordinazioa, menderakuntza dela eta, etab.) aztergai duten aditza eta honi dagozkion konplementuak zedarritzea da. Azpiesaldi hauetan aplikatuko dituzte analisi partzialeko teknikak. Tresna hau eskuz landutako zenbait aditzen azpikategorizazioaren datu-basea egiaztatzeko eta osatzeko erabilia da.

## IV.2 Zer ulertzen dugun oinarrizko argumentu-egituraz.

Aditzen adibideak aztertzeko arrazoiak aipatzerakoan —§ I.1 Motibazioa—, esan dugu azpikategorizazioak garrantzi handia duela LNPn. Eta LNPko aplikazioetarako lexikoietan informazio hau osatzeko premia larriaz gain, azpikategorizazio-egitura zehaztearen konplexutasunaz eta gai honi heltzeko dauden hurbilpen desberdinez jabetzen gara. Horren

[1] http://www.ilc.pi.cnr/sparkle.html

lekuko, EAGLES-ekoek *Preliminary Recomendations on Subcategorization* txostenean[1] eginiko lana aipa daiteke. Txosten horretan, zenbait teoria linguistikotan, LNPko lexikoietan, hiztegietan eta etiketaturiko corpusetan azpikategorizazioari dagozkion ezaugarri espezifikoen konparazioa egiten dute. Esate baterako, LNPko zenbait lexikoi konparatzerakoan honako ezaugarri hauei erreparatzen diete:

- lexikoiak oinarri teoriko espliziturik duen
- errepresentazio lexikalak zenbat deskribapen linguistiko maila dituen
- sarrera lexikalak duen modeloa
- sarreren identifikaziorako eta desberdinketarako irizpideak (adiera, kategoria, azpikategorizazio-patroia, etab.)
- argumentu kopurua
- argumentuen kategoria sintaktikoa
- argumentuen rol tematikoa
- kontrol eta igoera (*raising*) fenomenoa
- hautapen lexikala
- murriztapen morfosintaktikoak
- frame-alternantzia[2] eta argumentuen aukerakotasuna
- azaleko eta barneko egitura

Helburu gisa azpikategorizazioarekiko hurbilpen desberdin horien estandarizazioa erdiesteko eskema bat definitzea planteatzen dute. Eta horrez gain, definituriko eskema horrek Europako hizkuntza desberdinetarako izan dezakeen baliagarritasuna probatu nahi dute.

Esate baterako, LE-PAROLE proiektuaren baitan garatutako lexikoiak, informazio morfosintaktikoaren eta aditzen sintaxiaren adierazpenerako EAGLES-en gomendioetan oinarritu dira.

---

1 http://www.ilc.pi.cnr.it/EAGLES96/synlex/synlex.html.

2 Frame-ak sintaktikoki errepresentatzen ditu aditzarekin loturik dauden egitura sintaktikoak. Hau da, aditzak eskaturiko kontestu sintaktiko desberdinak eta portaera sintaktiko berarekin erlazionaturik daudenak, orokortu egiten dira frame-aren bidez.

Bestalde, azpikategorizazio-alternantzia edota diathesis gisa ulertzen da aditz baten argumentuak dituzten errealizazio (alternantzia) desberdinen multzoa. Hau da, aditz batekin konbinatzeko joera duten argumentuak agertzen diren azpikategorizazio-eskema desberdinak.

Jakina den bezala azpikategorizazioaren gaia oso zabala da, eta aurreko puntuan —ikus § IV.1— azpikategorizazioari dagokion informazioa eskuratzeko deskribatu ditugun hurbilpenek ere garbi erakusten dute halaxe dela.

Beraz, puntu honetan gure lanean zer hartuko den argumentu-egituratzat adieraziko dugu. Nahiz eta lehenagotik ere mugatua dugun aztergaia, hain zuzen ere lehenengo kapituluan motibazioa eta helburuak aurkezterakoan. Gure kasuan, esan bezala, aditzen[1] adibideetatik ahal den heinean oinarrizko argumentu-egitura eskuratu nahi dugu eta, ondoren, azterketa sakonagoak egiteko materiala eta metodologia landu.

Metodologia horren garapenerako tresna nagusietatik —ikus § I.3— EDBLn aditzen azpikategorizazioari dagokionean zer informazio lantzeko aukera dagoen deskribatzeari lotuko gatzaizkio. Informazio hori adierazteko Pin-Ngern-ek (1990) bere *Case-Frame* erlazioan egiten duen moduan, aditz baten argumentu bakoitzari buruz honako informazioa bil liteke *Azpikategorizazio_Eredua* deituriko eremuan: argumentuak perpausean betetzen duen funtzio sintaktikoa, berorri dagokion kasua (semantikoa), hautazko edo derrigortasuna, eta hautapen-murriztapenak, hots, argumentuaren "betetzaileak" izan behar duen ezaugarri semantikoa. Esate baterako, ingelesezko *open* aditzaren kasuan honako eredu hauek ematen ditu:

*(open, 2, 1, Subj, Agent, Obligatory, Animate)*

*(open, 2, 1, DObj, Patient, Obligatory, Something)*

*(open, 1, 1, Subj, Patient, Obligatory, Something)*

Antzeko zerbait egin daiteke aditzak perpaus-osagairik hartzen duen ala ez eta perpaus horri buruzko beharrezko den deskribapena egiteko ere.

Informazio hori lantzeko aukera baldin badago ere, oraindik eremu horiek betetzeke daude. Eremu horiek betetzeko tresnak garapenean daude, baina oraindik ezin dugu emaitzarik erdietsi. Hau da, esate baterako ontologia bat izatea oso erabilgarria litzateke argumentu baten tasun semantikoak adierazi ahal izateko. Baina, euskarazko ontologiarik ez dugu oraindik, nahiz eta hori erdiesteko lanak garatzen ari diren IXA taldean. Bestalde, maila sintaktikoan ere tresnak garapen mailan daude, lehenbiziko emaitzak eskuratzen ari garela.

Beraz, ditugun baliabideak kontuan izanik, momentuz gure lanaren bidez azaleko oinarrizko argumentu-egitura erdiesteko bidea jorratu dugu, beti ere, gogoan izanik lortzen den informazioa azterketa sakonagoetarako baliagarria izan daitekeela, eta metodologiaren baitan baliatzen diren tresnak sendotzen diren heinean emaitza osoagoak lor daitezkeela.

---

[1] Aditzez gain, izenek, adjektiboek eta preposizioek argumentu-egitura dute, baina gu aditzen argumentu-egituraz arduratuko gara.

Jo dezagun, bada, oinarrizko argumentu-egitura horren zehaztapena egitera. Gure azterketaren ikuspegitik, oinarrizko argumentu-egitura zehazteak esan nahi du, ditugun baliabideak kontuan izanik, argumentu-egiturari buruz zein informazio jasotzeari erreparatuko diogun. Hau da, kontzeptu zabal honen oinarriak finkatzea ez da gure helburua, baizik eta gure azterketaren lehentasunak zehaztea.

Hala ere, lehendabizi, zehaztapen hori egin aurretik argumentu-egitura nozioaren inguruko kontzeptu batzuk aipatu nahi genituzke (Saint-Dizier & Viegas, 95) autoreen azalpenei jarraituz. Autore horiek (Grimshaw, 1990)-n dakarrena hartzen dute abiapunutzat, eta horren arabera predikatu bakoitzari argumentu-egitura bat dagokio eta argumentu-egitura horrek predikatuak behar dituen argumentuen kopurua zehazten du. Hala ere, aditzen kasuan argumentu-egitura bat baino gehiago egon daitezke lotuak aditz bakoitzarekin. Argumentuek predikatuak deskribatzen duen ekintzan edo egoeran beharrezkoak diren elementuak errepresentatzen dituzte. Adjuntuak, aldiz, ez dira argumentu-egituran sartzen, nahiz eta ekintza edota egoeraren deskribapenean lagundu. Hala ere, argumentuak hautazkoak dira eta maiz ez dira gauzatuko esaldi batean.

Bi oinarrizko teoria daude predikatu eta bere argumentuen arteko harremana deskribatzeko: bata teoria logikoa, esaten duena argumentuak ordenatuak daudela eta beren baitan esanahi zehatza dutela; eta rol tematikoen distribuzio-sistema, argumentuei rol tematikoen etiketak jartzen diena.

Bestalde, maiz argumentu-egitura azpikategorizazio-frame eta rol tematikoen distribuzioaren nozioekin lotzen bada ere, argumentu-egiturak berez ez du ezer esaten argumentuen funtzio sintaktikoei buruz edota rol tematikoei buruz. Argumentu-egiturak soilik oso modu abstraktuan predikatu baten aritatea zehazten du, hots, predikatuaren definizioan parte hartzen dutenen kopurua. Hurbilpen honi jarraituz gero, aske gaude sailkapen sintaktikoetatik (trantsitiboa, ditrantsitiboa, etab.), eta esan dezakegu aditz baten aritatea 2 dela. Aditz baten argumentu-egiturak, aditz hori agertzen deneko egitura sintaktikoetan definituko ditu, modu batean edo bestean, agertu beharreko argumentuak. Aldiz, aditz baten adjuntuek informazioa gehitzen diote predikatuari, hala nola, denbora, lekua, modua, etab.

Argumentuak argumentu-egituran predikatuarekiko duten harremanaren bidez desberdintzen dira. Honi esker, argumentu-motak desberdintzen ahal dira. Argumentu-mota horiek sintaxian era desberdinetan gauzatu ahal izango dira. Harreman horien propietateak ere, era desberdinetara adieraz daitezke, rol tematikoak edo konbinaturiko elementu primitiboak erabiliz.

Rol tematikoen gaiak oso harreman estua du *The Lexical Conceptual Structure* (LCS)-ekin (Jackendoff, 1988,1990), baina LCSk are sakonago espezifikatzen eta errepresentatzen ditu rol

horiek. LCS-k sintaxiaren eta errepresentazio semantikoaren arteko harreman sakona ezartzen du.

Gertaera-egitura ere aipatzen da argumentu-egituraren definizioan, beti ere oso loturik denbora eta aspektuaren nozioekin. (Vendler, 1967)-en arabera gertaera-egiturak gertaera-mota bakoitza eta dagokion hitza edo esaldia identifikatzen laguntzen du, esate baterako: *achievement*, *accomplishment*, *state*, etab.

Argumentu-egitura zer den mugatzean batasunik ez izan arren hizkuntzalarien artean, aipatu ditugun autore horiek egindako kontsiderazioak egokiak direla uste dugu.

Hala ere, esan beharra dago onartua dela orokorki, aditzen azterketa lexikala egiten delarik, kontuan hartu direla aditz bakoitzak adierazten duen prozesuan edo egoeran partaide diren argumentuen:

- eginkizun semantikoak: agentea, eragilea, esperimentatzailea, lokatiboa, gaia, pazientea;
- kategoria gramatikala: izen-sintagma, kasu-sintagma, perpausa, ...
- izan ditzakeen murriztapen semantikoak: (+/- biziduna), (+/- zenbakarria), ...
- marka morfologikoak: kasuak bereziki.

Badituzte elementu hauek elkarren artean harremanak, eta hipotesi franko zabaldua da gaur egun argumentuak, parte handi batean, eginkizun semantikoen arabera antolatzen direla maila sintaktikoan. (Grimshaw 1990) eta (Pustejovsky 1993)-n, beste batzuen artean, esaten dute aditzen argumentu-egitura informazio lexiko-semantikotik eratortzen dela.

Argumentuak edo, halaber, argumentu-egitura aipatzen dugunean, bada anbiguotasun bat:

- alde batetik, ikuspegi sintaktiko batean, gramatika-funtzioez ari gara eta orduan aditzek islatzen duten egitura dela-eta subjektu argumentua edo objektu argumentua bezalako deiturak erabiliko ditugu, aditzen osagarriei datxekien egitura sintaktikoa gogoan izanez; edo, molde berean, kasuak izan ditzakegu gogoan, NOR-argumentua eta NORK-argumentua bezalako esamoldeak baliatuz.
- beste aldetik, aditzeko partaideen eginkizun semantikoez ari gaitezke: argumentu agentiboa, lokatiboa, pazientiboa, etab.

Hitz batez, argumentu-egituraren nozioak bateratzen baititu partaideen eginkizun semantikoak eta hauen araberako egituratze sintaktikoa eta markadura morfologikoa; ikusmoldearen arabera batzuei edo besteei behatzen ahal zaie.

Gure kasuan aditzaren oinarrizko argumentu-egitura jaso nahi dugula diogunean, batez ere, ondorengo ezaugarri hauek zehazteari edota jasotzeari erreparatuko diegu:

- Kasua eta funtzio sintaktikoa. Jasotzen ditugun balizko argumentu horien tasunetatik, nagusiki, kasuari eta funtzio sintaktikoari erreparatzen diegu. Hau kontuan izanez gero, adibideak aztertuz azpikategorizazio sintaktikoari heltzen dioten proiektuen ildotik ari garela esan dezakegu, adibidez, SNIV-Project (Martin eta beste, 92). Hala ere, jakin badakigu semantikak ere izan dezakeela eraginik, eta horregatik nolabait erabilgarria izan daitekeelakoan, adibideak antolatzerakoan hiztegian duten adiera-ikurraren berri jasotzen dugu adibide-multzo bakoitzean. Uste dugu hitzen erabileren eta adieren arteko harremana lantzeko dagoen ikergaia dela.

- Argumentu horiek perpausa diren ala ez bereiztu. Perpausa direnen kasuan, mendeko jokatu eta ezjokatuen arteko desberdintasunak ere jasoaz.

Horrez gain, bai aztertzen ari garen aditzerako bai bestelako aditzetarako, honako informazio hau jasotzen dugu: aspektua, modua, pertsona, laguntzaile-mota eta aditzaren funtzio sintaktikoa.

Lehenago esan bezala, oinarrizko argumentu-egitura jasotzeko xedeaz gain, ondorengo azterketetarako baliagarriak izan daitezkeen tasunak eskuratzen, eta eskuratzeko bideak lantzeaz arduratu gara. Hau da, gure helburua, batez ere, oinarrizko argumentu-egitura zehazteko baliabideak lantzea da. Hain zuzen ere, galdeketa-sistemari esker informazioa jasotzeko modua erraztea eta jasotako tasun horien konbinaketa desberdinak aztertzea ahalbideratzen da. Horrela bada, aditzen ahalik eta multzo esanguratsuenak osatzerakoan balia ditzakegun tasun horien azterketa erraztu nahi da.

Jasotzen diren ezaugarrien zerrenda osoa eta hauen erabilpena galdeketa-sisteman, zehatzago azalduko ditugu § V.1.7 eta § V.1.8-an.

Galdeketa-sistemaren bidez adibideetako informazioa modu antolatu batean aurkeztu nahi da, hau da, aditzen sailkapena egingo da oinarrizko argumentu-egituraren ikuspegitik. Hala ere, ditugun baliabideak aplikatzerakoan ikusiko dugu zenbateraino garen kapaz oinarrizko argumentu-egituratzat hartzen ditugun ezaugarri horiek jasotzeko. Bestalde, jasotzen dugun informazioa oso zabala izanik, ikuspegi desberdinetatik azter daiteke azpikategorizazioa. Hori dela eta, gure lanaren mugak zehazteko jasoriko informazio horietako batzuen arabera egingo dugu hiztegiko aditzen sailkapena, gainontzeko sailkapen-moduak lantzeko bideak zabalik utziko ditugula. Galdeketa-sistemaren bidez aditz bakoitzaren inguruan zenbat sintagma agertzen diren galde daitekeen arren, aditzen argumentu-kopuruaren zehaztapena sailkapenerako irizpideetatik at geratu da, esaterako. Lehenago ikusi dugun bezala, kopurua

zehaztean aditzaren aritatea lantzen dugun, baina ezaugarri hori esanguratsua den arren azterketa honetatik kanpo geratu da.

Horrez gain, azpimarratu nahi genuke azkenean eskuratutako informazio guztia giza erabiltzailearen galbahetik pasako dela, automatikoki eskuratzen den informazio hori egiaztatzeko. Horrela, C eranskinean jasotzen den aditz bakarreko adibideen sailkapena, automatikoki lorturiko emaitzen eskuzko orrazketaren ondorio da.

Ondorengo puntuan, aipatu informazioak jasotzeko dugun baliabideetariko baten oinarrian dagoen formalismoa aurkeztuko dugu: Constraint Grammar (CG), honetan oinarritzen baita gure metodologiako parte garrantzitsua den Euskararako Murriztapen-gramatika (EUSMG). Formalismo hau dela medio, azaleko sintaxia lantzen dugu hiztegiko adibideez osaturiko corpus "berezi" horretatik, ahal den heinean, azaleko argumentu-egitura eskuratzeko.

Lehenago aipatu dugun SPARKLE proiektuan ere, azaleko sintaxiaren garrantzia azpimarratzen dute LNPrako lexikoiak corpusetik eskuratutako informazioarekin hornitzen joateko, besteak beste, argumentu-egiturari dagokion informazioaz.

# IV.3 Analisi sintaktikorako joera nagusiak.

Adibideak aztertzeko aukeratu dugun *Constraint Grammar* (CG), testu errealak analizatzeko helburu orokorreko formalismo sintaktikoa dugu. Ondoren, CG formalismoaren ezaugarri nagusiak aipatu baino lehen, analisi sintaktikorako dauden joera nagusien ikuspegi orokorra aurkeztuko dugu, aukeraturiko formalismoa bera hobeki ulertzearren.

## IV.3.1 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. Testu errealez ez dira gehiegi arduratzen. Lantzen dituzten errepresentazio-ereduak handinahikoak izaten dira. 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.

## IV.3.2 Probabilitatean oinarritutako teknikak.

Analizatzaile probabilistikoen ezaugarri nagusiak hauexek dira:

- Etiketaturiko corpus batetik automatikoki ateratzen diren probabilitateetan oinarrituko dira predikzioak. Sistema hauek oso gutxi (edo batere ez ) erabiliko dute eskuz kodeturiko ezagutza gramatikala interpretazio desberdinen artetik egokia zein den erabakitzeko orduan.

- Oso azaleko analisia egiten dute, gehienetan kategoria gramatikalen etiketak esleitzean datzana. Badira zenbait lan parser sintaktiko estokastikoen arloan ere, hala nola, (Atwell, 1987; Schabes, Roth eta Osborne, 1993; Bod, 1993). Dena den, hurbilpen estatistiko hutsaren bidez analisi sintaktikoa bidera daitekeenaren auzia gai irekia da.

- Etiketatzaile estatistiko hoberenen emaitza uniformea dela azpimarra daiteke. Azken hamarkadan badira gutxienez hamar etiketatzaile %95-96 bitarteko arrakasta-kopurua lortu dutenak eta horrez gain ez da egon %97 baino emaitza hobeagorik lortu duenik. Joera arrakastatsu hau hizkuntza desberdinetan mantendu ahal izan da, beste batzuen artean, suedieraz (Eriksoson, 1991) eta txineraz (Chang eta Cheng, 1993), bietan %96ko arrakasta lortu zuten.

  %95-97 bitarteko arrakasta-kopurua aurrerapauso handia izan arren, ordura arte zeuden etiketatzaile-sistemen emaitzak hobetuz, emaitza kopuru hau ez da nahikoa. Esate baterako, parser baten aurreprozesatzailerako, errore-kopuru honekin testu errealago bateko esaldi luzeek (20-30 hitzekoak) bataz beste analisirik gabeko hitz bat izango bailukete, analisi sakonagoetarako oso ondorio kaltegarriak ekarriz.

  Zer errore-mota egiten du etiketatzaile estatistiko batek? Gehienetan, testuinguruaren araberako probabilitateak oso testuinguru mugatua hartzen dute, usu bi edo hiru hitzekoa gehienez. Zenbait anbiguotasun ebazteko ez da nahikoa izango testuinguru hori, eta egitura osoago baten berri beharko litzateke, hots, esaldia oso-osorik hartu beharko litzateke kontuan. Hain zuzen ere, etiketatzaile estatistikoen punturik ahulena esaldi osoko ezagutza linguistikoa erabiltzeko ezintasuna dugu. Askotan, bi edo hiru hitzeko testuingurua erabiltzea ere arazo-iturri bilakatzen da etiketatzaile estatistikoentzat. Beren predikzio gehienak probabilitate lexikaletan oinarrituko dira, hots, zein probabilitate duen i kategoria gramatikalak j hitza emanik.

### IV.3.3 Probabilitateetan eta gramatikan oinarrituriko hurbilpenak konbinatzen dituztenak.

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).

### IV.3.4 Azaleko parsing-eko teknikak.

Oro har, azaleko sintaxia lantzen duten sistemek analisi morfologikoan eta desanbiguatzean egiten dute lan. Helburu nagusienetarikoa da ahal den egitura sintaktiko gehien inferitzea informazio morfologikotik. Azaleko analisiaren xede tipikoa sintagmak eta oinarrizko buru/modifikatzaile harremanak atzematea da. Horrez gain, azaleko analizatzaile askok duten ezaugarria da corpus handietan aplikatzen direla. Horietako batzuk estokastikoak dira, hau da, estatistikan oinarritzen dira, esate baterako Church (1988) edo IBM/Lancaster-ekoen analizatzailea. TOSCA ingeleserako parserra, Nijmegen-en sortua (Oostdijk, 1991), azaleko parserren artean kokatzen da baita ere, erabiltzen dituzten erregelak corpusaren ikerketa sakonean oinarritzen baitira. Azaleko parserren ezaugarri nagusiak Abney (1997:125)-n dakarrenari jarraituz hauexek dira:

> "Partial parsing techinques aim to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis"

I.4.2-n ikusi dugun legez, CG formalismoa azaleko sintaxiaren arloan kokatzen du Abney-k. Eta CG-rekin batera azaltzen dituen azaleko sintaxiko parser desberdinak edota teknikak honako hauek ditugu: *chunk*-ak makoen artean kodetzeko teknikak (Church, 1988), Ramshaw & Marcus (1995); *Fidditch-en* parserra; *Brill*, *Copsy*, eta *Supertags.*; *Finite-State Cascades.*; *Longest Match.*

Gure kasuan, kontuan izanik analizatzaile sintaktiko murriztailea, hots, *Constraint Grammar* formalismoa baliatu dugula gure lanerako, ondorengo puntuetan formalismo hau hautatzeko arrazoiak eta ezaugarri nagusiak deskribatuko ditugu.

Gain-gainetik bada ere, azaleko sintaxiari eskaini diogun atal honi bukaera emateko, esan beharra dago gaur egun zabalik dagoen ikerlerroa dela mota honetako analizatzaileak analizatzaile sintaktiko sakonagoekin integratzea. Horrela bada, azaleko parserrak erabiliak dira

aurreprozesatzaile gisa zenbait arazori aurre egiteko, eta modu horretara analizatzaile sakonagoen lana errazteko, esate baterako anbiguotasuna jaitsiz.

## IV.4 *Constraint Grammar* formalismoaren aurkezpen orokorra.

Aipatu bezala, azaleko sintaxiaren arloan kokatzen da Helsinkin garaturiko *Constraint Grammar* (CG) formalismoa. Azaleko analizatzaile sintaktiko murriztailea dugu CG. Ondorengo puntuetan formalismo honen ezaugarri nagusiak deskribatuko ditugu.

### IV.4.1 CGren ezaugarri nagusiak.

Hona hemen bere ezaugarri nagusiak:

- Egitura gramatikala etiketen bidez adierazten da, sintagma-egituren parentesizazioaren ordez. Etiketa horien bidez ezaugarri morfologikoak eta funtzio sintaktikoak adierazten dira. Errepresentazio gramatikal honek estrukturalki erabaki ezinezko zenbait desberdintasun inplizituki uzten ditu. Nahiago dute azaleko analisi egokia, xehetasun askoko baina anbiguoa den analisia baino.

- Analisi-deskribapenak linguistek idatziak dira. Indar handia egiten dute idatziriko erregela gramatikalak testu errealetan probatzen. Parsing-erako idatziriko gramatikak batez ere testu errealetan azaltzen diren fenomeno linguistikoez arduratzen dira. Linguistikoki "interesgarri" diren fenomenoen berri ematen da, eta maiztasun nahikoa duten testu errealetan.

- Deskribapen morfologiko eta lexikalek garrantzi handia dute. Sistema honetan, bi mailatako morfologian oinarritutako formalismoa erabiltzen dute deskribapen morfosintaktikorako. Honekin batera, lexikoia da osagairik inportanteena, esate baterako, ingeleserako 56.000 sarrera dituzte, informazio morfologikoa eta sintaktikoa egokituz sarrera orori.

- Parsing-erako gramatika, erregela linguistiko, partzial, independente eta ordenarik gabeez osaturiko multzoa da. Partzial, erregela gramatikalek fenomeno partzial baina egiazkoak adierazten dituztelako maila desberdinetan.

- Parsing-a bi urratsetan garatuko den prozesu bezala uler daiteke:

    1) Bilaketa-mekanismo lexikala: hitz-forma orori posible dituen deskribapen guztiak esleituko zaizkio aukera bezala. Pauso honetan ez da kontestu sintaktikoa adierazten eta, horregatik, anbiguotasun handia azalduko da.

Lexikoian azaltzen ez diren hitzen analisia gauzatzeko, probabilitateetan oinarritutako erregelak erabiliko dira, hots, heuristikoak.

2) Analisi sintaktikoa. Ezinezkoak diren interpretazio-aukerak saihestuko dira, ordena linealean eta etiketetan oinarritzen diren murriztapenen bidez.

Erabaki gabe geratzen diren anbiguotasunak ebazteko, estatistikaz baliatzen diren erregelak erabiliko dira. Dena den, analisi gehienak ezagutza linguistikoan sustraituak daude, estatistikak paper osagarria betetzen duela.

Mota honetako sistema batek aurrera egin nahi badu, ezinbestekoa du errepresentazio gramatikal adierazgarria eta parsing-a gidatuko duten printzipioen deskribapen zehatza, hots, anbiguotasun-mota bakoitza nola ebatziko den zehaztu behar da.

CG erabili izan deneko zenbait hizkuntza aipatzen hasiz gero, ingeleserako eginiko lana izan da batez ere oparoena, (Voutilainen, Heikkilä eta Anttila, 1992; Karlsson, Voutilainen, Heikkilä eta Anttila (eds.), 1994).

Formalismo hau Fred Karlssonek proposatu zuen 1990ean, gramatikan oinarritutako testu errealetarako parser legez. Lana batez ere ingeleserako burutu da, baina beste hizkuntza batzuetarako ere garatu da formalismo hau, esate baterako, suomiera, suediera, danimarkera eta euskara. Euskarari dagokionez Murriztapen-gramatika (MG) (Aduriz eta beste, 1997).

- Gramatika etiketen bidez adierazten da. Erlazio gramatikalak etiketa morfosintaktikoen bidez adierazten dira.

  Etiketatze honen bidez azaleko analisi funtzionala lortzen dugu. Adibidez, *FRASE* suomierarako deskribapenean oinarrituriko sisteman, esaldiko hitz bakoitzari funtzioa adierazten duen etiketa funtzional bat esleitzen zaio. Esaldi-egituretarako notazioarekin alderatuz gero ahulagoa da, baina kategoria gramatikala esleitzen duten etiketatzaileena baino aberatsagoa edo osoagoa. Erlazio sintaktikoak landu izan dira etiketen bidez CGren aurretik ere, adib. Teleman (1974). Horrez gain, bestelako parsing-sistemetan ere erabiliak izan dira, adib. *English Slot Grammar parser* (ESG), Mc Cord eta Temperley (1991)-n, eta funtzioa adierazten duten etiketez gain, mendekotasun erlazioak adierazteko mekanismo esplizitua ere badute.

- Analisi sintaktikoa.

  Analisi sintaktikoa jarraiko lau moduluez osatuta dago:

1) Desanbiguatzaile morfologikoa: interpretazio-multzo anbiguoetako irakurketak testuinguruaren arabera zilegi ez baldin badira ezabatu egingo dira. Horretarako gramatikalariak desanbiguatze-erregelak idatziko ditu.

2) Esaldi-mugen ezarpena: testuinguruan oinarrituriko desanbiguatze-erregelekin tartekaturik, esaldi-muga etiketa esleituko dio parserrak zenbait hitzi baldin eta hitz horiek esaldiko klausula baten hasieran edo bukaeran dauden. Horretarako klausula hasieran edo bukaeran egon daitezkeen hitz-formen detekziorako erregelak landuko dira.

3) Mapaketa (*mapping*) morfosintaktikoa. Zenbait kasutan, hitz bati ezin izango zaio dagokion funtzio sintaktikoa zuzenean esleitu. Hau da, datu-basean ez dago aurreikusia kategoriari erreparatuz gero funtzio hori ezartzea, ez bailitzateke txukuna izango. Honelakoetan, testuingurua hartu behar da kontuan, esate baterako, determinatzaile kategoria duen sarrera bati buru baten modifikadore izango dela jarriko zaio funtzio sintaktikoen eremuan. Gerta daitekeena da buru hori ez azaltzea, eta, orduan, determinatzaile horren funtzioa modifikadorearena izan beharrean, subjektu izatera pasatzen dela. Funtzio sintaktiko hori esleitzeko erabiltzen diren erregelak mapaketa sintaktikoaren atalean sartzen dira.

   Desanbiguatzaile morfologikoaren irteera mapaketa morfosintaktikorako erregelen sarrera izango da. Mapaketaren bidez interpretazio morfologiko bakoitzari etiketa funtzional guztien zerrenda jarriko zaio, hau da, aukera sintaktikoen zerrenda. Esleituriko funtzioak gramatika deskribatzaileetan oinarritzen dira, esate baterako, ingelesaren kasuan (Quirk, Greenbaum, Leech eta Svartvik, 1985).

4) Murriztapen sintaktikoak.

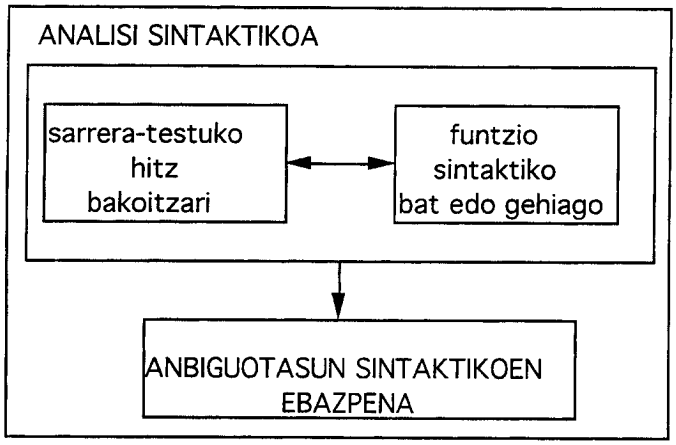
   Desanbiguazio morfologikoan erabiltzen diren murriztapenen antzekoak dira, desberdintasun bakarra litzateke funtzio-etiketak ezabatzen dituztela interpretazio morfologikoak ezabatu beharrean. Modulu hau ez dago desanbiguatze-erregelen modulua bezain landua. Fintzen eta garatzen ari dira etengabe. Erregela sintaktikoak aplikatu ondoren, hitzak %75-85 bitartean sintaktikoki desanbiguaturik gelditzuko dira.

CG jarraian aplikatuko diren azpigramatikez osatuta dago. Hauetako bakoitza ordenarik gabeko murriztapenez hornitzen da. Moduluetan banatze hau, deskribapen sintaktikoaren konplexutasunari aurre egiteko modu xinpleago bat hartu nahiak dakar, problema zatika aztertzean tratagarriagoa gertatuz.

Heuristikoetan oinarritutako murriztapenak aukerazkoak dira aplikatzerako orduan, eta konpontzeke geratzen diren anbiguotasunei aplikatuko zaizkie. Hauek aplikatu ondoren ere, gera daitezke desanbiguatu ezin izan diren hitzak.

## IV.4.2 Funtzio sintaktikoen esleipena.

Formalismoaren aurkezpen orokorra egiterakoan, ikusi dugu analisi sintaktikoa jarraiko lau moduluez osatuta dagoela. Funtzio sintaktikoen esleipenari dagokionean, hitz-forma orori funtzio sintaktiko bakarra esleitzean datza urrats honen eginkizuna. Hala ere, esan beharra dago funtzio sintaktikoen esleipenak bi urrats nagusi dituela. Lehenengoan, hitz-forma orori posible dituen funtzio sintaktiko guztiak esleitzen zaizkio. Ondorengoan, anbiguotasun sintaktikoen ebazpena burutuko da. Hona hemen analisi sintaktikoaren irudi eskematikoa.




**IV.1 Irudia**.-Analisi sintaktikoaren eskema.

Goiko eskematxo horretan erakusten ditugun asoziazio horien zehaztapenez osatuko da gramatika sintaktikoa. Batik bat, hitzen arteko aurrekaritza linealeko erlazioei buruzkoa izango da gramatika. Hitzetan oinarritutako errepresentazioa dugu, ez da eratzen arbolarik. CGren ezaugarri inportantea errepresentazioaren "azaltasuna" da. Azalera orientatutako analisi-gramatika. Adibidez: *Marik Jon etortzea nahi du* esaldian, perpasu nagusiaren ([Marik nahi du]) eta mendeko perpausaren ([Jon etortzea]) arteko erlaziorik ez da adierazi ahal izango. Horrelako erlazioak lan honen mugetatik kanpo daude. Murriztapenetan ez da mendekotasun-erlaziorik adierazten. Esaldi bat hitz-kate bat da. Horregatik, esaldien analisiaren emaitzan hitzak eta aurrekaritza-erlazioak ageriko zaizkigu.

Estrategia nagusia da, hitz bakoitzeko, eduki ditzakeen funtzio sintaktiko guztiak ematea. Teorikoki posible diren asoziazio guztiak zehazten dira (hitz-klaseak <---> funtzio

sintaktikoak). Horren ondoren, anbiguotasun funtzional izugarria suertatuko da. Nola konpondu? Bi bide hartuko dira anbiguotasun horri aurre egiteko:

1) Lan-espazioa zehaztu.

2) Anbiguotasuna gutxitu, testuinguruan oinarriturik. Horretarako, mapaketa (*mapping*) morfosintaktikoak eta murriztapen sintaktikoak baliatuko dira.

Hitz bakoitzak izan ditzakeen funtzio sintaktikoak lexikoiaren bidez adieraziko dira, edota mapaketa morfosintaktikoen bidez. Hain zuzen ere, mapaketa morfosintaktikoa deitzen den atalean erregela berezi batzuk izango ditugu zeinak kategoria bat funtzio sintaktiko batekin (edo gehiagorekin) harremanetan jarriko duten. Mapaketaren mekanismoa erabiliko da funtzio sintaktikoa zuzenean datu-base lexikalean adierazterik ez dagoenean. Funtzio sintaktikoak adierazteko etiketa sintaktikoak baliatuko dira. Etiketa bat baino gehiago agertzen denean hitz batean, orduan anbiguotasun sintaktikoaz hitz egingo da. Eta anbiguotasun sintaktikoa ebazteko, murriztapen sintaktikoak aplikatuko dira. Murriztapen sintaktikoen helburua hitz bakoitza etiketa sintaktiko bakarrarekin uztea izango da.

CG hizkuntza desberdinetarako garatzerakoan, kasuan kasuko hizkuntzarako, gramatikariek zehaztuko dituzte funtzio nagusiak eta modifikadoreak. CGko sintaxiaren bereizgarria funtzio-etiketadun dependentzia-sintaxia[1] (*functionally labelled dependency syntax*) da.

Aurrerago esan bezala, azaleko sintaxia da CGkoa. Morfologiak garrantzi handia du, baita ingelesa moduko hizkuntzetarako ere. Murriztapen sintaktikoak hitzen ezaugarri lexikaletan eta morfologikoetan oinarritzen dira. Eta azalekoa izateaz gain, laua da errepresentazio sintaktikoa. Esaldi baten deskribapen sintaktikoa etiketa sintaktikoez osatuta dago. Errepresentazio lauarekin zerikusia duen beste ezaugarri bat aditz-kateen nozioarena dugu. Aditz-lokuzioak kate lau bat izango dira egitura hierarkiko sakon bat adierazi beharrean. Horrela bada, adibidez ingelesez aditz jokatu/ezjokatu, eta nagusi/laguntzaileen artean bereizten duten etiketen bidez ingelesezko aditz-kate guztiak adieraz daitezke.

Corpusen anotazio sintaktikoan erabiltzen diren anotazio-eskema nagusien artean ere aurkeztua da CG. CG aipatua da Lancaster/IBM Scheme, Paris/IBM Scheme, TOSCA (*Tools for Syntactic Corpus Analysis*), UPenn eta SUSANNE (SUSANNE corpus, *Surface And Underlying Structural Analysis of Natural English*) eskemekin. Eskema hauek guztiek, sintagma-egiturako hurbilpena dute analisi sintaktikorako. CGk aldiz, dependentzia partzialeko hurbilpena du. CG gainera, erabat automatikoa den bakarra da.

---

[1] Azterketa gramatikaletan tradizio luzea du dependentzia sintaktikoaren kontzeptuak aro grekolatinotik. Hurbilagoko formalizazioak teoria sintaktikoan aipatzerakoan, Tésniere (1959) , Hays (1964), Mel'cuk (1988) eta Hudson (1990) aipa daitezke, besteak beste, dependentzia-sintaxiaren suspertzaile gisa arlo teorikoan.

Hiru ezaugarrik bereizten dute gainontzeko etiketatzaileetatik:

1) Edozein corpus etiketatzeko helburua du, ez da corpus partikularrak etiketatzeko.

2) Azaleko analisi sintaktikoa egiten du dependentzia-gramatiketan oinarritua. Gainontzekoek aldiz, konstituiente-egiturako analisia egiten dute.

3) Anbiguotasunak %3-7an ebatzi gabe utziko dira.

CGren oinarrian *dependency-oriented syntax* dago. Hau da, analisi sintaktikoa depedentzia analisia da, soilik hitzen arteko erlazioak kontuan hartzen diren zentzuan. Hitzak ez dira esplizituki beren gobernatzaileekin lotzen, baina etiketa sintaktikoek modu esanguratsuan murrizten dute analisi bateragarrien multzoa, eta ikus daitezke analisien anbiguotasun-klaseen errepresentazio gisa.

Bestalde, euskararako landutako funtzio sintaktikoen azalpenari heldu aurretik, talde finlandiarrekoek ingeleserako CG (Voutilainen eta beste, 1992) lantzerakoan harturiko zenbait erabaki aipagarriak iruditzen zaizkigu. Horien artetik, ebatzi ezin diren anbiguotasunak ez tratatzeko irizpidea nabarmenduko genuke. Esate baterako, *prepositional attachment* eta *attachment of adverbials* anbiguotasun-mota zailak direnez desanbiaguatzen, ez dituzte beren parsing-eskeman sartuko. Beraz, parserrak aukera handiagoak ditu analisi zuzenak eta ezanbiguoak sortzeko. Modu berean, ez dituzte bereizten ere jokaturiko eta ezjokaturiko klausulen subjektuak.

Hala eta guztiz ere, ohartzen gaituzte esanez beren errepresentazio gramatikalak ez daudela erabat aske egiturazko anbiguotasunetatik. Esate baterako, izen-sintagmek anbiguo izaten jarrai dezakete objektu eta konplementu preposizional funtzioengatik. Adibidez, *local societies* ondorengo esaldian (ikus oin-oharrean funtzio sintaktiko[1] horien esanahia): *They have established networks of state and local societies.*:

```
("<*they>"
  ("they" <NonMod> PRON PERS NOM PL3 (@SUBJ))
("<have>"
  ("have" <SVO> V PRES -SG3 VFIN (@+FAUXV)))
("<established>"
  ("establish" <SVO> PCP2 (@-FMAINV)))
("<networks>"
  ("network" N NOM PL (@OBJ)))
("<of>"
  ("of" PREP (@<NOM)))
("<state>"
  ("state" N NOM SG (@<P)))
```

[1] Adibidean azaltzen diren funtzio sintaktikoen esanahia honako hau da:

@SUBJ= subjektua; @OBJ= objektua; @<NOM= bestelako postmodifikatzailea; @AN>= adjektibo aurremodifikatzailea; @<P= preposizioaren bestelako konplementua; @CC= koordinatzailea; @+FAUXV= aditz laguntzaile jokatua; @-FMAINV: aditz nagusi ezjokatua.

```
("<and>"
  ("and" CC (@CC)))
("<local>"
  ("local" A ABS (@AN>)))
("<societies>"
  ("society" N NOM PL (@OBJ @<P)))
("<$.>")
```

Arau bezala planteatzen dute nahiago dutela honelako egiturazko anbiguotasunak esplizituki uztea, anbiguotasun hori konpontzearren kategoria berri bat *ad hoc* sartzea baino.

Bukatzeko, esan dezakegu aurkezturiko ezaugarrien arabera garbi dagoela formalismo hau *shallow syntax* edo *partial parsing* -aren —ikus § IV.3.4— alorrean kokatzen dela.

## IV.4.3 CG gramatika osatzen duten atalak eta horien idazkera.

Gramatika idazterakoan, exekutagarria izateko gramatikaren fitxategiak bost atal dauzka eta honako hauek dira: *SENTENCE-DELIMITERS*, *SET-DECLARATIONS*, *MORPHOSYNTACTIC-MAPPINGS*, *DISAMBIGUATION-CONSTRAINTS* eta *END*.

Parserrak, atal guztiak, hutsik bada ere, bertan daudela egiaztatzen du. Atalen zehaztapenak azalduko ditugu segidan.

### IV.4.3.1 Esaldien arteko muga-markatzaileak (*Sentence delimiters*).

Beharrezkoa den lehenengo atala da. Irteera esaldietan banatzen duen puntuazio-marka ziurren deklarazioaz osatua dago. Puntuekin batera, puntu eta koma, galdera-ikurra, etab. izango dira atal honetan definituko direnak.

### IV.4.3.2 Ezaugarri-multzoak (*Set declarations*).

Erregeletan erabiliko diren multzoak definitzen dira atal honetan.

### IV.4.3.3 Mapaketa morfosintaktikoak (*Morphosyntactic mappings*).

Mapaketen xede nagusia ezaugarri morfologiko bat funtzio sintaktiko batekin lotzea da. Mapaketak interpretazio morfologiko bati funtzio sintaktiko bat esleitzen dio. Datu-basetik ez datozen funtzio sintaktikoak esleitzeko erabiliko dira.

- mapaketa-erregelen formatua honako hau da:
  *eragilea (MAP), esleitu nahi den funtzio sintaktikoa, helburua, (IF), testuinguruko baldintzak.*
- eragilea: eragile bakarra dugu *MAP*.
- esleitu nahi den funtzio sintaktikoa: eragile horren ondoren helburu-interpretazioari esleitu nahi zaion funtzio sintaktikoa idatziko da.

- *target* hitza jarriko da helburua den ezaugarriaren aurretik. Eta, ondoren, esleitu nahi zaion interpretazioetan gerta daitezkeen ezaugarri morfologikoen lista.
- *IF*: testuinguruko baldintzen aurretik jartzen den hitza.
- testuinguruko baldintzak: testuinguruko baldintzak adierazten dira atal honetan. Baldintzak ezaugarriek osatzen dituzte.

Mapaketak sekuentzialki aplikatzen direla eta testuko hitzak ezkerretik eskuinera tratatzen direla kontuan hartuz, -1, -2, ... listetako zenbakiek ezkerreko testuinguruko hitz-formak adierazten dituzte. Kasu honetan uneko hitz-formatik hasita ezkerretan lehenengo eta bigarren dauden hitz-formak adierazten dituzte.

### IV.4.3.4 Desanbiguatzeko murriztapenak (*Disambiguation constraints*).

Desanbiguatzeko murriztapenak ezaugarri morfosintaktikoen desanbiguatzeaz arduratzen dira. Murriztapenek formatu hau jarraitzen dute:

*domeinua, eragileak, helburu-interpretazioa, IF, uneko hitzaren baldintzak, testuinguruko baldintzak*

- *domeinua:* desanbiguatu edo tratatu behar den elementua adierazten du. Hitz-forma konkretu bat adieraz daiteke ("<>" eta komatxoen artean idatziko da) edo aldagai baten bitartez edozein hitz-forma.
- eragileak:
  - *SELECT*: hau da, helburuan adierazten den ezaugarridun interpretazioa hautatuko da, baldin eta helburuaren baldintzak eta testuinguruarenak betetzen diren.
  - REMOVE: kontrako eragiketa burutuko da, ezabatu egingo da helburuan adierazten den ezaugarridun interpretazioa, baldin eta helburuaren baldintzak eta testuinguruarenak betetzen diren.
- helburu-interpretazioa: tratatu nahi den anbiguotasun-mota zehazten da ezaugarri morfologiko bat eragilearen ondoren jarriaz. Ezaugarri morfologiko horretan: kategoria, numeroa, etab. adieraz daitezke. Hori egin ahal izateko, ezaugarri hori erazagutua egon behar da gramatikaren multzoen atalean (*SET*).
- *IF*: hitza jartzen da testuinguruaren baldintzen atalaren aurretik.
- uneko hitzaren baldintzak: hitzak berak bete beharreko ezaugarrien multzoa. Horrela bada, baldintza horiek zerogarren posizioari dagozkionak izango dira.
- testuinguruko baldintzak: 0. posizioan dagoen hitz-formaren (uneko hitz-forma) arabera adierazten dira. Erregela aplikatu ahal izateko testuinguruan bete beharreko baldintzak adierazteko posizio positiboak (1, 2, 3,...) zein (-1,- 2, -3,...) negatiboak erabil daitezke. Zenbakien bidez adierazten da urrats batera, bitara, hirutara, etab. (eskuinetara zein ezkerretara) zer-nolako ezaugarriak bete behar diren. Testuinguru zabalagoa ere erabil daiteke, hainbat urratsetara eskuinera zein ezkerretara "*" ikurra jarriaz zenbakiaren

aurretik. Honekin adierazten da esaldian edozein posiziotan eskuinetara edo ezkerretara aurkitzen den hitz baten ezaugarriez ari garela.

Ohar orokor gisa, esan beharra dago baldintzak baiezkoak edo ezezkoak izan daitezkeela. Ezezkoak direnean baldintzaren posizioa adierazi aurretik *NOT* eragilea idatzi behar da. Eta, hartara, testuinguruan ezaugarri bat ez agertzea eskatzen da. Bestalde, baldintza horien aplikazio-eremua esaldia da, hau da, puntutik puntura dagoen hitz multzoa.

Dena dela, nahi izanez gero aplikazio-eremua muga daiteke. Horretarako *BARRIER* eragilea balia dezakegu:

- *BARRIER* posizioei dagokienean, bada eragile bat erregelaren eremua muga dezakeena: *BARRIER* eragilea. Eragile honen bidez, erregela batek adieraz dezakeen baldintza esaldi osoan kontuan hartu beharrean, interpretazio-helburutik *BARRIER* horrek mugatzen duen ezaugarriraino hartuko da. Bestalde baldintza absolutuez gain, hau da, baldintza helburuan adierazten den hitzarekin lotzeaz gain, baldintza erlatiboak adieraz daitezke. Hau da, posizio batean adierazten den baldintza beste posizio batekoarekin lot daiteke *LINK* eragilearen bidez.

  Posizioarekin zerikusirik duten beste puntu hauek ere kontuan hartu behar dira. *Aplikazio modu normala edo kontu handiko modua*. Desanbiguatzeko erregelek interpretazioak baztertzeko testuinguruko hitz-formen interpretazioak kontuan hartuz egiten dute nahiz eta testuinguruko hitz-forma horiek anbiguoak izan. Parserrak bi aukera uzten ditu: testuinguruko hitz-forma ez-anbiguoetan oinarritzea uneko hitz-forma desanbiguatzeko (modu *kontu handikoa* edo *ziurra*, C karakterea posizioaren atzetik idatziz adierazten da) edo testuinguruko hitz-formen interpretazioetan oinarritzea nahiz eta hauek anbiguoak izan (modu normala).

### IV.4.3.5 Erregela sintaktikoak (*Syntactic rules*).

Funtzio sintaktikoen desanbiguazioaz arduratzen dira, eta azalpenerako bereizi ditugun arren, esan beharra dago desanbiguatze-erregelekin batera azalduko direla *DISAMBIGUATION-CONSTRAINTS* atalean. Erregelen formatuaz desanbiguatze-erregelei buruz esandakoak balio du hauen kasuan ere. Desberdintasun bakarra da ezaugarri morfosintaktikoak desanbiguatzeaz arduratu beharrean, anbiguotasun sintaktikoak ebazteaz arduratzen direla.

## IV.4.4 Gramatikaren kudeaketa.

CGren arazoetariko bat erregelen kudeaketarena dugu. Arazo hau areagotu egiten da gramatika handitu ahala. Lehenengo parserrean (Karlsson, 90) erregela guztiak multzo berean zeuden, eta horrek erregelen kontrola zaildu egiten zuen. Esate baterako, honelako itaunak sortzen dira gramatikaren inguruan: nola aurkitu eta zuzendu murriztapen desegokiak? Bi murriztapen emanik, nola jakin kontraesanean daudenentz, edota batak bestea estaltzen duen (partez bada

ere)? Kontraesanean badaude, zer esan nahi du horrek? Ba al da modu orokorrik gramatikaren koherentzia egiaztatzeko? Nola kendu erredundantzia?

Galdera horien guztien pean ezkutatzen den arazoa murriztapenen aplikazio-ordenarena dugu. Murriztapenak zein ordenaren arabera aplikatzen diren, emaitzak era batekoak nahiz bestekoak izan daitezke eta. CG-2 parserrak (Tapanainen, 96) hain zuzen ere, erregelak multzokatzeko aukera eskaintzen du. Horrela, aplikazio-ordenarena hein batean bederen kontrola daiteke.

Esate baterako, gure kasuan, erregelak ziurtasunaren arabera multzokatu ditugu, lehenengo multzoan ziurrenak jarriz.

## IV.4.5 *English Constraint Grammar* (ENGCG) eta *Constraint Grammar* formalismoaren ebaluaketa.

ENGCGk, sistema probabilistikoekin konparatuz gero, errore-tasa txikiagoa du. Arrakasta honen zergatiak aipatzerakoan, honako hauek azpimarratzen dira:

- Gramatikaren errepresentazio-ereduak gramatikagilea konpondu ezin diren anbiguotasunetatik aske uzten du.
- Errepresentazio gramatikala oso ondo espezifikatzen da, deskribapenaren ebaluaketa eta jarraipena erraztuz.
- CGren arabera, murriztapen bakoitza independentea da. Horrela bada, murriztapen bakoitzaren ondorioak banan-banan konprobatu ahal izango dira. Jakina, independente izate horrek erraztu egingo du lana, eta aldaketa zein erregela berri bat idazterakoan ez dugu besteengan izan dezakeen eraginetan pentsatu beharrik.
- Desanbiguaziorako gramatika eraikitzeko corpusaren erabilera oparoa egin dute (3 milioi hitz, testu errealetatik atereak). Corpus hau eskuz desanbiguatu zen lehendabizi.
- CG formalismoak ondoko hitzei erreferentzia egiteaz gain, esaldian dagoen edozein hitzi erreferentzia egiteko ahalmena du. Erreferentzia egin dakioke hitz baten posizio finkoari, adib. ezkerretik hasi eta hirugarren posizioan dagoen hitza. Murriztapenetan kontuan hartzen den testuingurua hitz batekoa edo bikoa izatea ez da nahikoa anbiguotasun-mota oro ebazteko. Hori dela eta, urruneko testuingurua ere kontuan har daiteke, esate baterako, esan daiteke ezkerrean dagoen hitzen bat determinantea dela baldin eta, determinantea eta aztertzen ari garen hitzaren artean izenik aurkitzen ez bada.

Hala ere, (Voutilainen, 1994b)-n dakarrenari jarraituz ENCG-k baditu hobetu beharreko zenbait ezaugarri:

- Klausula-mugen ezarpenerako mekanismoa.
- *Constraint Grammar* sekuentziala izatea.
- Anbiguotasunaren errepresentaziorako hitzean oinarritzea.
- Sintagmei erreferentzia egiteko desegokia izatea.
- CG formalismoaren adierazgarritasun falta: aurrerago aipatu ditugun ezaugarriengatik, gramatikalariak zailtasunak izango ditu gramatika erabiltzeko.

Arazo horiei guztiei aurre egiteko *Finite-State Intersection Grammar* garatu zuten (Voutilainen eta Tapanainen, 1993). *Finite-State Intersection Grammar*-en ingurunean lortzen den deskribapen sintaktikoa CG-rekin lortzen denaren antzekoa da, baina ez da hain azalekoa. Ingurune horretan eginiko hobekuntza nagusien artean, analisirako unitatea esaldia hartzea dugu.

## IV.5 Dependentzia-Gramatikan oinarritutako parserra (*Dependency Grammar Parser*).

CG formalismoaren ezaugarri nagusiak labur-labur aurkeztu ondoren, formalismo horren ahaide gisa har daitekeen parser horren inguruko ezaugarri orokorrak aurkeztuko ditugu. Formalismo honi hurbiltzeko arrazoiak bi ditugu: oinarrian CGren bidetik jarraitzen duela, eta CG landu dutenek bide honetatik jo dutela sintaxia aurrera egin ahala. Hona hemen, (Jarvinen eta Tapanainen, 1997:1)-n jasotzen diren hitzak, aipatu berri dugunaren inguruan:

> "The new dependency parser belongs to a continuous effort to apply rule-based methods to natural languages. It can been seen as a relative of the Constraint Grammar framework (Karlsson et al., 1995), since many features of the system have been derived from it. The syntactic description (Järvinen, 1994) in the English Constraint Grammar (ENGCG) is implicitly dependency oriented; it contains tags for heads and modifiers but not explicit links between them."

*A Dependency Parser of English* (Järvinen eta Tapanainen, 1997) lanean adierazten den bezala, parser horren ideia nagusia da erlazio sintaktikoak esplizitu egitea, hau da, dependentzia funtzionalak loturen (*link*) bidez adieraztea. Horretarako oinarri teoriko gisa Tesnière-ren dependentzia-gramatika (Tesnière, 1959) hartzen dute. Hona hemen, (Jarvinen eta Tapanainen, 1997)-n aipatu lanean azaltzen diren ezaugarri nagusiak:

- oinarrizko elementu sintaktikoa ez da hitza, gunea baizik.
- elementu bakoitzak gune bakar bat du.

- emaitza zuhaitz bat izango da.
- dependentzia funtzionalak *link*-en bidez, hots, loturen bidez adierazten dira.
- modifikadoreak ez dira derrigorrezkoak.
- edozein input analizatuko da, nahiz eta zuzena ez izan.
- deskribapena laua da, maila sintaktiko bakarra izango da.

## IV.6 Sintagmak atzemateko zenbait tresna.

Puntu honetan sintagmak osatzeko gai diren zenbait tresnaren berri emango dugu. Lehenbiziko biak, CG analisietatik abiatzen direnak dira, eta Helsinkiko taldekoek garatuak. Azkenik, Tacat ere aipatzen dugu, azken hau corpus etiketatuen parentesizatzaile bat da.

NPtool ingelesezko modifikadore/buru egiturak atzemateko programa da. Horretarako ENGCG desanbiguatzaile morfologikoa, eta beste zenbait gramatika eta parser Atro Voutilainen-ek eta Pasi Tapanainen-ek eginak baliatuko ditu. Egitura horiek ezagutzeaz gain, NPtool izen-sintagmak atzemateko sistema dugu. Eta sistema honen muina CGren filosofian oinarritzen da. Analisi-estrategia berdintsua du CGrekiko, baina xedea desberdina da. NPtool-ek izen-sintagmak ezagutu eta analizatu nahi baititu. Horrez gain, egitura konplexuagoak ere trata ditzake, esate baterako posmodifikadore egiturak (*postmodifying constructions*) eta aditz/objektu egitura sinpleak (*simple verb-object constructions*). Helburuan bereizteaz gain, sintaxirako gramatika-eskema hobetu du sistema honek. Deskribapen sintaktikoari gagozkiola, esan beharra dago anbiguotasun sintaktiko guztiak lexikoiaren bidez sartuko direla zuzen-zuzenean, beraz, ez dago gainerako modulu baten beharrik eginkizun honetarako.

NPtool-ek egoera finituzko parserra (Voutilainen, 93) baliatuko du onartezinak diren analisiak baztertzeko. Egoera finituzko parser honek gramatikarekin bat datozen esaldi-interpretazio guztiak sortzen ditu.

Kontuan hartu behar da bakarrik hitz-mailako anbiguotasunak tratatzen dituela. Orduan, geratzen diren anbiguotasunei aurre egiteko mekanismoa du. Mekanismo honi esker, gai da anbiguotasunik gabe analizatu diren egiturak, zeinak bat datozen gramatikako egiturekin, eta gramatikako egiturekin bat etor daitezkeen analisien artean bereizteko. Anbiguotasunik gabe analizatu direnei *OK* etiketa jartzen zaie, eta balizko analisiek *?* etiketa izango dute.

Analizaturiko testu batetik atera ditzakeen egiturak modu honetakoak dira: hainbat adjektibo aurremodifikadore koordinaturik egon daitezkeenak, edota izenak hauen ondorik buru nominal bat eta ondoren hautazko ezanbiguoa den PP posmodifikadore bat (zeinak izango dituen preposizioa eta koordinaturik egon daitezkeen hainbat determinatzaile eta aurremodifikadore),

eta azkenik buru nominala. Sistema honek bestelako egiturak ere atzeman ditzakeela esaten dute egileek, horretarako aurkikuntzarako gakoa aldatzea baino besterik ez litzateke egin behar. Probatu dugun NPtool-en bertsioak bi irteera sortzen ditu:

1. CGn oinarrituriko izen-sintagma analizatzailearen irteera.

2. Modifikadore-buru egiturez osaturiko zerrenda.

Esate baterako, NPtool-ek ondorengo esaldirako : *The cast iron cylinder block is integral with the upper half of the crankcase, the lower half of which is formed by the pressed steel sump.*; honako irteera hau sortuko du:

```
the
 the <*> DET CENTRAL ART SG/PL @>N ;; @>N = determiner or premodifier
cast_iron
 cast_iron N NOM SG @>N
cylinder
 cylinder N NOM SG @>N
block
 block N NOM SG @NH ;; @NH = nominal head
is
 be V PRES SG3 VFIN @V ;; @V = verbal
integral
 integral A ABS @AH ;; @AH = adverb / adverbial / adjectival head
with
 with PREP @AH
the
 the DET CENTRAL ART SG/PL @>N
upper
 upper <Attr> A ABS @>N
half
 half N NOM SG @NH
of
 of PREP @N< ;; @N< = postmodifying PP of a structurally obvious case.
the
 the DET CENTRAL ART SG/PL @>N
crankcase
 crankcase N NOM SG @NH
@comma
the
 the DET CENTRAL ART SG/PL @>N
lower
 low A CMP @>N
half
 half N NOM SG @NH
of
 of PREP @N<
which
 which PRON WH NOM SG/PL @NH
is
 be V PRES SG3 VFIN @V
```

```
formed
 form PCP2 @AH ;; note that the "@V / @AH" distinction
by              ;; is left blurred in this description
 by PREP @AH    ;; because it has few consequences for
the             ;; NP extraction. The grammar should be
 the DET CENTRAL ART SG/PL @>N ;; modified if we were also
pressed         ;; interested e.g. in V -- NP constructions.
 press PCP2 @>N
steel
 steel N NOM SG @>N
sump
 sump N NOM SG @NH
@fullstop
OK      cast iron cylinder block
OK      lower half
OK      pressed steel sump
OK      upper half of the crankcase
```

Adibide honetan ikus daiteke nola lehenengo esaldiaren analisia egiten duten, eta ondoren *OK* laburdura aurretik jarrita eta lerroka ezagutu diren izen-sintagmak bereizten dituzte.

Tresna hau erabilgarria izan daiteke, beste batzuen artean, honako aplikazio hauetarako: informazio-eskurapeneko (*Information Retrieval*, IR) sistemetako, itzulpen-unitatea aurkitzeko, eta aplikazio terminologikoetarako. NPTool-i buruzko oso deskribapen zehatza dago (Voutilainen, 93)-n.

NPTool tresna hobetuz, EngLite parserra garatu dute Helsinkiko taldekoek. Bere aurreko sistemarekiko honako hobekuntza hauek ditu:

- Arkitektura sinpleagoa. Eta era berean, errepresentazio sintaktikoa informatzaileagoa da, esate baterako, aditz-kateak zehatzago deskribatzen dira.
- Doitasun orokorra hobea da.
- Softwareari dagozkion osagai guztiak eraginkorrago berridatzi dira.
- Hobekuntza hauek direla eta, NPtool-a baino azkarragoa da, eta bere analisi linguistikoak ere NPToolenak baino informatzaileagoak eta zehatzagoak dira.

Bestalde, TACAT (*TAgged Corpus Analyser Tool*) ITEM (TIC96-1243-C03-02) proiektuaren barnean garaturiko parentesizatzailea ere aztertu dugu. TACAT-ek testu etiketatuak analizatzen ditu, eta bere helburu nagusia hau da: ahalik eta gizakiaren lan txikienarekin parentesizaturiko eta sintaktikoki analizaturiko corpus handiak erdiestea (corpus orokorrak zein domeinu espezifikokoak). Corpus handiak erabat edo partzialki sintaktikoki etiketatuak, edota bakarrik parentesizatuak (*tree banks*) oso beharrezkoak dira lengoaia naturalaren ezagutza-iturriekin erlazionaturiko hainbat ikasketa-prozesutarako. Sarrerako testuak etiketaturik egon behar du, eta testu horren gainean lan egin ahal izateko formatu jakin

bat eskatzen da. Baldintza horiek bete ondoren, tresna honek testuingururik gabeko (*context free grammar*, CFG) idazteko aukera ematen du. Gramatika horretan zehazten denaren arabera eraikiko ditu egitura parentesizatuak. Hau da, esaldian, idatziriko gramatikaren arabera ezagutu ahal dituen egitura sintagmatikoak parentesi artean jarriko ditu tresna honek. Hona hemen adibide bat:

```
{{el_del_gato_n}_sn  come_v  {pescado_n}_sn}_oracio
```

## IV.7 *Constraint Grammar* formalismoa aukeratzeko arrazoiak.

Formalismo hau aukeratzearen erabakia hartzeko taldean eginiko azterketaren berri emango dugu atal honetan. Horrela bada, sintaxiari heldu beharra iritsi zenean, analisi sintaktiko konputazionalaren arloko hurbilpen batzuk hartu ziren kontuan eta (Abaitua eta beste, 1993) lanean jaso ziren. Lan horretan, euskarazko sintaxiaren deskripziorako interesgarrientzat jotzen ziren formalismoak konparatu genituen euskararen tratamendurako egokitasunari erreparatuz. GPSG (Gazdar et al., 1985) eta LFG (Kaplan, Bresnan, 1982) formalismoak eta beren inplementazioak (*ALVEY Natural Language Tools* (ANLT)) (Carrol et al., 1991) eta GFU-LAB (Ruiz, 1991) LFG teorian oinarritzen dena, arreta bereziz aztertu ziren. HPSG -k (Pollard et al., 1987) eta *Constraint Grammar Parser* -ek (Karlsson, 1990) ere, beren tokian izan zuten azterketan.

Aipatu egokitasuna ikusteko, hizkuntzaren ezaugarri bereizgarrienetako batzuk hartu ziren kontuan, hala nola 1) perpaus mailako osagai sintagmatikoen ordena librea, 2) aditzaren azpikategorizazioa eta 3) aditzaren komunztadura subjektu, objektu zuzen eta zehar-objektuarekin. Honekin batera, analisiaren helburuak egitura ez-konfigurazionala eta laua izan behar du, osagaien ordena librea bermatuko duena.

Azterketa honetan ikusi zen GFU-LAB sistemak ez duela arazorik hiru fenomenoak deskribatzeko, izan ere, euskara bezalako hizkuntzak aztertzeko sortua izan zen. ANLT sistemarekin egindako analisiaren ondorioa da hiru fenomenoak tratatzeko GPSG teorian orokortzat hartzen diren printzipio batzuk birplanteatu egin behar direla. HPSG gramatikek, GPSG eta LFG gramatiketako ezaugarriak bateratzen dituztenez, teorian ematen du hiru fenomenoak deskribatzeko gai izango dela, baina praktikan ez zen probatu.

ENGCG (*English Constraint Grammar*) (Voutilainen eta beste, 1992) ingeleserako gramatika eta lexikoak osaturik dago. Arras diferentea da oinarrizko egituran, alegia, ez da oinarritzen testuinguruarekiko independente diren gramatiketan, erregeletan baizik, egoera finituko automatetan kodifikatzen diren erregeletan, hain zuzen ere. Informazio morfologikoa zutabetako bat izateak erakargarri egiten zuen guretzat, kontuan hartzen badugu euskararen aberastasun morfosintaktikoa. Helburuetan ere bat gentozen, izan ere, testu errealak tratatzea

dute helburu. Automaten bidez inplementatua egoteak, gainera, eraginkortasun handia ematen dio.

Arrazoi hauengatik CGren alde egitea erabaki genuen euskal sintaxiaren tratamendu zabalerako gramatika idazterakoan[1].

Horrez gain, EHko aditzen adibideen azterketak duen helburu nagusiari erantzuteko bide egokia dugula uste dugu. Azaleko parserrak, besteak beste, azpikategorizazio-ereduak eskuratzeko baliatu izan baitira. Hona hemen, azaleko parsing-ak izan ditzakeen aplikazioei buruz Abney-k (1997:134) dioena:

> "Partial parsing has been put to use in a variety of ways, including bootstrapping a more complete parser, terminology and multi-word term extraction for information retrieval, and as a component of data extraction systems.
>
> The chief goal in bootstrapping is the acquisition of lexical information needed for more complete parsing. The type of information to be acquired is primarily collocational, particularly subcategorization frames and selectional restrictions."

Horrela bada, azaleko sintaxia burutu dugu, CG formalismoa jarraituz, euskararako garatu dugun Murriztapen-gramatika (EUSMG).

# IV.8 Euskararako Murriztapen-gramatika (EUSMG).

Atal honetan euskararako garatu dugun Murriztapen-gramatikaren atal nagusienak deskribatuko ditugu: desanbiguatze-erregelak, euskararako funtzio sintaktikoak eta funtzio sintaktikoen esleipenerako erregelak.

## IV.8.1 Desanbiguatze morfosintaktikoa.

Desanbiguatze morfosintaktikoa izendatu dugun atal honetan funtzio sintaktikoen esleipenari ekin ahal izateko, egin beharreko desanbiguatze-lanaz arituko gara oro har (Aduriz eta beste, 97).

Analizatzaile morfologikoak (zatikatzaile morfologikoak) hitz-forma orori posible dituen analisi guztiak eta hauei loturiko informazioa asignatzen dio. Analisiaren emaitza dira hitzaren analisi-aukera guztiak, zeinetan morfema bakoitzari lexikoian adierazitako ezaugarriak atxikiko zaizkion: kategoria, azpikategoria, kasua, numeroa, mugatasuna, funtzio sintaktikoa, etab. Urrats honetan testuinguruaren arabera interpretazio-multzo horretatik zein aukera den zilegi

[1] Gaur egun, formalismo hori lantzeaz gain, bestelako ikuspegirik ere landu da taldean sintaxiari ekiterakoan (ikus Aldezabal eta beste, 1999). Lan horretan *chart-parsing* eta egoera finituko *parsing* teknikak konbinatzen dituzte.

erabakitzeko erregelak aplikatuko ditugu. Horrela bada, erregela hauen xedea anbiguo den hitz oro analisi bakarrarekin uztea izango da. Era berean CGren printzipioetariko bati jarraituz, ehuneko ehunean ziurrak diren erregelak garatzen saiatu behar dugu, hots, kasu guztietan analisi zuzenarekin asmatuko dutenak. Ahalik eta zurrunen izateko, hurbileko testuinguruan aplikatuko direnak, eta, ahal bada, dauden interpretazio posibleetarik bat behintzat saihesten duten modukoak lantzea komeni da. Jakina, estrategia nagusia hori izan arren, kasu batzuetan urruneko testuinguruaz baliatu beharra izango dugu zenbait anbiguotasunen arazoa konpontzeko.

### IV.8.1.1 Anbiguotasun-mota nagusiak.

Ondoren aurkeztuko ditugun anbiguotasun-multzoak, EDBL eta analizatzaile morfologikoak emandako informaziotik aterata daude. Hau da, aurretik, bai batean eta bai bestean hartutako erabaki eta irizpide jakin batzuen ondorio dira. Eginiko deskribapen linguistikoak baldintzatzen baitu anbiguotasuna. Oro har, forma bat anbiguo izango da interpretazio bat baino gehiago duenean, hau da, kategoria bat baino gehiago, kasu edo mugatasunaren mailan aukera bat baino gehiago dituenean, edota esaldian funtzio sintaktiko bat baino gehiago betetzen duenean. Gerta daiteke, bestela ere, forma bat anbiguo izatea lema desberdinetatik sor daitekeelako. Adibidez:

limoi -> 1. lima + oi

2. limoi

Anbiguo ziren elementuak aztertu ondoren, hiru multzotan sailkatu genituen: anbiguotasun kategoriala, morfosintaktikoa eta sintaktikoa. Hala ere, esan beharra dago multzo hauen arteko mugak ez direla gardenak, eta askotan zail dela mugak ezartzea, batez ere, anbiguotasun morfosintaktiko eta sintaktikoari dagokienez.

- Anbiguotasun kategoriala: hitz-forma bakarra, kategoria bat baino gehiagokoa izan daitekeenean (bi, hiru edo lau kategoriatakoa izan daitekeenean). Hona hemen, multzo nagusietako batzuk:

  •ADL (aditz laguntzailea)/ADT (aditz trinkoa) : dio, du, da, ...

  •ADB (adberbioa) /ADJ (adjektiboa) : aldrebes, azkar ...

  •ADI (aditza) /ADJ (adjektiboa) /ADB (adberbioa) : bizkor, ...

  •ADB (adberbioa) /ADI (aditza) : ados, ...

  •IZE (izena) /ADJ (adjektiboa) /ADI (aditza) : ordena, zati, txiki, ...

  •ADIOIN (aditzoina) /PART (partizipioa): egin, esan, ...

Anbiguotasun-mota hau aztertzeko kategoria nagusiak hartu ditugu kontuan (19 kategoria, ikus B eranskina), eta corpusetan eginiko azterketen arabera 1.55 interpretazio izango ditugu hitz bakoitzeko kategoria nagusiak kontuan hartuz gero.

- Anbiguotasun morfosintaktikoa: hitz-forma batek deklinabidea edota beste ezaugarri morfosintaktiko batzuk (numeroa, aspektua, etab.) direla medio interpretazio bat baino gehiago dituenean, anbiguotasun morfosintaktikoaren aurrean gaudela esan dezakegu. Dena den, morfologia eta sintaxiaren arteko muga ez da oso gardena euskaraz. Hizkuntza eranskarietan, morfologia eta sintaxiaren arteko harremanak oso estuak dira; hori dela eta, anbiguotasun morfosintaktikoa deitu diogu bigarren multzo honi. Esate baterako, absolutibo pluralaren eta ergatibo singularraren artean erabakitzerakoan, sintaxirako urratsa ematen ari gara. Hau da, aditz iragankor batekin azaltzen baldin bazaigu halako anbiguotasuna, absolutiboa ala ergatiboa den erabakitzerakoan ondorengo urratsean (funtzio sintaktikoen esleipenean) zer funtzio izango duen hautatzen ari gara zeharka bada ere.

  Tratatu beharreko multzo honetako anbiguotasunen artean koka dezakegu ere, izen batek izen-sintagman izan dezakeen posizioa kontuan hartuz gero sortzen den anbiguotasuna. Adibidez, *aho* hitz-formak analisi morfologiko hau du:

  "<Aho>" [1]
  *"aho" IZE ARR DEK ABS MG @OBJ @SUBJ HAS_MAI*
  *"aho" IZE ARR ZERO HAS_MAI*

  Anbiguotasuna IZE ZERO vs IZE ABS bezala deskribatu dezakegu. Interpretazio batean IZE ZERO analisia izango du hitz honek, baldin eta ez bada izen-sintagma horretako kasua, numeroa eta mugatasuna eramango dituen elementua. Adibidez: *Zure **aho** beltz hortatik ateratzen diren hitzek ez naute ikaratuko.* Bestalde, izen-sintagmako azken osagaia denean berak hartuko ditu kasua eta mugatasunaren ezaugarriak (eta baita funtzio sintaktiko nagusiak ere). Hona hemen adibide bat: *Hainbat **aho** aztertu zituen goiz hartan haginlariak.*

  Anbiguotasun honen neurketa egin dugu 10.000 testu-hitzeko corpus batean, eta anbiguotasun-tasa 3.05ekoa dugu oro har. Hitz ezezagunen kasuan 7.05 interpretaziotara ere igo daiteke anbiguotasun-tasa. Anbiguotasun-mota hau dugu zailena ebazten.

---

[1] Analisian agertzen diren laburduren esanahia: ABS: absolutiboa; ARR: arrunta; DEK : deklinabide-morfema; HAS_MAI: maiskulaz hasten den hitza; IZE: izena; MG: mugagabea; ZERO: kasurik gabea; @OBJ: objektu funtzioa; @SUBJ: subjektu funtzioa.

- Anbiguotasun sintaktikoa: aurrerago aipatu dugun legez morfologia eta sintaxiaren arteko harremanak oso handiak izanik, zail da hauen artean mugak ezartzea. Hala ere, anbiguotasun sintaktikoaz hitz egingo dugu hitz-forma batek funtzio sintaktiko bat baino gehiago dituenean. CG formalismoan funtzio sintaktikorako etiketek arroba sinboloa eramaten dute aurretik, adibidez: @SUBJ, @OBJ, @+JADNAG, ...

### IV.8.1.2 Desanbiguatze-erregelak.

Desanbiguatze-erregela bakoitzak kategoria gramatikalez osaturiko kate bat definitzen du. Hau da, hitz-forma batek dituen interpretazioetariko bat aukeratzen edota ezabatzen dugunean testuinguruko beste hitz-forma baten interpretazioen arabera, eta beste horrenbeste egiten dugunean azken honen interpretazioekin, egiten ari garena definizio gramatikal bat zehaztea dela esango nuke. Modu honetara uler daitezke erregelak, eta honela argiago gera daiteke erregela hauek denek batera gramatika bat osatzen dutela. Baina, ez da debekuz osaturiko gramatika bat, baizik eta definizio partzialez osaturikoa. Murriztapen bat datuei buruzko "predikzioak" egiten dituen hipotesia da. Adibidez (ondoren datozen adibideak hobeto ulertzeko, —ikus IV.4.3 atala.—):

> Aditz-izen, izen (hiztegiko sarrera), eratorri (-te/-tze atzizki lexikalaren bidez) formen artean erabakitzeko, lehenengo urrats batean desanbiguatze-erregela baten bidez esan dezakegu ez dela aditz-izena izango baldin eta desanbiguatu nahi dugun hitz horren eskuinetara urrats batera adjektibo edo determinatzaile bat dagoen:
>
> ```
> SELECT (ARR) IF (0 ZERO AND ADIZE) (1 ADJ OR DET);
> ```
>
> Adib.: *ILUNTZE alferra izan dut gaur ia beti legez.*

Erregela hauek idazteko egin diren urrats nagusiak hauexek dira:

- Lehendabizi, anbiguotasun-motak aztertu ditugu. Honetarako, EDBL eta analizatzaile morfologikoaren irteera izango dira iturri nagusiak. Azterketa honetan, anbiguotasun-mota desberdinak bereizteaz gain, maizen gertatzen diren fenomenoetan jartzen dugu arreta. Hau da, erregelak diseinatzerakoan maizen gertatzen diren anbiguotasunak ebazteari ekiten zaio aurrenik. Horretarako, anbiguotasun hori corpusean neurtuko da. Era berean, corpusa baliatuko da ebatzi nahi den anbiguotasun horri aurre egiterakoan, zehaztu beharreko testuinguruak erabakitzen laguntzen baitigu. Corpus orokorrez gain, gramatikak eta hiztegiak ere erabiliko ditugu hitz baten erabilerak erregela bidez definitu nahi ditugunean.

- Eskuz desanbiguaturiko corpusa: badugu 24.000 hitzez osaturiko corpus desanbiguatu bat. Corpusa bi linguista desberdinen artean desanbiguatua izan da, eta emaitzak beraien artean konparatu dituzte adostasunera iristeko asmoz. Metodo hau

*double blind* gisa ezagutzen da (Sammuelsson & Voutilainen, 97). Eskuz desanbiguaturiko corpus hau oso garrantzitsua da:

- Adosturiko etiketatze-eskema bat definitzen duen heinean. Errepresentazio gramatikal hau kontsultarako edota desadostasun kasuak konpontzeko erabiliko da.
- Etiketatzaile automatikoaren bidez lorturiko emaitzak egiaztatzeko balioko du.

- Erregelen diseinua. Erregelek lehenago deskribaturiko anbiguotasunak ebatzi beharko dituzte. Erregela hauek definitzeko desanbiguaturiko corpusaren zati bat erabiliko da. Gainontzeko zatia ez da aztertuko, eta honetan aplikatuko dira erregelak, hauen sendotasuna frogatzeko. Jakina, erregelen emaitzak hobetu nahi izanez gero, berriz ere birfindu egin beharko ditugu.

Adibideen analisirako osatu dugun gramatikak ia 700 erregela morfosintaktiko ditu, horietatik 298 hitz konkretuak tratatzeko dira. Erregela hauek, anbiguotasun morfosintaktikoari dagokionez, interpretazio erdiak kentzen dituzte %97,86ko zuzentasunarekin. Eta, soilik kategoria nagusiei erreparatuz gero, herena desanbiguatzen dugu %99,12ko zuzentasunarekin (Aduriz eta beste, 97).

Erregela hauek hiztegiko adibideen gainean aplikatzerakoan, erregela hauen sendotasuna ere handiagotu egiten dela esan daiteke. Zeren eta, testu-mota desberdinetan aplikatu ondoren, emaitza onak lortzen baitira. Hala ere, hau desanbiguatze kategorial eta morfosintaktikoari dagokionez esan dezakegu. Sintaktikoari dagokionez, aldiz, emaitzak erabat frogatzeke daude, zeren eta ez dugu eskuz desanbiguaturiko etiketa sintaktikodun corpusik. Corpus hori oso lagungarria izango litzateke emaitzak konprobatzeko eta hiztegiko adibideen ezaugarriengatik sortu diren erregelen eragina neurtzeko. Batez ere, kontuan izanik adibide anitz ez direla esaldi osoak. Eta esaldi osoak ez izateak eragin franko izan ditzake desanbiguatze sintaktikoari heltzerakoan.

Aurrerago aipatu ditugun anbiguotasun-mota nagusiei aurre egiteko darabilzkigun erregela horietako batzuk aurkeztuko ditugu ondoren:

i) Anbiguotasun kategoriala tratatzen dutenen artetik, aditz laguntzaile/trinkoen arteko anbiguotasunari aurre egiteko zenbait erregela deskribatuko ditugu. Aditz laguntzaile eta trinkoaren artean erabakitzeko, zenbait erregela ditugu aditz laguntzailearen interpretazioa hartzen dutenak baldin eta ez dagoen esaldian aditz-lokuzioetako osagairik (nahi, behar...), ez aditz konposatuetako osagairik (*lo egin, min hartu*, ...), ez aditzoinik, ez aspektu burutu, ezburutu edo geroko aditz-formarik, ez *izan*, *izaten*, *izango* formarik, ez aditzarekin konposizioa osa dezaketen forma batzuk (*berri*, *zahar*, ...). Adibidez:

```
SELECT (ADL) IF (0C ADL/ADT) (-2 NOTDEK)
                        (NOT -1 ADIELK) (-1 PRT-ZIU OR ADITZMULTZO)

                        (NOT -1 ADPOSAG;
```

Adib.: *Mikel estropadetatik etorri omen DA lan egitera.*

```
SELECT (ADL) IF (0 ADL/ADT) (NOT -1 ADPOSAG)
                            (-1 EZEZKOAK) (1C SIN) (NOT 1 ADP)
                            (NOT 2 ADPOSAG) (NOT 1 IZANTENGO);
```

Adib.: *ez ZUEN egin*

```
SELECT (ADL) IF (0C ADL/ADT) (NOT *1 ADL)
                (NOT *1 ADPOSAG BARRIER PUNTUAZIOA)
                (*1C BURU BARRIER PUNTUAZIOA LINK NOT 1 ADIKONP);
```

Adib.: *ez GENUEN atzo goizean kalean ikusi*

ii) Anbiguotasun morfosintaktikoari dagokionez bi anbiguotasun ebazteko murriztapenak azalduko ditugu. Batetik, ABSolutibo/ERGatibo anbiguotasuna ebazteko erregelak, eta bestetik, ZERO (hau da, ez kasu ez mugatasunik daraman izen, adjektibo, determinatzaile, etab.), eta elementu horiek berek kasu absolutibo mugatua edo mugagabea eraman dezaketenean.

a) ABS/ERG anbiguotasuna ebazteko planteatzen da ergatiboa saihestea baldin eta ez eskuinetara ez ezkerretara ez dagoen ergatibo horrekin komunzta dezakeen laguntzailerik edota aditz iragankorrik. Absolutiboa ere kendu egingo da ez dagoenean beste ergatiborik, eta ezta ere berarekin komunzta dezakeen NOR laguntzailerik. Adibidez:

```
REMOVE (ERG) IF (0 ABS) (NOT 0 DET)
                        (NOT *-1 (NK_HU) OR NOR_NORK OR NOR_NORI_NORK)
                        (NOT *1 (NK_HU) OR NOR_NORK OR NOR_NORI_NORK);
```

Adib.: *GALBURUAK ezker eskuin kulunkatzen dira.*

```
REMOVE (ABS) IF (0 ERG) (NOT *1 PUNTUAZIOA) (NOT *-1 ERG)
                    (NOT *1 ERG) (NOT *1 NR_HK) (*1 NK_HU);
```

Adib.: *KAREAK ongarria ahuldu egiten du.*

b) ZERO / ABS MUGM S edo ABS MG, ZERO markak adierazten du elementu hori ezin dela sintagma batean kasua eta numeroa daramatzana izan, orduan ideia da interpretazio hau hautatzea hiru aukera horietatik, baldin eta bere eskuinetara urrats batera deklinaturik dagoen izen, adjektibo edo determinatzaileren bat agertzen bada, edota

elementu horiek berak deklinatu barik. Ez du onartuko eskuinetara urrats batera posposizioen artean kontsideratzen den osagairik (adib. arte, alde, atze, etab.). Ezkerretara urrats batera ez da azalduko determinatzailerik edo mailakatzailerik ZERO interpretazioa hobesten bada. Ezkerreko posizio horretan ager daitezkeenak beste izen batzuk edota genitibo kasuan dauden elementuak izango dira. Adibidez:

```
SELECT (ZERO) IF (0C IZE + (ARR))
          (NOT O AORG OR GEN/GEL OR POSPOZ OR ADIKONP)
          (NOT -1 DET OR MAILAKATZAILEAK)
          (NOT 1 POSPOZ) (1C DEK);
```

Adi. *HITZ eratorriez hizkuntza aberastu*

Erregela honi buruz komentatu behar da, tratatzen duen hitzak izen arrunta izan behar duela, *a* organikodunik gabea, genitibozko kasuan ez doana, eta ez dela posposizio edo aditz konposatuetako osagaien multzokoa izango. Baldintza horiek betez gero, eta ezkerretara ez baldin badago determinatzailerik edo mailakatzailerik; eta eskuinetara ez badago posposiziorik baina bai deklinaturik dagoen elementuren bat, orduan ZERO etiketa duen interpretazioa hautatuko du erregela honek.

```
REMOVE (ZERO) IF (0 ABS) (NOT -1 ABS) (NOT 1 DEK)
                        (NOT 0 PART) (1 ADI) (NOT 2 DET);
```

Adib.: *HIZKUNTZA aberastu ...*

iii) Desanbiguatze-erregelen artean badira hitz konkretuen anbiguotasunaz arduratzen direnak ere. Adibidez, *eta* juntagailua batez ere esaldi bukaeran menderagailu legez (MEN KAUS) etiketarekin analizatzen dugu, hau da esaldi kausala osatzeko menderagailu gisa. Interpretazio hori batez ere bere eskuinetara urrats batera puntua edota puntu eta koma dituenean suertatuko da. Horrela ez denean (MEN) interpretazio hori kendu egingo dugu eta loturazko juntagailuaren interpretazioarekin geratuko gara. Ikus ditzagun *eta* juntagailuari dagozkion zenbait erregela:

```
"<eta>" REMOVE (MEN) IF (-1 PUNT/KOMA/PKOMA/BEREIZ);
```

Adib.: *Atzo goizean heldu zen, ETA berehala hasi zen lanean*

```
"<eta>" REMOVE (MEN) IF (1 IZE/ADJ/PAR)
                            (-1 IZE/ADJ/PART);
```

*Batean hasi ziren hazi biak, ona ETA gaiztoa, agertzen.*

```
"<eta>" SELECT (MEN) IF (-1 ADL/ADT) (1 PUNT/PKOMA);
```

Adib.: *Bihar ez naiz azalduko ez dut gogorik ETA.*

## IV.8.2 Euskararako funtzio sintaktikoak.

Era berean, EUSMG erabiliko da behin desanbiguatze morfosintaktikoa burutu ondoren; kasurik hoberenean, morfosintaktikoki erabat desanbiguaturiko testuaren gainean lan egingo genuke funtzio sintaktikoak esleitzerakoan.

Etiketa sintaktiko[1] hauek hitzei esleitzen zaizkie ezaugarri morfologikoak esleitu zaizkien modu berean. Adibidez:

```
"<Berri>"
 "berri" IZE ARR ZERO HAS_MAI @KM>
"<haiek>"
 "haiek" DET ERKARR NUMP DEK ABS NUMP MUGM @OBJ
"<entzunik>"
 "entzun" ADI SIN AMM PART ERL MEN MOD NOTDEK @-JADNAG_MP
"<adoretu>"
 "adoretu" ADI SIN AMM PART ASP BURU DA-DU NOTDEK @-JADNAG
"<zen>"
 "izan" ADL B1 NOR NR_HU @+JADLAG
"<mutila>"
 "mutil" IZE ARR DEK ABS NUMS MUGM @SUBJ
"<$.>"
```

EUSMGn erabakitako funtzio sintaktikoak *Euskararako murriztapen-gramatika: lehen urratsak* (Aduriz eta beste, 1996)-n azaltzen direnetan oinarritzen dira. Hala ere, badira desberdintasun batzuk: mendeko perpausen funtzioetan, mendekoaren funtzioa zehatzago adierazten dugu, eta atributu funtzioaren (@ATRIB) ordez predikatua (@PRED) baliatuko dugu[2].

Erabakiak hartzerakoan, anbiguotasun sintaktikoaz gain, aurkitu dugun arazo nagusia esaldi nagusi eta mendeko perpausen arteko erlazio sintaktikoen adierazpena izan da. Ondoren, egindako lanean gai hauen inguruan aurkitutako arazoak eta hartutako erabakiak azaltzeari ekingo diogu. Azalpenean zehar zenbait adibide eta gramatikaren funtzio sintaktikoak

---

1 Etiketa sintaktikoek @ ikurra daramate aurretik.

2 Azaldu beharra dago adibideak aztertu ziren garaian @ATRIB funtzio sintaktikoa jasotzen zela EDBLn. Beraz, zenbait adibidetan funtzio hori ager dakiguke. Hala ere, @PRED funtzioa aparteko mapaketa-gramatika baten bidez esleitu diegu adibideei. Kontua da, adibideei gramatika aplikatzerakoan ordura arte landutako deskribapen linguistikoarekin bat ez gatozenean aldaketak egin ditugula.

aurkeztuko ditugu, eta azkenik, analisi sintaktikoaren azterketa dela medio atera ditugun ondorioak.

Oinarrian, ENGCGren etiketa sintaktikoen filosofia jarraitzen dute (Aduriz eta beste, 1996)-n landuriko funtzio sintaktikoek. Hala ere, zenbait erabaki hartu, eta zenbait aldaketa egin behar izan ditugu. Batetik, izen-sintagmaren barneko dependentziak adierazteko, @KM> funtzio sintaktikoa, hots, kasu-markadunaren modifikadorea sortu dugu. Zer adierazten du funtzio "bitxi" honek? Esan gabe doa, gainontzeko CGko funtzioen estiloari jarraituz, eta kontuan izanik hitz-forma orori funtzio bat esleitu behar zaiola, ez dela funtzio sintaktiko "tipiko" horietakoa. Kontua da, izen-sintagmaren osagai izan daitezkeen elementuetatik batek bakarrik hartzen dituela kasua eta numeroa, eta hain zuzen ere izen-sintagmaren kasua daraman osagai hori izango da buru eta horrekiko lotura adierazten dugu funtzio honekin. Horrez gain, funtzio horretan azaltzen den norabidea ere adierazgarria dela esango genuke, eskuinera dagoen osagai bati lotua dagoela adierazten baitu.

Bestetik, perpaus jokatu eta ezjokatuei dagozkien etiketetan hauek bete ditzaketen funtzioak bereizi ditugu. Hau da, ENGCGri jarraituz gero, ezin izango genuke adierazi mendeko ezjokatu batek perpaus nagusian zer funtzio bete dezakeen. (Aduriz eta beste, 1996)-en ere, ez zen medeko perpausaren funtzio sintaktikoa oso zehatz adierazten. Funtzio sintaktiko horietan, alde batetik, mendeko esaldiaren barruan aditzari dagokion funtzio sintaktikoa azaltzen da, eta bestetik, esaldi nagusiarekiko funtzio sintaktikoa adierazten da. EUSMG-n, ordea, mendekoaren funtzio sintaktikoa zehatzago adierazten da[1]. Esate baterako, *Erretzeak kalte egin dit* adibidean, *Euskararako murriztapen-gramatika: lehen urratsak* eranskinean proposatzen genuenari jarraituz gero, *erretzeak* @-JADNAG_MP (Aditz nagusi ezjokatu / Mendeko perpausa) etiketarekin azalduko litzaiguke. EUSMG-n proposatzen ditugun etiketen bidez *erretzeak* @-JADNAG_MP_SUBJ (Aditz nagusi ezjokatu / Mendeko perpausa subjektua) gisa etiketatuko dugu anbiguotasun sintaktikoa ebatziz gero.

Ikusi dugun legez —ikus § IV.4—, CGren analisi-prozeduran burutzen den azken urratsa funtzio sintaktikoen asignazioa da. Eta helburua hitz bakoitzari funtzio sintaktiko bakarra asignatzea da. Adibidez, *Udaberria etorri da* esaldiaren analisia ondokoa litzateke:

Udaberria (@SUBJ)

etorri (@-JADNAG)

da (@+JADLAG)

---

[1] Aurrerago esan dugun bezala, zenbait kasutan EDBLn zegoen deskribapen linguistikoa aditz-adibideen azterketarako aplikatzerakoan zenbait aldaketa egin ditugu. Mendeko funtzio sintaktikoen kasuan, adibideen azterketa egin zenean, EDBLn ez zen mendekoak nagusian zuen funtzio sintaktikorik zehazten. Hori dela eta, gure azterketaren mesederako zirelakoan, mapaketa-gramatika baten bidez mendekoen funtzio sintaktiko zehatzagoak esleitu dizkiegu adibideei. Beraz, zenbait kasutan bi bide horietatik datozen mendekoen funtzio sintaktikoak aurki daitezke.

(@SUBJ = subjektua @ -JADNAG = aditz nagusi ezjokatua; @+JADLAG = aditz laguntzailea jokatua).

Hitz bakoitzari funtzio bakarra asignatzearen beharra dela eta, anbiguotasun sintaktikoa adierazterik ez dago zenbait kasutan. Azter dezagun honako hau: *Haurrak eraman ditu*; kasu honetan esaldiaren bi analisi sintaktiko posible daude, haurrak (@OBJ) edo haurrak (@SUBJ) (@OBJ = objektua; @SUBJ = subjektua). CGren helburua hitz bakoitzari funtzio sintaktiko bakarra asignatzea denez, aukeratu beharko genuke *haurrak* hitzari zer asignatu, @OBJ edo @SUBJ. Erabakia hartzeko ezagutza pragmatiko-semantikoa behar da nahitaez. Soluzioa izan liteke gure analizatzaileak inolako erabakia ez hartzea honelako kasuetan, eta emaitza bezala bi funtzio sintaktikoak uztea. Dena den, orain har dezagun bigarren adibide bat: *Zortzigarren udako ikastaroa hasi da*; esaldi honen analisia ondokoa litzateke:

Zortzigarren (@IA>)

udako (@IZLG>)

ikastaroa (@SUBJ)

hasi (@-JADNAG)

da (@+JADLAG)

(@IA> = ezkerreko izenondoa; @IZLG> = ezkerreko izenlaguna).

Kasu honetan, *zortzigarren* bai *udako* bai *ikastaroa* hitzen adjektibo modifikadorea izan daiteke. Oraingo honetan soluzioak ezin du izan emaitza bezala bi funtzio sintaktiko uztea, bi posibilitate horiek adierazteko ahalmenik ez baitago CGn; @IA> funtzio sintaktikoak *zortzigarren* hitza bere eskuinean dagoen izen baten modifikadorea dela esan nahi du, baina ez du adierazten zein izenen modifikadore den, are gutxiago bi izen desberdinen modifikadore izan daitekeenik.

Etiketa sintaktikoak bi motakoak dira: modifikadoreak edota beren buruak markatzen dituztenak, eta nagusiak. Modifikadore-etiketek beren burua zein norabidetan dagoen adierazten dute. Adibidez, @IZLG> etiketa izango dugu bere eskuinetara dagoen izen bat modifikatzen duten izenlaguntzaileentzat, eta etiketa honek adierazten du modifikatzen duen burua eskuinetara dagoela. Funtzio sintaktiko nagusiek, hots, buruei dagozkien etiketek, perpauseko osagai tipikoak errepresentatzen dituzte, hala nola: subjektua, objektua, zeharkako objektua, subjektuaren predikatu konplementua, ... Oro har, funtzio hauek gramatika tradizionaletik oso gertu daudenak direla esan genezake.

Funtzio sintaktikoei buruzko kontsiderazio orokor horien ondorik, funtzio sintaktikoak azaltzeari ekingo diogu. Lau multzotan aurkeztuko ditugu: funtzio sintaktiko nagusiak, izen-sintagma barruko dependentzia sintaktikoak adierazteko funtzioak, aditzen funtzio sintaktikoak, eta, azkenik, bestelako funtzio sintaktikoak. Hala eta guztiz ere, bereizketa

nagusia funtzio sintaktiko nagusien eta modifikadoreen artekoa da. Nagusiak dira hain zuzen ere etiketa sintaktikoan >/< ikurrik ez daramatenak, hau da, adierazten dituzten funtzioak perpaus osoari dagozkio. Modifikadoreak aldiz, >/< ikur horien bidez beren buruarekiko lotura adierazten dute, zein norabidetan aurkitzen den burua adieraziz. Bestalde, lehenago aipatu dugun bezala —ikus § IV.4—, funtzio sintaktikoak lexikoian adieraziko dira, edota mapaketa morfosintaktikoen bidez esleituko zaizkie hitzei.

### IV.8.2.1 Funtzio sintaktiko nagusiak.

Atal honetan funtzio sintaktiko nagusiak jasotzen dira:

| | | |
|---|---|---|
| @SUBJ | Subjektua | *Haize eraso* ***batek*** *abarrikatu du txabola* |
| @OBJ | Objektua | ***Kiratsa*** *aditzen duzu* |
| @ZOBJ | Zehar objektua | ***Euskarari*** *beti agertu dion atxekimendu beroa* |
| @ADLG | Adizlaguna | *Sugetzar bat agertu zitzaion* ***bidean*** |
| @PRED | Predikatua | ***Ordezkari*** *hautatu dutenari* |

Funtzio sintaktiko nagusi hauen zerrendako azkenari buruz esan beharra dago etiketa hau hautatu genuela atributu eta predikatuaren arteko bereizketarik egin gabe. Hau egitearen arrazoia, batez ere ondorengo azterketarako bata eta bestearen arteko bereizketak ez duelako eraginik. Baina, jakina, oso inportantea da funtzio hau izatea argumentu-egituren berri eman nahi denean. Bestalde, ENGCGkoen etiketen artean @PCOMPL-S (*Subject Complement*) eta @PCOMPL-O (*Object Complement*) etiketak ditugu subjektu edo objektuaren predikatua den bereiztuz.

Esate baterako, ondorengo adibideetan subjektuaren edo objektuaren konplementua den bereiztuko lukete CGkoek:

| | | | | | | |
|---|---|---|---|---|---|---|
| *Ni* | *pozik* | *nago* | */ Nik* | *pozik* | *ikusten* | *zaitut* |
| | @PCOMPL-S | | | @PCOMPL-O | | |

Guk aldiz, bietan @PRED etiketarekin analizatuko genuke *pozik* konplementua.

Bestalde, (Zabala, 1993)-n lehenengo mailako eta bigarren mailako predikatuak bereizten dira. Lehenengo mailakoetan, adizki-predikatuak eta izenki-predikatuak (atributuak) izango genituzke, bigarrengoan izenki-predikatuak (predikatiboak). Lehenengo mailako predikatuak, beren argumentu guztiak kasuaz hornitzeko behar den aparailu funtzionala islatzen dutenak dira, eta horrelako aparailua isla dezaketen aditzen osagarri modura txertatzen direnak dira. Bigarren mailako predikatuak, berriz, aukerakoak dira, eta adjuntutzat hartzen dira. Izenki-

predikatuek, orokorki ezin gara dezakete beren subjektua kasuaren bidez zilegiztatuko duen kategoria, edo, bestela, aukera hori dutenean, subjektu/predikatu harremana, subjektua zilegiztatzen deneko maila baino beherago ezartzen da. Adizki-predikatuak, beren argumentu guztiei kasua esleitzea baimenduko lieketen islapen hedatuak garatzeko gai dira. Dena den, gure kasuan azaleko analisi sintaktikoa egiten ari garenez gero, etiketa bakarra izango dugu predikazioari erreferentzia egiteko: @PRED. Gainontzeko funtzio sintaktikoak, funtzio sintaktiko klasikoak direla esan dezakegu, eta, azterketa honen azken xedeari gagozkiola, aditzak eska ditzakeenak direla ohar gaitezke.

Funtzio sintaktiko hauek guztiak lexikoian adierazten dira.

### IV.8.2.2 Izen-sintagmaren barruko dependentzia sintaktikoak.

Hiru modu ikusten ditugu izen-sintagmaren barruko funtzio sintaktikoak adierazteko.

Hona adibide bat:

| *izurda* | *azkar* | *bat* | *zebilen* | *hondartza* | *hartan* |
|---|---|---|---|---|---|
| @ISG | @<IA | (@<ID @SUBJ) | @+JADNAG | @ISG | (@<ID @ADLG) |
| @SUBJ | @<IA | @<ID | @+JADNAG | @ISG | (@<ID @ADLG) |
| @KM> | @<IA | @SUBJ | @+JADNAG | @KM> | @ADLG |

Hiru aukerak aztertu eta gero, hirugarrena hartu dugu ondoko arrazoiengatik:

1. @ADLGrekin bezalaxe jokatzen da @SUBJ funtzioarekin (edo @OBJrekin), beti kasu-marka daraman elementuari esleituz sintagmak perpausean betetzen duen funtzio sintaktikoa. (Hau lehen aukeran ere betetzen da).

2. Askoz sinpleago egingo du funtzio sintaktiko nagusi horien asignazioa, testuingurua aztertzen ibili gabe, lexikotik bertatik morfemari egokitutako funtzioa hartuko baitugu. (Hau lehen aukeran ere betetzen da).

3. Hitz bakoitzari funtzio sintaktiko bakarra asignatzen zaio.

Ikus dezagun orain, nola analizatuko genituzkeen ondoko sintagmak, hartutako aukeraren arabera. (Oharra: *zuberotar* adjektiboa izenaren ezkerraldean jar daiteke)

| *zuberotar* | *mutil* | *jatorrek* |
|---|---|---|
| @IA> | @KM> | @SUBJ |
| *zuberotar* | *mutilek* | |
| @IA> | @SUBJ | |

| *Zuberoako* | *mutil* | *jatorrek* |
|---|---|---|
| @IZLG> | @KM> | @SUBJ |

Ondoren, hartutako aukera kontuan izanda, izen-sintagmaren barruko dependentzia sintaktikoak adierazteko definitu ditugun funtzio sintaktikoak azaltzeari ekingo diogu. Hona hemen:

@KM> Kasu-markadun elementuaren modifikadorea:

***emakume** bat kezkaraziko lukeen bihozkada*

@IZLG> Ezkerreko izenlaguna:

*Bere herritarrak **euskararen** egoeraz kezkarazteko*

@<IZLG Eskuineko izenlaguna:

*orri **deboziozko** madarikatua leitu dinat gaur goizean*

@IA> Ezkerreko izenondoa:

***lekeitiar** antzarrek kale egin zuten*

@<IA Eskuineko izenondoa:

*emakume **bikain** haren aurrean ez zen berehalakoan kikildu*

@ID> Ezkerreko determinatzailea:

*Lazaro **lau** egunetako hil kirastua piztu zuenean*

@<ID Eskuineko determinatzailea:

*euskaldun **gutxi** batzuek ez dezaten dena eman*

### IV.8.2.3 Aditzen funtzio sintaktikoak.

Aditzei ere funtzio sintaktikoa esleitzea, hitz-forma orori funtzio sintaktikoa esleitu behar izatetik dator.

#### IV.8.2.3.1 Nagusiak.

| | | |
|---|---|---|
| @+JADNAG | Aditz nagusi jokatua | *Bihotz-bihotzean **daramatza** kezkak iltzaturik* |
| @-JADNAG | Aditz nagusi ezjokatua | *Ardoak adimena **iluntzen** du* |
| @+JADLAG | Aditz laguntzaile jokatua | *Berri hark bihotza ilundu **zion*** |

| | | |
|---|---|---|
| @-JADLAG | Aditz laguntzaile ezjokatua | *Euskal Herriko zokorik galduenetan jaso **izan** diren hitzak* |

**IV.8.2.3.2 Mendeko perpausak.**

Atal honetan aditz jokatu eta ezjokatuekin sortzen diren mendeko perpausen funtzio sintaktikoak[1] jasoko ditugu. Etiketa sintaktiko hauetan mendeko perpausa zer-nolako aditzak osatzen duen adierazteaz gain, mendeko perpaus horrek nagusian bete ditzakeen funtzioak zehaztuko dira. Euskararako Murriztapen-gramatikaren lehenbiziko urratsetan, CGren filosofiari hertsiki jarraituz ez genuen honelako bereizketarik egiten. Kontua da, gramatika garatu ahala analisi finago baten beharra sumatu dugula. Hau da, bereizketa hauek egitea beharrezkotzat jotzen dugu, bai aditzen adibideen azterketarako bai etorkizunean egin daitezkeen sintaxiko lan sakonagoetarako. Alabaina, esan gabe doa, funtzio sintaktikoak fintze horrek desanbiguatze sintaktikoaren urratsean izango diren zailtasunak areagotzen dituela. Hots, kontua da, oso zaila dela nagusi eta mendekoen arteko harremanak bereiztea hain fin jokatuz gero.

- Aditz jokatuekin osatzen diren mendeko perpausen funtzio sintaktikoak:

| | |
|---|---|
| @+JADLAG_MP_IZLG> | Aditz laguntzaile jokatua / Mendeko perpausa izenlagun gisa<br>*Egin **duen** laguna ez da zurruna* |
| @+JADLAG_MP_ADLG | Aditz laguntzaile jokatua / Mendeko perpausa adizlagun gisa<br>*Etorri **delarik** hartu du horietarik* |
| @+JADLAG_MP_OBJ | Aditz laguntzaile jokatua / Mendeko perpausa objektu gisa<br>*Eta zuhurrena horixe **dela** nik uste* |
| @+JADNAG_MP_IZLG> | Aditz nagusi jokatua / Mendeko perpausa izenlagun gisa<br>*Zera **dakarren** hura joango da zerura* |
| @+JADNAG_MP_ADLG | Aditz nagusi jokatua / Mendeko perpausa adizlagun gisa<br>*Izenak artikulu erantsia **daramanean*** |
| @+JADLAG_APOS | Aditz laguntzaile jokatua / Aposizioa<br>*Jon, etorri **dena,** alaia da* |
| @+JADNAG_APOS | Aditz nagusi jokatua / Aposizioa<br>*Mutila, hor **dagoena,** bere anaia da* |

[1] Mendeko perpausen funtzio sintaktikoak adierazteko etiketetan azaltzen diren funtzio-bikoteak batera suerta daitezkeenak dira; batera suertatzeak ez du inola ere adierazten funtzioen arteko anbiguotasunik dagoenik, hitzak bi funtzioak batera betetzen dituela baizik. Bi funtzioak etiketa berean biltzen ditugu.

• Aditz ezjokatuekin osatzen diren mendeko perpausen funtzio sintaktikoak :

@-JADNAG_MP_KM> Aditz nagusi ezjokatua / Mendeko perpausa kasu-markadunaren modifikadore gisa

*Zuk hori **pentsatze** hutsak ikaratu egin nau*

@-JADNAG_MP_SUBJ Aditz nagusi ezjokatua / Mendeko perpausa subjektu gisa

*Hiru gisatan gertatzen da **horditzea***

*Haren **urruntzeak** guztiz ninduen hagorandu*

*Zuk hori **eginak** kezkatu egiten nau*

@JADNAG_MP_OBJ Aditz nagusi ezjokatua / Mendeko perpausa objektu gisa

*Elurrak galerazi egiten du menditik ibilbtzea*

*Etorkizunari **antzematerik** berriz ez dago*

*Zuk **egina** ongi iruditzen zait*

@-JADNAG_MP_ZOBJ Aditz nagusi ezjokatua / Mendeko perpausa zeharkako objektu gisa

*Lehenbizi **jaiotzeari** darraizkion eskubidea*

*Zuk **esanari** ez diot kasurik egin*

@-JADNAG_MP_IZLG> Aditz nagusi ezjokatua / Mendeko perpausa izenlagun gisa

*Dirua **arriskatzearen** saria, hamarretik bat*

*Bere etsaia **galtzeko** bidea asmatu zuen*

*Zuk **esan** kontua ez dut gogoratzen /*

*Badu nori **eman** dirua*

*Ekaitzak **egindako** kalteak balioztatu*

@-JADNAG_MP_ADLAG Aditz nagusi ezjokatua / Mendeko perpausa adizlagun gisa

*Lana **bukatzeko** amari eskatu nion laguntza*

*Ez **betetzekotan** , ez berba*

*Zuk hori **esanaz** ez naiz harritzen*

*Zuk hemen **ikasiarekin** aurrera egingo duzu*

*Zuk **idatziagatik** ez dute iritzia aldatuko*

*Zuk **idatziarren** ez dute kasurik egingo*

@-JADNAG_MP_PRED Aditz nagusi ezjokatua / Mendeko perpausa predikatu gisa

***Joatekoa** naiz / **Joatekoak** gara*

@-JADLAG_MP_SUBJ Aditz laguntzaile ezjokatua / Mendeko perpausa subjektu gisa

*opa **izanak** ematen ditu lanak*

### IV.8.2.4 Beste funtzio sintaktiko batzuk.

Atal honetan azaltzen diren funtzio sintaktikoak era askotarikoak ditugu: partikulei esleiturikoak (bat berezia *ez* partikularentzat eta beste bat oso zabala, askotan funtzio sintaktiko garbia zehazten zail suertatzen den partikulen kasuetarako (@PRT); juntagailuei edota menderagailuei eskainiak; graduatzaile funtzioa bete dezaketen hitz-formentzat eta funtzioari baino gehiago egitura bati erreferentzia egiten diona (@APOS) ere hemen kokatu dugu.

| | | |
|---|---|---|
| @EZ_PART | Ezezko partikula | *ez dugu deus ere jakin nahi* |
| @APOS | Aposizioa | *horiei, **Pasaiakoei,** irabaziko diegu* |
| @GRAD> | Ezkerreko graduatzailea | *Satanek **hagitz** ongi eskarniatzen du aingeru onaren figura* |
| @<GRAD | Eskuineko graduatzailea | *gorri **askoak** dira* |
| @PJ | Koordinaziozko juntagailua | *Peru **eta** Mari* |
| @MP | Menderagailua | *Sutan egon **arren** erretzen ez dena* |
| @PRT | Partikula | *Hantxe epeldu **ohi** zuen ohe purtzil bat gauero* |

## IV.8.3 Funtzio sintaktikoen esleipenerako erregelak.

CGren estrategia aurkezterakoan esan bezala —ikus § IV.4—, erregela sintaktikoen xedea hitz-forma bakoitza etiketa sintaktiko bakarrarekin uztea da. Hori erdiesteko erregela sintaktikoak ditugu, aurreko desanbiguatze-erregelen funtzionamendu berdina dutenak. Desberdintasuna izango da ezaugarri morfosintaktikoen artean erabaki beharrean, etiketa sintaktikoekin lan egiten dutela. Hala ere, erregela-multzo hauek, morfologikoak eta sintaktikoak, elkarrekin badute harremanik. Funtsezkoa izango baita desanbiguatze morfologikoan eginiko lana ondorengo urratsean funtzio sintaktiko zuzena hautatu ahal izateko. Aurreko puntuetan aurkeztu ditugu erabiliko diren funtzio sintaktikoen etiketak, eta orain zenbait erregela sintaktiko deskribatuko ditugu, hauek ebatzi beharreko anbiguotasun sintaktikoetariko batzuk azalduz:

- Kasu-markadunaren modifikadorea @KM> / funtzio sintaktiko nagusiak (@SUBJ, @OBJ,...). Adibidez, *a* organikodun izenen kasuan esango dugu ez direla izango kasu-markadunaren modifikadoreak baldin eta bere eskuinetara urrats batera izen bat kasu adlatiboan duen, eta ez duen adjektibo interpretaziorik posizio berean:

  ```
  REMOVE (@KM>)    IF     (0 AORG) (NOT 1 ADJ) (1 IZE-ALA);
  ```

  Adib.: *Bere ARIMA zerura airatu da*

- Izenaren determinatzailea @ID> / funtzio sintaktiko nagusiak (@SUBJ, @OBJ,...). Hau da, determinatzaileek ez baldin badute beren inguruan determina dezaketen burua, orduan beraiek hartuko dute funtzio sintaktiko nagusia. Gainontzeko kasuetan, bururen bat dutenean inguruan, determinatzaile funtzioa beteko dute. Adibidez, determinatzaile zehazgabe batek bere eskuinetara urrats batera aditz laguntzaile bat, aditz trinko bat edota aditz bat baldin badu, baldintza hau betez gero izenaren determinatzailearen funtzioa ezabatu egingo dugu. Honela adierazten dugu erregela bidez:

  ```
  REMOVE (@ID>) IF (0 DZG) (1 ADL/ADT OR ADI);
  ```

  Adib.: *Jaun HORIEK bazutela*

- Subjektu eta objektuaren arteko anbiguotasuna. Adibidez, absolutibo plurala subjektua dela esateko jartzen ditugun baldintzak hauexek dira: ez egotea NOR-NORK edo NOR-NORI-NORK laguntzailerik, eta hirugarren pertsona pluraleko "haiek" azaltzea NOR motako laguntzailean. Hau adierazten duen erregela, honela formulatzen dugu:

```
SELECT (@SUBJ) IF (0C ABS) (0 NUMP)
(NOT 0 ERG OR GEN OR PRT OR PART) (*1 NR_HK BARRIER KOMA)
(NOT *1 NOR_NORK OR NOR_NORI_NORK BARRIER KOMA);
```

Adib.: *TXORIAK tarrotuz gero hegaldatzen dira ohantzetik*

- Esaldi batean ezin dira bi subjektu egon jarraian, adibidez *Gorputza* (@SUBJ) *abaildua* (@SUBJ) *gelditu zitzaion*.. Kontua da, bigarren elementua predikatu gisa analizatuko dugula, baina morfologiatik partizipio deklinatua dela dugu, eta, absolutiboan dagoenez, objektu edota subjektu funtzioekin azaltzen zaigu. Beraz, planteatzen duguna da, subjektu bat baldin badugu izena dena, eta ondoren partizipio bat azaltzen baldin bazaigu subjektu legez, bigarren funtzioa kendu egingo dugu:

```
REMOVE (@SUBJ)   IF (0 PART) (-1 IZE + (@SUBJ));
```

Adib.: *Gorputza ABAILDUA gelditu zitzaion*

Bukatzeko, gogora dezagun azpikategorizazioaren inguruko zenbait puntu azaldu, CG formalismoaren aurkezpen orokorra egin, eta EUSMG deskribatu ditugula. Ondoren, hurrengo kapituluan hiztegiko adibideak aztertzeko landu dugun metodologiaz mintzatuko gara. Lehenago esan bezala, laugarren kapitulu honetan azaldurikoak metodologia horren garapena ulertzen lagunduko dutelakoan gaude.

# V. EHko aditzen adibideen azterketarako metodologia.

Bosgarren kapitulu hau, adibideak automatikoki aztertzeko erabili dugun metodologiaren (Arriola eta beste, 99) garapenari eskainiko diogu. Aurrera baino lehen, azpimarratu behar da kapitulu honetan azalduko dugun azterketa gehiago dela adibideak azpikategorizazioaren gaiari heldu ahal izateko moduan paratzeko eginiko lana, adibideetako aditzen azpikategorizazioaren azterketaren emaitzak baino. Hau da, adibideak aztertzeko prestatu ondoren —ikus § III.2, hiztegiaren prestatze-lanen deskribapenarekin erakutsi dugun legez— ja adibideen azterketari helduko zaio metodologia bat garatuz. Adibideak aztertzeko emaniko urrats nagusiak honako hauek dira: analisi morfologikoa, desanbiguatze morfologikoa, funtzio sintaktikoen esleipena, aditz-kate eta sintagmen osaketa, analizaturiko adibideak SGML-ez kodeturiko ezaugarri-egituren bidez errepresentatzea, eta, azkenik, ezaugarri-egitura horien gainean lan egiteko galdeketa-sistemaren garapena (ikus V.1 Irudia).

Bestalde, metodologiako azken bi urrats horien garrantzia ere nabarmendu nahi genuke. Alde batetik, analisiaren emaitzak ezaugarri-egituren bidez errepresentatzearena, eta bestetik, horren gainean garaturiko galdeketa-sistemaren diseinua. Errepresentazioari dagokionez, —§ V.1.7— deskribatzen dugu EUSMGko urrats desberdinen analisiaren emaitza errepresentatzeko hautaturiko bidea, hots, ezaugarri-egiturena eta analisietatik zer-nolako informazioa jasoko dugun. Errepresentazio-modu honek analisien emaitza testu huts izatetik errepresentazio aberatsago batera moldatzean, ustiapena errazteko bideak irekitzen ditu. Hain zuzen ere, ustiapen hori diseinatu dugun galdeketa-sistemaren bidez burutzen da. Hau da, galdeketa-sistema —ikus § V.1.8— honetan definitutako galderek analisiaren emaitzak aztertzen lagunduko digute. Urrats hauen bidez lorturiko emaitzak VI. kapituluan aurkeztuko ditugu.

Bukatzeko, landu dugun sintaxiaren atala TACAT tresnaren bidez erdiesten den errepresentazio sintaktikoarekin konparatu dugu.

## V.1 Hiztegiko aditzen adibideak aztertzeko erabilitako metodologiaz.

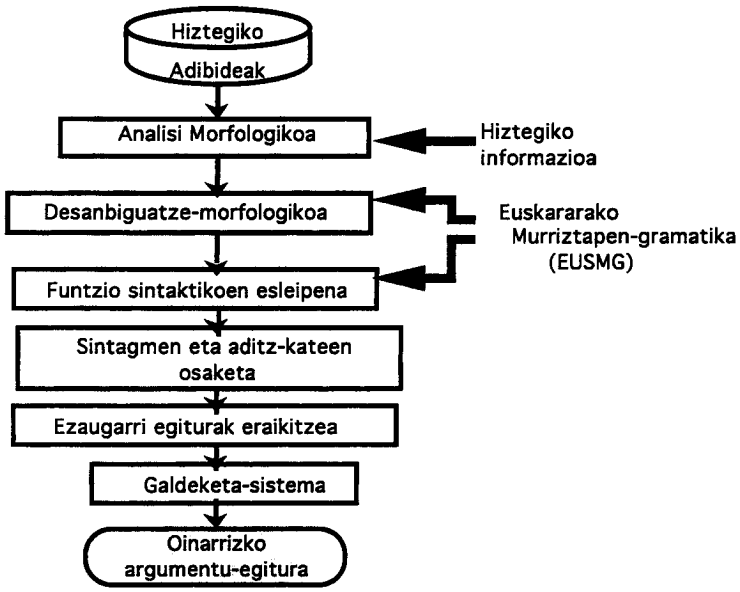
Azpikategorizazioa automatikoki lantzen duten zenbait proiektu deskribatzerakoan —ikus § IV.1—, ikusi dugu irtenbide nagusia corpusarena dela; hala ere, hiztegiko adibideak corpus berezitua diren heinean, gure proiektuak MRDtik abiatuz aurre egin nahi dio azpikategorizazioaren eskuratzeari, bai lehenengo kapituluan motibazioak —ikus § I.1—

azaltzerakoan emandako arrazoiengatik, eta baita ere hemendik jasotako informazioa corpus zabalagoetan azterketa gidatzeko baliagarria izango delakoan baikaude.

Bestalde, corpusetatik abiatzen diren proiektuetan analisi sintaktikorako bi joera nagusi azaldu dira: analizatzaile partzialak erabiltzearena eta analizatzaile orokorrez baliatzearena. Gure kasuan, bigarren bide honi jarraitu diogu, eta CG formalismoa (Karlsson eta beste, 95) baliatuko dugu euskararako Murriztapen-gramatika (EUSMG) lantzeko, formalismo honek morfologiaren gainean lan egiten baitu azaleko errepresentazio sintaktikoa erdiesteko. Eta esan beharrik ez dago hori zein ongi egokitzen zaion euskara bezalako hizkuntzei, non morfologia eta sintaxia hain loturik dauden. CGren bidezko analisiaren parte inportanteenetako bat desanbiguatze morfologikoa da, analisi morfologikotik irtendako emaitza anbiguoa tratatu ezagutza linguistikoan oinarritu murriztapenen bidez doitzeko helburuarekin, alegia. Beraz, soilik hiztegiko adibideetarako baliagarria den gramatika garatu eta analisirako teknika partzialak erabili beharrean, parser orokor bat erabiltzeko bidea hartu dugu, eta zeudenen artean CGz landu dugu gure gramatika euskararen ezaugarriei hobekien lotzen zelako formalismo honen estrategia —ikus § IV.7—.

Ildo beretik ari diren talde gehiago ere badira, analizatzaile orokor bat erabiltzen dutenak, beste batzuen artean: (Montemagni, Vanderwende 92) eta (Dolan, Vandewende, Richardson, 93).

Ondorengo irudian, gure metodologiaren urrats desberdinen eskema ikus daiteke:




**V.1 Irudia.-** Adibideak aztertzeko jarraituriko urratsak.

Aurkeztu dugun metodologiako 3. urratsa, desanbiguatze-lanarena, eta 4.ena, funtzio sintaktikoen esleipenari dagokiena, Euskararako Murriztapen-gramatika (EUSMG) baliatuz burutu diren bezalaxe, 5. urratsa ere, aditz-kateen eta sintagmen osaketari dagokiona, formalismo honen irteerako funtzio sintaktikoetan oinarrituko da. Bosgarren urrats horretan lantzen dugu bereziki azaleko sintaxia.

## V.1.1 Adibideak prestatu.

Adibideak prestatzeak, hiztegiko analisiari esker ezagutu ditugun aditzen adibideak aztertu ahal izateko behar ditugun eremuekin eta antolamenduarekin jasotzea esan nahi du. Hiztegiko sarrerei erreparatuz gero azpikategoria sarrera-mailan, edota adiera-ikurraren (adiera-ikurra, ñabardura, adiera-multzoa, etab.) ondorik agertuko zaigu. Sarrera-mailako azpikategoria lehenengo adieraren ondorik datozen adibideena dela esan dezakegu harik eta adiera-ikur baten ondorik azpikategoria desberdin bat azaldu arte. Irizpide horri jarraituz, adiera-ikur bakoitzari azpikategoria bat dagokio, eta halaxe jarriko dugu esplizituki adibideak jasotzerakoan. Ikus dezagun adibide baten bidez:

**adabatu, adaba, adabatzen**. du ad. (1808; adobatu 1571; cf. adabakitu 1832). Adabakia ipini; konpondu, antolatu. Soinekoak, abarkak adabatu. Zaku urratuak adabatzeko betarik ez zuelako. Ontzian sareak adobatzen ari zirela. Tresna urratua adabatu. II Bere galeoi adabatuari aingurak altxaturik. • G. er. Egin duen kaltea adabatu behar du.

**V.2 Irudia.-** EHko *adabatu* sarrera MRDn azaltzen den bezala.

MRDko sarreran ikusten denez, *du* laguntzaile-mota (guk azpikategoria deitu duguna) bakarrik sarrera-mailan azaltzen zaigu; hala ere, prestatze-urrats honetan adiera-ñabardura bakoitzari azpikategoria hori esleitzen diogu, desberdinik azaldu ez denez gero. Eta modu honetara esplizituki geratzen da adierazia adiera-ikurraren eta azpikategoriaren arteko lotura. Adiera-ñabardura horiek *@@Ad_string* deituriko eremuan adierazten ditugu *N1* eta *N2* ikurren bidez, adibide horiek bi ñabardura adierazten dituztela azalean adieraziz. Lotura hau azalerazteak begien bistan jartzen digu adibide bakoitza zein adiera-ikurri eta azpikategoriari dagokion, eta horrela azaltzea lagungarria dugu ondorengo azterketarako.

```
@@lema    adabatu, adaba, adabatzen.
@@Kategoria    ad.
@@Azpikategoria     du
@@Adibidea     Soinekoak, abarkak adabatu. Zaku urratuak
adabatzeko betarik ez zuelako. Ontzian sareak adobatzen ari
zirela. Tresna urratua adabatu.
@@Ad_string    N1.
@@Azpikategoria     du
@@Adibidea          Bere galeoi adabatuari aingurak altxaturik.
@@Ad_string    N2.
@@Azpikategoria     du
@@Adibidea          Egin duen kaltea adabatu behar du.
```

**V.3 Irudia.-** EHko *adabatu* sarrera analisi morfologirako prestatua.

Bestalde, esan beharra dago prestatze-lan hau erabat automatikoki egin dela. Hau da, ezagutu diren aditzak aztertzerakoan zer-nolako informazioa jasoko den, eta nola jasoko den zehaztu dugu programa batean. Aditz hauen adibide-eremuko adibideez gain, honelako informazioa jaso dugu aditz bakoitzerako: lema, kategoria, azpikategoria (laguntzaile-mota) eta adiera-ikurra. Adibidez, *kidetu*-ren aditz-sarrera honela jasotzen dugu:

```
@@lema    kidetu, kide edo kidetu, kidetzen.
@@Kategoria    ad.
@@Azpikategoria     du
@@Ad_string    A1.
@@Azpikategoria     du
@@Adibidea     Ez bila kidetzea zu baino goragokoekin.
@@Ad_string    A2.
@@Azpikategoria     du
@@Adibidea     Oinetakoak kidetu.
```

**V.4 Irudia.-** EHko *kidetu* sarrera analisi morfologirako prestatua.

Lehenago aipatu bezala, adiera-ikurrak ezinbestean hartu behar izan ditugu kontuan adibideak antolatzeko, hiztegian azpikategoria-etiketak adiera-ikur, azpiadiera, edota ñabarduraren arabera alda daitezke eta. Aurreko adibidean (ikus V.3 Irudia), adieraren eremuan, hots, @@*Ad_string* delakoan ñabardurak jasotzen genituen *N1*, *N2* ikurren bidez, eta oraingoan (V.4 Irudian) adierak jasotzen ditugu *A1* eta *A2* ikurren bidez. Horrez gain, aurrekoan bezalaxe sarrera-mailan agertzen den *du* laguntzaile-mota adiera-ikur horiekin lotzen dugu. Modu honetara, adibide bakoitza adiera-ikur eta azpikategoria jakin batekin lotuta

dagoela azalerazten dugu. Bi ikur horien arteko lotura esplizitu egitea, adierak zer-nolako eragina duen argumentu-egitura osatzerakoan aztertzeko baliagarria izan daiteke.

Esan beharra dago hiztegiak eskaintzen digun informazioak mugatzen gaituela, adibidez irizpide semantikoen arabera hautatu nahi bagenitu adibideak ez genuke horretarako kode semantiko lagungarririk izango.

## V.1.2 Analisi morfologikoa.

Analizatzaile morfologikoaren bidez aditzen adibideak morfologikogi analizatu ditugu. Oinarrian analizatzaile morfologikoa izango dugu, eta honekin batera lexikorik gabeko analisiak egiteko aukera ere, eta hori garrantzitsua da gure azterketarako adibideetan hitz ezezagunak, hau da, lexikoan ez dauzkagunak, ager daitezke eta. Euskal morfologiaren tratamendu automatikoa gauzatu ahal izateko (Alegria, 96; Urkia, 97), bi mailatako morfologian oinarritutako analizatzaile morfologikoa dugu. Zehaztu behar da, analizatzaile baino segmentatzaile edo zatikatzaile morfologikoa dugula. Hau da, testu-hitza osatzen duten elementu guztiak banatu eta bakoitzari dagokion informazio morfologikoa erakusten du segmentatzaileak, EDBLtik zuzenean hartuta. Analizatzaile morfosintaktikoaren izena, Morfeus, segmentatzaile morfologikotik datorkion informazio gordina landuko duen tresnarentzat utziz, (Aduriz eta beste, 1999). Analisiaren lehen urrats bezala, zatikatzaile morfologikoaz baliatuz adibideen analisi morfologikoa burutuko da. Horrela bada, hitz ororen zatiketa morfologikoak izango ditugu, zatiketa horren ondorioz sortzen diren lema eta morfema desberdinei dagokien informazioa erantsiko zaie.

Aitzitik, V.4 irudian erakutsi dugun *kidetu*-ren aditz-sarrera, oraingo honetan zatikatzaile morfologikotik pasa ondoren lorturiko emaitza dugu ondorengo adibidean[1]:

```
/<@@lema    kidetu, kide edo kidetu, kidetzen. >/<ID>/
/<@@Kategoria     ad. >/<ID>/
/<@@Azpikategoria       du>/<ID>/
/<@@Ad_string     A1.>/<ID>/
/<@@Azpikategoira       du>/<ID>/
/<@@Adibidea>/<ID>/
```

---

1 Analisian azaltzen diren laburduren esanahia: ABS: absolutiboa; ADB: adberbioa; ADI: aditza; ADIZE: aditz-izena; ADJ: adjektiboa; ADOARR: aditzondo arrunta; ADOIN: aditzoina; ALGARR: adizlagun arrunta; AMM: aspektu-mota morfema; AORG: *a* organikoa; ARR: arrunta; ASP: aspektua; ATZ: atzizkia; BURU: burutua; DEK: deklinabide-morfema; DU: *du* laguntzaile-mota; EGI: egiatasunezkoa; ELI: elipsia; ERG: ergatiboa; ERL: erlazio atzizkia; GEL: genitibo leku-denborazkoa; GRA: graduatzailea; HAS_MAI: maiuskulaz hasten den hitza; IOR: izenordea; IZE: izena; IZO: izenondoa; JNT: juntagailua; KONP: konpletiboa; LOT: lokailua; MEN: menderagailua; MG: mugagabea; MUGM: mugatua; NOTDEK: deklinabide-morfemarik gabeko interpretazioa; NUMP: numero plurala; NUMS: numero singularra; PART: partizipioa; PERARR: izenordain pertsonal arrunta; PRT: partikula; SIN: sinplea; SOZ: soziatiboa; TE_TZE: aditz-izenei esleitzen zaien etiketa; ZERO: kasurik hartzen ez duen interpretazioa; ZU: bigarren pertsona singularra; @-JADLAG_MP: aditz laguntzaile ezjokatu / mendeko perpausa; @-JADNAG_MP: aditz nagusi ezjokatu / mendeko perpausa; @<IZLG: eskuineko izenlaguna; @ADLG: adizlaguna; @ATRIB: atributua; @IZLG>: ezkerreko izenlaguna; @OBJ: objektua; @PJ: koordinazio juntagailua; @SUBJ: subjektua.

```
"<Ez>"
  "ez" IZE ARR DEK ABS MG @OBJ @SUBJ HAS_MAI
  "ez" IZE ARR ZERO HAS_MAI
  "ez" PRT EGI HAS_MAI
"<bila>"
  "bila" ADB ADOARR AORG
  "bilatu" ADI SIN AMM ADOIN NOTDEK
"<kidetzea>"
  "kidetu" ADI SIN AMM ADIZE DEK ABS NUMS MUGM @OBJ @SUBJ @ATRIB DU
  "kidetu" ADI SIN AMM ADOIN ERL MEN KONP @-JADNAG_MP DU NOTDEK
  "kidetu+tze" ADI SIN ATZ IZE ARR DEK ABS NUMS MUGM @OBJ @SUBJ @ATRIB
  TE_TZE
"<zu>"
  "zu" IOR PERARR NUMS ZU DEK ABS MG @OBJ @SUBJ
"<baino>"
  "baino" LOT JNT @PJ
"<goragokoekin>"
  "gora" ADB ALGARR GRA KONP DEK GEL @IZLG> @<IZLG @ADLG DEK SOZ NUMP MUGM
  @ADLG AORG
  "gora" ADB ALGARR GRA KONP DEK GEL @IZLG> @<IZLG @ADLG ELI DEK SOZ NUMP
  MUGM  @ADLG AORG
  "gora" ADJ IZO GRA KONP DEK NUMS MUGM DEK GEL @IZLG> @<IZLG @ADLG DEK
  SOZ NUMP MUGM @ADLG AORG
  "gora" ADJ IZO GRA KONP DEK NUMS MUGM DEK GEL @IZLG> @<IZLG @ADLG ELI
  DEK SOZ NUMP MUGM @ADLG AORG
"<$.>"
  PUNT_PUNT
/<@@Ad_string     A2.>/<ID>/
/<@@Azpikat       du>/<ID>/
/<@@Adibidea>/<ID>/
"<Oinetakoak>"
  "oin" IZE ARR DEK MG DEK GEL @IZLG> @<IZLG @ADLG DEK ABS NUMP MUGM @OBJ
  @SUBJ  @ATRIB  HAS_MAI
  "oin" IZE ARR DEK MG DEK GEL @IZLG> @<IZLG @ADLG DEK ERG NUMS MUGM @SUBJ
  HAS_MAI
  "oin" IZE ARR DEK MG DEK GEL @IZLG> @<IZLG @ADLG ELI DEK ABS NUMP MUGM
  @OBJ @SUBJ  @ATRIB HAS_MAI
  "oin" IZE ARR DEK MG DEK GEL @IZLG> @<IZLG @ADLG ELI DEK ERG NUMS MUGM
  @SUBJ HAS_MAI
  "oin" IZE ARR DEK NUMP MUGM DEK GEL @IZLG> @<IZLG @ADLG DEK ABS NUMP
  MUGM @OBJ @SUBJ @ATRIB HAS_MAI
  oin" IZE ARR DEK NUMP MUGM DEK GEL @IZLG> @<IZLG @ADLG DEK ERG NUMS MUGM
  @SUBJ HAS_MAI
  "oin" IZE ARR DEK NUMP MUGM DEK GEL @IZLG> @<IZLG @ADLG ELI DEK ABS NUMP
  MUGM  @OBJ @SUBJ @ATRIB HAS_MAI
  "oin" IZE ARR DEK NUMP MUGM DEK GEL @IZLG> @<IZLG @ADLG ELI DEK ERG NUMS
  MUGM  @SUBJ HAS_MAI
  "oinetako" IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @ATRIB HAS_MAI
  "oinetako" IZE ARR DEK ERG NUMS MUGM @SUBJ HAS_MAI
"<kidetu>"
  "kidetu" ADI SIN AMM PART ASP BURU DU NOTDEK
  "kidetu" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ DU
  "kidetu" ADI SIN AMM PART DU NOTDEK
"<$.>"
  PUNT_PUNT
```

## V.1.3 Hiztegian oinarritutako desanbiguazioa.

Hiztegiko informazioa baliatu dugu hitzen interpretazioei hiztegitik bertatik jasoriko informazioa esleituz, EUSMGri bidea errazteko bai desanbiguazio morfologikoan bai funtzio sintaktikoenean:

a) Aditzek dituzten interpretazioei, aditzen adibideetako laguntzaile-mota etiketa gehitu (da, du, da-du, ...). Adibidez, *kilimatu*-ren sarrerako azpikategoria-etiketa, da-du, *kilimatu* aditza agertzen deneko analisi morfologikoaren lerroetan erantsi dugu, eta errazago ikusteko letra lodiz nabarmendu dugu:

```
/<@@lema    kilimatu, kilima edo kilimatu, kilimatzen. >/<ID>/[2]
/<@@Kategoria     ad. >/<ID>/
/<@@Azpikategoria       da-du>/<ID>/
/<@@Adibidea>/<ID>/
"<Nori>"
  "nor" IOR IZGGAL DEK DAT MG @ZOBJ HAS_MAI
"<,>"
  PUNT_KOMA
"<haurtzaroko>"
  "haurtzaro" IZE ARR DEK NUMS MUGM DEK GEL @IZLG> @<IZLG @ADLG DEK ABS MG
  @OBJ @SUBJ
  "haurtzaro" IZE ARR DEK NUMS MUGM DEK GEL @IZLG> @<IZLG @ADLG
"<oroipenek>"
  "oroipen" IZE ARR DEK ERG MG @SUBJ
  "oroipen" IZE ARR DEK ERG NUMP MUGM @SUBJ
"<bihotza>"
  "bihotz" IZE ARR DEK ABS NUMS MUGM @OBJ @SUBJ @ATRIB
"<kilimatu>"
  "kilimatu" ADI SIN AMM PART ASP BURU DA-DU NOTDEK
  "kilimatu" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ DA-DU
  "kilimatu" ADI SIN AMM PART DA-DU NOTDEK
"<ez>"
  "ez" IZE ARR DEK ABS MG @OBJ @SUBJ
  "ez" IZE ARR ZERO
  "ez" PRT EGI
"<diote>"
  "*edun" ADL A1 NOR_NORI_NORK NR_HU NI_HU NK_HK
  "*edun" ADT A1 NOR_NORI_NORK NR_HU NI_HU NK_HK
  "*io" ADT A1 NOR_NORK NR_HU NK_HK
"<$?>"
  PUNT_GALD
```

[2] Analisian azaltzen diren laburduren esanahia: A1: indikatibozko orainaldia; ABS: absolutiboa; ADI: aditza; ADL: aditz laguntzailea; ADT: aditz trinkoa; AMM: aditz-mota morfema; ARR: arrunta; ASP: aspektua; BURU: burutua; DA-DU: *da-du* laguntzailea; DAT: datiboa; DEK: deklinabide-morfema; EGI: egiatasunezkoa; ERG:ergatiboa; GEL: genitibo leku-denborazkoa; HAS_MAI: maiuskulaz hasten den hitza; IOR: izenordea; IZE: izena; IZGGAL: izenordain galdetzailea; MG: mugagabea; MUGM: mugatua; NI_HU: nori 3. pertsona singularra; NK_HK: nork 3. pertsona plurala; NOR_NORI_NORK: nor-nori-nork; NOR_NORK: nor-nork; NOTDEK: deklinabide kasurik ez duen interpretazioa; NR_HU: nor 3. pertsona singularra; NUMP: plurala; NUMS: singularra; PART: partizipioa; PRT: partikula; SIN: sinplea; ZERO: kasurik hartzen ez duen interpretazioa; @<IZLG: eskuineko izenlaguna; @ADLG: adizlaguna; @ATRIB: atributu; @IZLG>: ezkerreko izenlaguna; @OBJ: objektua; @SUBJ: subjektua; @ZOBJ: zehar objektua.

b) Aditzek dituzten interpretazioetarik zenbait ez dira egokiak izango edota aukera gutxiago dutenak dira: berez aditz bati dagokion adibidea izanik, aditz bezala agertu behar baita eta ez adjektibo bezala, esate baterako (posible dena partizipioen kasuan). Orduan aukera gutxiko interpretazioei hori adierazten duen etiketa esleituko zaio, hots, *aukera gutxiko interpretazioa* (AGI). Adibidez, *kirastu*, *kirats edo kirastu*, *kirasten* artikuluaren adibideetan *kirastu* aurkitzen badugu bere interpretazio-multzoan aditza ez den interpretazio batekin, orduan lerro horri AGI etiketa esleituko zaio:

```
/<@@lema    kirastu, kirats edo kirastu, kirasten. >/<ID>/[3]
/<@@Kategoria     ad. >/<ID>/
/<@@Azpikategoria       du>/<ID>/
/<@@Ad_string     A1.>/<ID>/
/<@@Azpikat       du>/<ID>/
/<@@Adibidea>/<ID>/
"<Urak>"
  "ur" IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ @ATRIB HAS_MAI
  "ur" IZE ARR DEK ERG NUMS MUGM @SUBJ HAS_MAI
"<kirastu>"
  "kirastu" ADI SIN AMM PART ASP BURU DU NOTDEK
  "kirastu" ADI SIN AMM PART DEK ABS MG @OBJ @SUBJ DU
  "kirastu" ADI SIN AMM PART DU NOTDEK
  "kirastu" ADJ IZO DEK ABS MG @OBJ @SUBJ AGI
  "kirastu" ADJ IZO AGI
"<ziren>"
  "izan" ADL B1 NOR NR_HK ERL MEN @+JADLAG_MP
  "izan" ADL B1 NOR NR_HK ERL MEN ERLT @+JADLAG_IZLG>
  "izan" ADL B1 NOR NR_HK ERL MEN ZHG @+JADLAG_MP
  "izan" ADL B1 NOR NR_HK
  "izan" ADT B1 NOR NR_HK ERL MEN @+JADNAG_MP
  "izan" ADT B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG>
  "izan" ADT B1 NOR NR_HK ERL MEN ZHG @+JADNAG_MP
  "izan" ADT B1 NOR NR_HK
  "zira" IZE ARR DEK GEN NUMP MUGM @IZLG> @<IZLG DEK ABS MG @OBJ @SUBJ
  AORG
  "zira" IZE ARR DEK GEN NUMP MUGM @IZLG> @<IZLG AORG
"<$.>"
  PUNT_PUNT
```

Aipaturiko etiketa horiek desanbiguatzearen urratsean baliagarriak dira. Laguntzaile-moten kasuan, erabilgarriak dira komunztadura dela medio kasu-anbiguotasuna ebazten laguntzeko. Eta AGI etiketa erabilia izango da gainontzeko informazio linguistikoa nahikoa ez denean anbiguotasunari aurre egiteko, orduan, etiketa hori duen interpretazioa ezabatu egingo dugu. Azken batean, nolabaiteko heuristiko gisa baliatzen dugu azken etiketa hau.

---

[3] Analisian azaltzen diren laburduren esanahia: ABS: absolutiboa; ADI: aditza; ADJ: adjektiboa; ADL: aditz laguntzailea; ADT: aditz trinkoa; AORG: *a* organikodun hitza; ARR: arrunta; ASP: aspektua; B1: indikatibozko lehenaldia; BURU: burutua; DEK: deklinabide-morfema; DU: *du* laguntzailea; ERG: ergatiboa; ERL: erlazio atzizkia; ERLT: erlatibozkoa; GEN: genitiboa; HAS_MAI: maiuskulaz hasten den hitza; IZE: izena; IZO: izenordea; MEN: menderagailua; MG: mugagabea; MUGM: mugatua; NOR: *nor* sailekoa; NOTDEK: deklinabide-morfemarik gabeko interpretazioa; NR_HK: nor 3. pertona plurala; NUMP: plurala; NUMS: singularra; SIN: sinplea; ZHG: zehargaldera; @<IZLG: eskuineko izenlaguna; @ATRIB: atributua; @IZLG>: ezkerreko izenlaguna; @OBJ: objektua; @SUBJ: subjektua; @+JADLAG_IZLG>: aditz laguntzaile jokatua/ ezkerreko izenlaguna; @+JADLAG_MP: aditz laguntzaile jokatua / mendeko perpausa; @+JADNAG_IZLG>: aditz nagusi jokatua / ezkerreko izenlaguna; @+JADNAG_MP: aditz nagusi jokatua / mendeko perpausa.

## V.1.4 Desanbiguatze morfosintaktikoa EUSMG baliatuz.

Urrats honetan, EUSMGn garatutako desanbiguatze-erregelak (ikus C eranskina) aplikatuko ditugu hiztegiko adibideen anbiguotasun morfosintaktikoari aurre egiteko. Oro har, aditzen adibideetan 3,7 interpretazio ditugu hitzeko, eta informazio morfosintaktiko guztia kontuan hartuta (etiketa sintaktikoak[4] ezik) %60 desanbiguatzen dugu murriztapen-gramatikaren bidez. Hau da, erregelak aplikatu ondoren %77 ez da anbiguo izango, desanbiguatzen dugunari, hasieratik anbiguo ez direnen kopurua gehitu behar baitzaio.

Anbiguotasun-tasa jaisteaz aparte, esan behar da aukerak baztertzerakoan analisi zuzenak ken daitezkeela. Gure kasuan, aditz-kateen inguruan dauden sintagmak ezagutzeko xedea dugunez, errore-tasaren neurketa lortu nahi dugun egitura bakoitzeko egingo dugu —ikus § V.1.6.4—.

Aurrerago esan bezala, urrats honetan hiztegian oinarritutako desanbiguazioan esleitutako etiketak, laguntzaile-motarenak (DA, DA-DU, ...) eta *aukera gutxiko interpretazioarenak* (AGI) lagungarriak ditugu. Hau da, etiketa horiek baliatuko ditugu desanbiguatzeari ekiterakoan. Alde batetik, laguntzaile-mota etiketak oso erabilgarriak dira aditz batek *Nor* edo *Nork* motako subjektua duen jakiterakoan.

Bestalde, *aukera gutxiko interpretazioa* (AGI) etiketa baliatuko dugu testuinguruaren arabera informazio nahikoa ez dagoenean hitz baten interpretazioen artean erabakitzeko.

## V.1.5 Funtzio sintaktikoen esleipena EUSMG baliatuz.

Desanbiguatze morfosintaktikoa egiten den legez, hitz bakoitzak dituen funtzio sintaktiko desberdinak erregela sintaktikoen (ikus C eranskina) bidez desanbiguatuko dira. Desanbiguatze sintaktikoa arrakastatsua baldin bada hitz orok funtzio sintaktiko bakarra izango du. Analisi sintaktikoak, soilik, hitzen arteko harremanak lantzen dituenez, dependentzia-analisia gisa ezagutzen da. Hona hemen, Abney-k (1997:129) zer dioen:

> "Word are not explicitly associated with their governors, but the syntactic-function annotations significantly constrain the set of compatible analyses, and can be seen as representing an ambiguity class of analyses."

Bai desanbiguatze morfosintaktikoa, bai funtzio sintaktikoen esleipena oso garrantzitsuak dira ondorengo urratsean burutuko dugun aditz-kateen eta sintagmen osaketarako.

Aurrerago esan bezala, informazio morfosintaktiko guztia kontuan hartuta (etiketa sintaktikoak izan ezik) murriztapen-erregelak aplikatu ondoren %23ko anbiguotasun-tasa

---

[4] Anbiguotasuna zenbatean jaisten den neurtzerakoan informazio morfosintaktikoa hartu dugu kontuan. Hala ere, esan behar da, desanbigutze sintaktikoa ere gauzatu dugula. Hau da, murriztapen sintaktikoak ere aplikatu ditugu.

izango dugu. Ondorengo puntuan (V.1.6) ikus daitekeenez, anbiguotasun sintaktiko orokorra neurtu beharrean funtzio sintaktiko nagusien eta modifikadoreen arteko anbiguotasuna neurtu dugu. Neurketa hori egiterakoan hitz-forma bakoitzeko interpretazio-multzoa izan dugu kontuan. Anbiguotasun sintaktiko orokorra neurtzeko anbiguotasun sintaktikoaren ikerketa sakonagoa egitea mereziko luke.

## V.1.6 Aditz-kateen eta sintagmen osaketa EUSMGko analisitik abiatuz.

Murriztapen-gramatikaren bidez lorturiko egitura sintaktikoa laua eta azalekoa izateaz gain, ez du sintagma noziorik esplizituki adierazten, ez dago egitura sintagmatikorik. Formalismo honetan ez dago perpaus-egiturarik. Deskripzio-lengoaia ona da behe mailako egitate sintaktikoak adierazteko: hitzen ordena lineal absolutua (2 hitz aurrerago, 7 hitz atzera), hitzen ordena erlatiboa (Xren ondoren nonbait ...; X eta Y hitzen artean ...), hurrengo ISren buruan adierazi behar izaten da formalismo honetan. Lehenengo egitatetik hirugarrenera zailtasuna areagotzen da. Esate baterako, subjektuak ezagutzeko CGn erabiltzen den estrategiari erreparatuz gero, esan dezakegu teorikoki aszetikoagoa dela, zeren eta hitza, hitz-klasea, juxtu aurretik, ... bezalako kontzeptuak erabiltzen dira, ez IS bezalakoak. Argi eta garbi, behe mailako egitateak deskribatzeko formalismoa dugu. Honen arrazoi nagusiak bi aldetakoak ditugu:

1) Praktikoa: jatorrian, LNPko aplikazio praktiko baterako garatua delako, informazioaren berreskurapenerako. Horrekin batera, corpus handien analisi eta desanbiguazioa ditu helburu. Sendotasuna eta eraginkortasuna dira bere alde onak, deskripzio-egokitasuna baino. Ikuspegi konputazionaletik hizkuntzalariari "deskribatzeko askatasun" handiegia ematea ez litzatekeela komeni uste dute formalismo honen garatzaileek.

2) Teorikoa: desanbiguazio eta analisiari begira, kontzeptu linguistiko asko baliagarri ote direnaren eszeptizismo sakona dute (Karlsson, 1990). Kontzeptu sofistikatuz beteriko gramatika ez da egokia benetako testuak (ez "laborategiko esaldiak" bakarrik) analizatzeko.

   CGren arrakastaren gakoa, goi mailako kontzeptu teorikorik ezean datza. Behe mailako datu asko baliatzen ditu: hitz-formak, kategoriak, ezaugarri morfologikoen funtzioak, etab.

Egitura sintagmatikorik adierazita ez egon arren, deskribapen linguistiko horretan inplizituki daude adieraziak elementuen arteko harremanak, eta informazio horretan oinarrituz aditz-kateak eta gainontzeko sintagmak eraiki ditzakegu. Horretarako, hitz-forma ororen anbiguotasun morfosintaktikoa zein sintaktikoa ebatzita egotea komeni da, kateen osaketa *garatu ahal izateko*.

Beraz, urrats honetan egin nahi dena da, funtsean inplizituki dauden harreman sintaktikoak esplizitu egitea. Horretarako, aditz-kateen osaketarako erregelak definitu ditugu lehendabizi, eta aditz-kateak bereiztu ondorik, hauen inguruan dauden sintagmak markatu ditugu batik bat.

Aditz-kateak zein sintagmak definitzeko murriztapen sintaktikoen aplikazioaren ondorik geratzen diren etiketa sintaktikoetan oinarritu gara. Etiketa sintaktiko nagusi eta modifikadoreen arteko bereizketan dago gakoa urrats honi ekiterakoan. Esaldia hitzez osaturiko katea dugu, eta kate hori osatzen duten hitzen arteko dependentzia-harremanak inplizituki adierazita daude etiketa sintaktiko horiek direla medio. Kontua da harreman horiek azaleratzea. Jakina, horretarako garrantzitsua da oso, funtzio sintaktiko nagusi eta modifikadoreei erreparatuz dagoen anbiguotasuna txikia izatea. Horrela bada, 84.963 hitzetik, 1.145 hitzek izango dute modifikadoreari dagokion etiketaren bat eta nagusiari dagokion etiketaren bat. Hau da, %98.65 ez da anbiguo aditz-kateen eta sintagmen osaketarako darabilgun irizpidearekiko. Nahiz eta, desanbiguatze morfologikoaren ikuspegitik, zenbait hitz-forma anbiguo izan, funtzio sintaktiko nagusi bi baldin badituzte edota bi modifikadore, guretzat urrats honetan ez dira anbiguo izango.

Horrez gain, azaleko sintaxia garatzerakoan gure azterketaren arrakastaren alde izan ditugun bestelako ezaugarrien artean, honako hauek azpimarratuko genituzke:

- esaldien luzera (6,44 hitz ditugu esaldiko),beraz, esaldi ezkonplexuak espero ditugu;

- eta hiztegian azaltzen den laguntzaile-mota txertatu dugula aditzen analisi morfologikoan —§ V.1.3—. Kontua da, aditz ezjokatuen adibideetan analisi zuzena egiteko ez dugula laguntzaileak aditz jokatuen kasuan ematen duen laguntzarik. Horren ordez laguntzaile-mota etiketaz baliatzen gara. Esate baterako, *Bere salgaiak kalez kale aldarrikatuz* adibidean, *Bere salgaiak* objektu gisa hautatzeko, hiztegitik jasotako *du* etiketa baliatzen dugu. Horrela, ikus dezakegu zein baliagarria den horrelako informazioa hastapeneko azaleko analisi sintaktikoan aurrera egiteko.

Ikus ditzagun hurrengo puntuetan aditz-kateak eta sintagmak definitzeko erabilitako irizpide nagusiak.

### V.1.6.1 Aditz-kateak.

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. Aditz-kate jarraiak markatzeko erabili ditugun etiketak hauek dira:

- %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-kateari esleitzen diogun etiketa.

Hona aditz-kate jarrai baten adibidea:

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

Esan bezala, badira elementu bakarreko aditz-kateak ere, eta beraz kasu hauetan etiketa bakarra izango dute: %ADIKAT, honela adierazten dugu aditz-kate baten aurrean gaudela. Adibidez:

*Egiazko tresora lan segituan* *datzala*

**%ADIKAT**

Aditz-kate ezjarraiei dagozkien etiketak hauek dira:

- %ADIKATHAS: aditz-kate jarraiei jartzen zaien etiketa bera erabiliko dugu aditz-kate ezjarraien lehenengo elementua markatzeko.

- ADIKATETEN: aditz-kate ezjarrai baten bigarren osagaiari esleitzen diogun etiketa.

- ADIKATETENBU: aditz-kate ezjarrai baten azken elementuari esleitzen diogun etiketa.

Hona hemen, aditz-kate ezjarrai baten adibidea:

```
"<Ez>"
  "ez" PRT EGI HAS_MAI %ADIKATHAS
"<dio>"
  "*edun" ADL A1 NOR_NORI_NORK NR_HU NI_HU NK_HU @+JADLAG %ADIKATETEN
"<maitasun>"
  "maitasun" IZE ARR ZERO @KM>
"<horrek>"
  "hori" DET ERKARR NUMS DEK ERG NUMS MUGM @SUBJ
"<gogoa>"
  "gogo" IZE ARR DEK ABS NUMS MUGM @OBJ
"<mehartu>"
  "mehartu" ADI SIN AMM PART ASP BURU DA-DU NOTDEK @-JADNAG
  %ADIKATETENBU
  "mehartu" ADI SIN AMM PART DA-DU NOTDEK @-JADNAG %ADIKATETENBU
"<batere>"
  "batere" ADB ADOARR @ADLG
```

```
  "batere" IOR IZGMGB
"<$.>"
```

Aipatu etiketok esleituz, aditz-kate jarrai nahiz ezjarrai bat esplizituki markatzeko murriztapen-gramatika batzuez baliatu gara (gramatika horiek guztiak B eranskinean ikus daitezke).

Ondoren, adibide gisa, aditz-kate jarraiak identifikatzeko hiru erregela erakutsiko ditugu:

- Aditz-kate hasiera markatzeko mapaketa-erregela:

```
MAP (%ADIKATHAS) TARGET EZ_OSAG
IF (NOT -1 ADIKAT_NAG_EZJOK OR ADIKAT_OSAG)
(1 ADITZAK OR ADIKAT_PART);
```

Adib. ***Ez** sartu hemen ...*; ***Ez** dut egin*; ***Ez** du ogirik*; ***Ez** du nahi*; ***Ez** omen doa.*

Mapaketa-erregela honen bidez hauxe esaten da: markatu %ADIKATHAS etiketarekin *ez* partikula, baldin eta bere ezkerretara urrats batera ez dagoen ez aditz nagusi ezjokaturik ez aditz lokuzioetako osagai izan daitekeen elementurik, adib. *behar*, *nahi*, etab.; eta eskuinetara urrats batera aditz, aditz laguntzaile edo aditz trinko kategoriako elementuren bat dagoen, edota hauetako partikularen bat: *omen*, *ote*, *ohi*, *ahal*, *al*, *baldin*.

- Aditz-kate bukaera markatzeko mapaketa-erregela:

```
MAP (%ADIKATBU) TARGET ADIKAT_LAG
IF (NOT -1 ADIKAT_OSAG OR EZ_OSAG)
(-1 ADIKAT_NAG_EZJOK OR ADIKAT_PART OR EZ_OSAG);
```

Adib. *alboratu **balitz**; ... egon ohi **zela***

Mapaketa-erregela honen bidez zera adierazten da: %ADIKATBU etiketa jarri aditz laguntzaileari, baldin eta, ez duen urrats batera bere ezkerretara aditz lokuzioetako osagai izan daitekeen osagairik edota *ez* partikula; eta bere ezkerretara pauso batera aditz nagusi ezjokatu, *ez* partikula edota *omen*, *ote*, *ohi*, *ahal*, *al*, *baldin* moduko partikularen bat izatea.

- Elementu bakarreko aditz-katea markatzeko mapaketa-erregela:

```
MAP (%ADIKAT) TARGET ADIKAT_MP_EZJOK
IF (0 ADIKAT_MP_EZ_JOK) (NOT -2 ADIKAT_OSAG)
(NOT -1 EZ_OSAG OR ADIKAT_NAG_EZJOK)(NOT 1 ADIKAT_OSAG);
```

Adib. ...**entzunik** adoretu zen

Azken adibide honetako erregelak honako zeregin hau du: elementu bakarreko aditz-kate etiketa %ADIKAT, esleituko dio aditz ezjokatu mendeko bati, baldin eta, ezkerretara urrats bitara aditz lokuzioetako osagai izan daitekeen elementurik ez duen, norabide berean baina urrats batera ez duen *ez* partikula edota aditz nagusi ezjokaturik; eta eskuinetara, urrats batera, ez du izan behar aditz lokuzioetako osagai izan daitekeen elementurik.

Horrela bada, adibide batean agertzen diren aditz-kate desberdinak ezagutzeko gai izango gara, aditz-kate bakoitza non hasten den eta non bukatzen den adieraziz. Garbi gera bedi, urrats honetan aditz-kateak markatu baino ez dugula egiten. Ondorengo urratsean, kateak osorik jasotzerakoan ezinbestekoak ditugu aditz-kateen eta sintagmen osaketa deitu dugun urrats honetan jarririko aditz-kate etiketak. Hain zuzen ere, hauei esker dakigu aditz-kate bat non hasi eta bukatzen den, edota elementu bakarreko aditz-katea den. Beraz, hurrengo urratsean kate hauei unitate gisa egin diezaiekegu erreferentzia, eta hauei bere osotasunean dagokien informazioa jasoko dugu.

Aditz-kateen artean, izen/adjektibo + izan/ukan aditz-esapideak (*behar izan*, *nahi izan*, *bizi izan*, *balio izan*, *maite izan*, *komeni izan*, *merezi izan* ...) ditugu, eta aditz-lokuzio hauetako analisia ematerakoan esan beharra dago *behar dut* bezalako aditz lokuzioetako *dut* aditz trinko legez analizatu dugula. Eta hori horrela egin dugu, indikatibozko zein subjuntibozko eta aginterako formetan. Horrela bada, *balio lezake*, *hala bedi*, *uste dezagun* eta honelakoetan *edin* eta *ezan* aditzak *izan* eta *edun* trinkoen supleziozko formatzat har daitezke. Aditz-lokuzioak kate bezala hartzeko, horietako osagai izan daitezkeenen multzoa zehaztu dugu, hauetariko bakoitza dagokion kategoriarekin. Hona hemen bildu ditugun aditz-lokuzioetako osagaietariko batzuk: (ageri ADJ ABS MG), (aiher IZE ABS MG), (balio IZE ABS MG), (behar IZE ABS MG), (bizi ADJ ABS MG), (falta IZE ABS MG). Jasotzen ditugun osagai guztiak B eranskinean ikus daitezke.

Hala ere, esan beharra dago *lo egin*, *negar egin*, ... modukoak ez ditugula aditz-kate gisa ezagutzen. Honelakoetan, aurreko osagaiari absolutibo mugagabe analisia eman diogu; adibidez, *lo* izen absolutibo mugagabe legez analizatuko dugu *lo egin* moduko aditz konposatuetan azaltzen zaigunean.

Ezagutzen ditugun aditz-kateak ilustratzeko, horien eskematzat har daitezkeen hogei eredu aurkeztuko ditugu. Erregela tankerako eredu hauek lagungarriak suertatu dira aditz-kateak markatzeko mapaketa-erregelak idazterakoan. Eredu horiek zerrendatu aurretik, zenbait xehetasun egin beharra dago:

- Denak elementu terminalak dira.
- /: aukera adierazteko erabili dugun ikurra.

- 0: zero zenbakia, elementurik ez dela agertuko adierazteko erabili dugu.
- *: edozein elementu ager daitekeela adierazteko. Halako ikurra azaltzen deneko adibideak aditz-kate ezjarraienak ditugu.

Hona hemen jarraian ezagutzen ditugun kateak, aditz-funtzio etiketez, eta aditz-kateetako parte izan daitezkeen zenbait osagaiz osatuak:

1. @-JADNAG_MP / @+JADNAG_MP

*Izerditzea ona dela osasunerako / Egiazko tresora lan segituan datzala*

2. @-JADNAG (@+JADLAG / @+JADLAG_MP)

*Asko belutu barik laster hasiko zela diru-banatzea*

*Bere inozotasuna aurpegiratu zion*

3. @-JADNAG (BEHAR / NAHI/0) (@+JADNAG / @+JADNAG_MP)

*Erresumako haur guztiei frantsesa ikaserazi behar zietela*

*Herriko guztiek bideak adelatu behar dituzte, urtean bi aldiz*

*Betiko bizitza iritsi nahi baduzue*

4. @-JADNAG @-JADNAG (@+JADLAG / @+JADLAG_MP)

*Israeldarrak Egypton gehiegitu egin zirelako*

*Liburu hark pentsarazi egin zion*

5. @-JADNAG @-JADLAG @+JADNAG_MP

*Bere zerbitzariaren odola haren eskuetan mendekatu izan baitu*

6. @-JADNAG @-JADNAG_MP

*Bertutezko bideak gu makur erazteko*

7. (EZ / EZIN) @-JADNAG_MP

*Entzunak arinki ez kanporatzea / Elkarri ezin adieraziz*

8. EZIN @-JADNAG @+JADLAG_MP

*Senarraren griña txarrak ezin irozo dituela*

9. (ARI / EZ) @+JADNAG_MP

*Artzainarekin jardunean ari zela / Hots hori ez duela euskarak berea, eta bai erdaratik erantsia*

10. @-JADNAG EZ (@+JADNAG_MP / @+JADLAG_MP)

*Gizonak oraindik sua ezagutzen ez zuenean*

11. @-JADNAG (EZ / NAHI /0) (@+JADLAG_IZLG>/ @+JADNAG_IZLG>)

*Inorekin etsaitzen ez direnak*

*Apaizen mailara igo nahi duena*

*Lehengo bideari* ***segitzen diotenak***

12. @-JADNAG_MP (ARI/ 0) (@+JADNAG / @+JADNAG_MP)

*Nekeak eta goseak txit* ***ahulduta zeuden***

13. (EZ / 0) (@-JADNAG / @+JADNAG)

*Euskaldunak, behin berbatuak, bihotza* ***ez dugu*** *geurea*

*Lurraren umea lurrak hazi* ***daroa***

*Han* ***daude****, nork nori erasoko*

*Goizean lanari* ***oratu***

14. EZ @+JADLAG @-JADNAG

*Gainbegiratu bat egin diot baina* ***ez dut irakurri***

15. NAHI @+JADNAG @-JADNAG

*Zuk manatzen dituzun hitzak* ***nahi ditut obratu***

16. EZIN @-JADNAG @+JADLAG

*Juduak eta mairuak* ***ezin herriratu ziren*** *Bizkaian*

Ondorengo ereduaren bidez aditz-kate jarrai nahiz ezjarraiak deskribatzen ditugu:

17. EZ @+JADLAG_MP (*/0) @-JADNAG

*Mairuak* ***ez zirela*** *inoiz Euskal Herrian* ***nagusitu***

*Manu hau* ***ez dezazula ahantz***

Azkeneko hiru aditz-kate abidideak, aditz-kate ezjarraienak ditugu:

18. EZ * @-JADNAG

***Ez*** *niri horrelako aitzakiarik* ***paratu***

19. (EZ/EZIN/0) @+JADLAG * @-JADNAG

***Ezin daiteke*** *horrelakorik* ***pentsa*** *ere*

***Ez ziren*** *gaizki esateari lekurik ematera* ***benturatu***

20. BEHAR @+JADNAG * @-JADNAG

*Baina* ***behar zen*** *latinean hobeki* ***trebatu***

### V.1.6.2 Sintagmak.

Sintagmak ezagutu ahal izateko, aditz-kateekin egin dugun bezalaxe, hauen hasiera (%SIH, sintagma-hasiera) eta bukaera (%SIB, sintagma-bukaera), eta hitz bakarreko sintagmak (%SINT) etiketez baliatu gara. Etiketa hauek esleitzeko mapaketa-erregelez osaturiko gramatika dugu (B eranskinean jasotzen ditugu). Mapaketa-erregelen oinarriak honako hauek dira:

- sintagma-etiketak aditz-funtzio ez diren gainontzeko funtzio sintaktikodun elementuei esleituko zaizkie.

- arau nagusia etiketa horiek esleitzeko honako hau da: elementuak lotzen joatea harik eta funtzio sintaktiko nagusi bat aurkitu arte. Hau da, funtzio sintaktikoak bi multzo nagusitan bereizten ditugu, modifikadoreak (> / < ikurra dutenak etiketa sintaktikoan) eta nagusiak (@ etiketa dutenak), —ikus § IV.8.2—. Bereizketa honi esker, ezartzen ditugu sintagma-hasierak eta sintagma-bukaerak, eta baita ere elementu bakarreko sintagmak. Hau da, modifikadoreek ezin dute beraiek bakarrik sintagmarik osatu, modifikadore direnez beti beste elementuren bati eragiten diote. Modifikatua den elementu honi modifikadorearen burua ere esaten zaio, hain zuzen ere, > / < ikurrek modifikatzen duten burua zein norabidetan dagoen adierazten dute. Beraz, modifikadore den elementu oro ezin izango da sintagma-bukaera izan (%SIB). Funtzio sintaktiko nagusidunak aldiz, sintagma bukaera (%SIB) etiketa eramango dute, edota elementu bakarreko sintagmaren etiketa (%SINT). Har dezagun adibide bat:

| *Soro* | *handi* | *bat* | *maneatu behar duzu* | *garia* | *ereiteko;* | ... |
|---|---|---|---|---|---|---|
| @KM> | @<IA | @OBJ | | @OBJ | | |
| %SIH | | %SIB | | %SINT | | |

Arau nagusi honek badu salbuespenik, sintagma koordinatuak osatzerakoan funtzio sintaktiko nagusidun juntagailua azaltzen baitzaigu, baina, jakina, hau ezin da inola ere sintagma baten bukaera izan. Adibidez:

| *Igande* | *eta* | *jaiegunetan* | *lanean* | *aritu* | *naiz* |
|---|---|---|---|---|---|
| @KM> | @PJ | @ADLG | | | |
| %SIH | | %SIB | | | |

Bestalde, sintagmen artean zenbait postposizio-sintagma ezagutzeko gauza garela aipatu behar da. Adibidez: ***Inoren aurka*** *dabil beti.*

Oinarriak eta zenbait adibide erakutsi ondoren, mapaketa-erregela batzuk aurkeztuko ditugu:

• Sintagma-hasiera etiketa esleitzeko mapaketa-erregela:

```
MAP (%SIH) TARGET GUNE
IF (NOT 0 (&POS_HAS)) (0 HAS_MAI)) (1 @PJ);
```

Mapaketa-erregela honen bidez adierazten duguna da: %SIH sintagma-hasiera etiketa jarriko zaiola gune etiketadunei (hots, GUNE= @KM>), baldin eta esaldiaren hasieran azaltzen diren, &POS_HAS (postposizio-hasiera) markarik gabe, eta urrats

batera eskuinetara perpaus-juntadura etiketadun elementu bat dagoen. Adibidez: ***Oinaze** eta arazoak askotuko dizkizut*

- Sintagma-bukaera (%SIB) etiketa asignatzeko mapaketa-erregela:

```
MAP (%SIB) TARGET NAG
IF (NOT 0 (&POS_HAS)) (NOT 1 MOD_EZK OR (@PJ))
(-1 MOD OR GUNE OR MOD_ESK);
```

Funtzio sintaktiko nagusia duen elementu bati %SIB etiketa jarriko zaio baldin eta &POS_HAS postposizio etiketarik ez duen, eskuinetara urrats batera ezkerreko modifikadorerik edo perpaus-juntadurako elementurik ez duen, eta ezkerretara urrats batera modifikadoreren bat edo gunea duen. Adibidez: *Baserri **zaharrak** ...*

- Elementu bakarreko sintagmaren etiketa, hau da, %SINT asignatzeko erregela:

```
MAP (%SINT) TARGET NAG
IF (NOT 0 ADPOSAG OR (&POS_HAS))
(NOT -1 (@PJ)) (NOT -1 MOD OR GUNE OR MOD_ESK);
```

Mapaketa honen bidez funtzio sintaktiko nagusia duen orori %SINT etiketa esleituko zaio baldin eta aditz-lokuzioetako osagaia ez den, edota ezta ere &POS_HAS etiketaduna, eta ezkerretara ez duen ez perpaus-juntadurako elementurik, ez modifikadorerik ez gunerik. Adibidez: ***Gizonak** egin du ...*

Ezagutzen ditugun izen-sintagmen egitura sinplifikatua honako elementu hauez osatuko litzateke (ikus § IV.8.2-n deskribatzen diren funtzio sintaktikoak):

- MOD = MODIFIKADOREAK= @IZLG> @<IZLG, @ID>, @<ID, @IA>, @<IA.
- GUNE =@KM>.
- FNAG= FUNTZIO SINTAKTIKO NAGUSIAK= @SUBJ, @OBJ, @ZOBJ, @ADLG, @PRED.
- +-ek esan nahi du behin edo gehiagotan.
- ( )-ek adierazten du aukeran dagoela elementu hori.
- derrigorrezko osagaiak, funtzio sintaktiko nagusidunak dira (FNAG-ekoak): @SUBJ, @OBJ, @ZOBJ, @ADLG, eta @PRED.

Ondorengo taulan, ezagutzen diren sintagmen egitura sinplifikatua adierazi dugu:

| |
|---|
| (MOD)+ (GUNE) (MOD)+ (GUNE) FNAG |
| (MOD)+ (GUNE) (MOD)+ (GUNE) FNAG SJ (MOD)+ (GUNE) (MOD)+ (GUNE) FNAG |

Egitura sinplifikatu horrek erakusten duen legez, derrigorrezkoa dena da beti funtzio sintaktiko nagusidun elementu bat izatea, gainontzeko elementuak hautazkoak ditugu.

Ezagutzen diren sintagmak honako hauek dira: izen-sintagmak, postposizio-sintagmak, sintagma koordinatu batzuk eta adizlagunak. Zenbait adibide emango ditugu ondoren:

*Bi aldiz*

*ale gorriz*

*bere emaztearekin*

*beste hertsigunerik*

*Emakume zahar bat*

*Aldizkari baten zenbaki bat*

*silize kristaldu mota bat*

*bere ahaleginaren ondorio gaiztoak*

*haren lehengo aurpegi argi eta ederra*

*ogia eta ardoa*

*borroka harrigarri baten ondoren*

### V.1.6.3 Aditz-kateen eta sintagmen osaketaren ebaluazioa.

Kateak markatzeko urratsaren ondorik, 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.

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. Adibidez:

*Erro honen gainera eraiki izan dira **geroztikako** aurrerapenak*

Adibide horretan *geroztikako* hitz-formak bi funtzio sintaktiko ditu, bata nagusia @ADLG eta bestea eznagusia edo modifikadorea @IZLG>. Anbiguotasun hau dela eta, bi modutara etiketa daiteke hitz-forma hori kate-osaketaren ikuspegitik, funtzio sintaktiko nagusiari %SINT etiketa dagokio, eta eznagusiari %SIH.

Aldiz, hitz berean bi funtzio sintaktiko nagusi edota bi funtzio sintaktiko modifikadore agertzeak ez dio kalterik egiten zatiak osatzeko urratsari. Jakina, baldin eta funtzio horiek zuzenak diren zatien ikuspegitik. Hau da, hitz bat beregaina bada sintagma osatzeko funtzio sintaktiko nagusia izango du, aldiz, beste elementuren batekin joan behar baldin badu sintagma bat osatzeko, orduan modifikadorearen funtzioa izan behar du. Adibidez:

***Orok** behar genuke horretara erori, **orok** bat eginez*

*Orok* hitz-formak bi interpretazio-lerro ditu absolutibo gisa @OBJ eta @SUBJ funtzioekin, eta ergatibo gisa @SUBJ funtzioarekin. Funtzio sintaktiko nagusiak ditu kasu guztietan, eta sintaktikoki subjektu ala objektu den desanbiguatzeke dagoen arren, izen-sintagma osatzeko anbiguotasun horrek ez du eraginik.

- Desanbiguatze-erroreak. Atal honetan, funtzio sintaktiko desegokiak hautatzerakoan suertatzen direnak hartuko ditugu kontuan. Hauek baitira zatiak osatzeko urratsari eragiten diotenak. Adibidez:

*Bakailao puska egosiak **oihal** batean xuka itzazu*

*Oihal* hitz-formak @OBJ funtzioa duenez ezin dugu [**oihal batean**] sintagma osatu. Hitz-forma horri @KM> (kasu markadunaren modifikadorea) funtzio sintaktikoa zegokion.

- 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. Adibidez:

*Emazteei beren urreriak galdeginik moldatu zioten **Araoni** aratxe bat, ...*

*Araoni* hitz-formaren lema gisa *Araoni* interpretazioa duena geratu zaigu desanbiguatze-urratsaren ondorik, eta horrexegatik @KM> funtzioa du esleitua @ZOBJ funtzioa izan beharrean.

Behar ez duen funtzioa esleitua duenez behar ez duen sintagma osatzen dugu: [**Araoni aratxe bat**]

- 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.

- Bestelako erroreak. Multzo honetan ditugu beste batzuen artean, aurreko urratsetatik garraiatutako akatsak, esate baterako, kasu batean adibide gisa aditzaren kategoria azaltzen zaigu. Hau da, kategoria gisa ezagutu beharrean adibide gisa hartu dugu hiztegiaren prestatze-lanetako lehenbiziko urratsetan. Eta horrez gain, badira hiztegiko adibideetan azaltzen diren zenbait forma errore-iturri bilakatzen direnak, adibidez: *zaite*, *nitekeen*, *itzatzu*, *zitzizkien*, *itxadon* ...

Bestalde, lotuta, hots, katea osatuz agertu beharko liratekeen elementuei dagokienez, 26 adibide baztertu beharrean gaude, hauetan lotuta agertu beharreko elementu bat edo gehiago solte azaltzen baitira. Lotzeke geratzen diren elementuak batez ere postposizio-egiturak, sintagma koordinatuak eta aditz-kate ezjarraiak ditugu. Hauek atzemateko baditugu zenbait erregela, baina azaldu zaizkigun kasu hauek harrapatzeko hobetu egin beharko lirateke. Dena dela, lehenago aipatu dugun bezala gure xedea batez ere kate jarraiak ezagutzean datza.

### V.1.6.4 Sintagmei esleituriko funtzio sintaktikoen ebaluazioa.

Ezagutu diren sintagmetan funtzio sintaktikoak ongi esleituak dauden neurtzeko, lagin osoaren ezaugarriak betetzen dituen lagin bat hautatu dugu ausaz, eta, horren gainean sintagma bakoitzari dagozkion funtzio sintaktikoak eskuz esleitu ditugu. Lagin hori eskuz analizatu ondoren, automatikoki markatutakoarekin konparatu dugu. Aipatu ausazko laginak 1211 adibide ditu; eta, horietarik konparaketarako aditz bakarrekoak direnak hartu ditugu kontuan: 646. Eta, irizpide gisa honako hauek erabili ditugu:

- Funzio hauei erreparatu diegu: subjektua, objektua, zehar objektua eta adizlaguna.

- Kontuan hartzen dena da eskuzko analisiaren bidez eta analisi automatikoaren bidez esleitutako funtzio sintaktikoak zenbat kasutan bat datozen edo ez. Bat ez etortze edo huts egitearen arrazoiak oker markatzea edo markatu gabe uztea izan daitezke.

Hona hemen ebaluazio honen emaitzak jasotzen dituen taula:

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 177 | 126 | 51 |
| OBJEKTU GISA MARKATUAK | 358 | 251 | 107 |
| ZEHAR OBJEKTU GISA MARKATUAK | 21 | 20 | 1 |
| ADIZLAGUN GISA MARKATUAK | 220 | 213 | 7 |

**V.1.6.4 Taula**: sintagmen funtzio sintaktikoen ebaluazioaren emaitzak.

Taulan ikus daitekeenez, zehar objektuen eta adizlagunen esleipena arrakastatsua da. Aldiz, alderik ahulena subjektu eta objektuen esleipenarena dugu. Dena den, hauetarako erdietsitako emaitzak ere ontzat jotzen ditugu, kontuan harturik %70 ongi etiketatzen dela, eta desanbiguatze sintaktikoaren alorra oraindik ere garapen egoeran dagoela.

Subjektuen eta objektuen kasuan egiten diren akatsen arrazoietariko bat, funtzio sintaktikoak esleitzeko aditz ezjokatuen kasuan dugun zailtasuna da. Hau da, nahiz eta aditz bakarrekoak izan, horietan ez dugu jokatuetan laguntzaileak eskaintzen duen laguntza, komunztadura dela medio zatien kasuak eta funtzioak erabakitzen baititugu. Edota, hiztegitik jaso dugun laguntzaile-mota ez da lagungarria suertatzen (ikus V.1.3 atala), esaterako, DA-DU. Begiratu batean oso esaldi sinpleak izan arren, ditugun baliabideekin ez dago asmatzerik aditz ezjokatuak dituzten adibide horietan zati bakoitzaren funtzio zuzena zein den. Horretarako, azpikategorizazioaren informazioa beharko genuke lexikoian. Adibidez: *Lana banatu.*

*Lana* objektu dela esateko, *banatu* aditzak zer-nolako objektuak hartzen dituen zehaztu beharko litzateke lexikoian. Zehaztapen horretan objektuaren paper tematikoa zehaztu beharko litzateke. Hartara, ahalko genuke objektua eta subjektua bereiztu, hau da, *lana* subjektu gisa ez etiketatzeko automatikoki, aditz horren agentea biziduna dela esan beharko litzateke lexikoian. Objektua edo gaia, aldiz, bizigabea izango litzateke. Honetarako, oso garrantzitsua litzateke

aditzen paper tematikoak zehaztuak izatea. Edota subjektu edo objektu izateko gai den elementu horren zein tasunek egiten duten subjektu edo objektu izan ahal izatea.

Bestalde, taula horretan jasotzen diren emaitzez gain, esan beharra dago analisi automatikoaren bidez erdietsitako sintagmen kopurua bat ez datorrenez eskuz analizatu denarekin (ikus V.1.7.3, eta, gogoratu %79 ezagutzen dela ongi), funtzio jakin batekin markatu diren sintagmen kopurua handiagoa edo txikiagoa izan daitekeela automatikoki markatutako laginean. Automatikoki markatu dugun laginean 40 sintagma gehiago ditugu eskuz aztertutakoan baino. Beraz, horren ondorioz aditz bati dagozkion sintagmak, eta hauei dagozkien funtzio sintaktikoak eta kasuak jasotzerakoan, azaleko patroi desegokiak eratuko ditugu. Esaterako, *Meza azkendu zen arte* adibidean, bi subjektu azaltzen zaizkigu: *Meza* eta *arte*, eta, horren ondorioz, bi subjektudun aditz gisa sailkatuko litzateke adibide hori.

## V.1.7 Analisiaren emaitzatik jaso den informazioa eta nola errepresentatu den.

Kateen osaketarako urratsean ezagutu ditugun kateak jasoko ditugu aditzen azterketarako baliagarri irizten diegun ezaugarriekin. Lehenik, gogora dezagun metodologiaren helburu nagusia, aztertzen ari garen adibideetako aditz-sarreren inguruan dauden sintagmak eta aditz-kateak jasotzea dela. Beraz, kate 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 SGMLratzeko urratsa eta ondorengo galdeketa-sistemarena. Azken urrats hauetatik at geratzen diren adibideak (1.473) eskuz aztertu beharko dira.

Zilegizko etiketak dituzten zatien adibideak, hau da, zatiek hasierako eta bukaerako etiketak adierazita dituztenean, hots, ez dagoenean zatia markatzeko etiketarik ixteke; orduan automatikoki aztertzeko prest daude. Baina, lehenbizi adibide hauetatik zer-nolako informazioa jasoko den zehaztu beharra dago. Aditzen adibideetatik baztertu beharrekoak eskuz analizatzeko utzi ondoren, prozesu automatiko baten bidez zer-nolako informazioa jaso den horietatik zehaztu, eta zehaztapen hori egiteaz gain, emaitza horiek nola errepresentatu ditugun deskribatuko dugu ondorengo lerroetan.

Lehenik, jasoko den informazioa zehazteari ekin diogu. Kontua da adibideetan oso informazio aberatsa dugula, eta dena jaso beharrean, batez ere aditzen azterketarako baliagarrienak deritzogunak jaso ditugula. Esate baterako, erlazioa (KONP (konpletiboa), HELB (helburuzkoa), DENB (denborazkoa), etab.), hau da, erlazio-atzizkien bidez sor daitezkeen mendekoei dagozkien erlazio-motak ez ditugu jasoko. Informazio hau jasotzerakoan, erlazio morfema batek dituen aukera guztiak jaso beharko genituzke (adib. -la (konpletiboa, moduzkoa, denborazkoa)), ez baitugu horien artean aukeratzeko modurik harik

eta aditzak zer eskatzen duen zehaztu arte. Hain zuzen ere, aditzaren inguruan dauden elementuak aditzak azpikategorizatuak diren ala ez aztertzeko laguntza eman nahi dugu, eta horretarako uste dugu aditzei dagokienean nahikoa dela inguruan dagoen aditz-katearen funtzio sintaktikoa jasotzea. Hau da, aditz-kate hori mendeko jokatua ala ezjokatua den jasoko dugu, eta horrez gain ondoren zehaztuko ditugun zenbait tasun.

Kontua da aztertzen ari garen aditz-sarrera bakoitzerako aditz-kateak (sarrerako aditz-kateak zein bestelako aditz-kateak) eta hauekin batera azaltzen diren sintagmak jasotzea. Aditz-kateak zein sintagma kateak jasotzeko sintagma-osaketarako urratsean esleituriko etiketez baliatuko gara, hots, aditz-kateak jasotzeko (%ADIKAT, %ADIKATHAS, %ADIKATBU, %ADIKATETEN, %ADIKATETENBU), eta sintagma kateak jasotzeko (%SINT, %SIH, %SIB) etiketez.

Kontsiderazio orokor gisa azpimarratu nahi genuke kate ororen posizioa jaso dela, hau da, zati orok perpausaren ordena linealean duen posizioa zehaztu da zenbaki baten bidez.

Kateak jasotzeaz gain, kate hauetatik zenbait informazio-mota jasotzeari interesgarri deritzogu (ondorengo puntuetan aipatzen diren ezaugarrien zerrenda osoa B eranskinean jaso dugu):

- Aditz-kateekin batera jaso diren ezaugarriak:
  - azpikategoria: hau da laguntzaile-mota, DA, DU, DIO, ZAIO, DA-DU.
  - funtzio sintaktikoak: aditzen funtzio sintaktikoak, @+JADNAG, @-JADNAG, @-JADNAG_MP_OBJ, @+JADLAG_IZLG>, ...

    Katea osatzen duten elementuen funtzio sintaktikoak jasotzen dira. Horrez gain, aditz-laguntzailea aditz-kate baten parte denean, bere erroa zein den adieraziko da.

    Aditzen funtzio sintaktikoak ikusi nahi dituenak, jo beza IV.6.2.3 puntura, atal horretan deskribatzen baitira.
  - pertsona-ezaugarriak: NR_NI (nor ni denean), NI_NI (nori ni denean), NK_NI (nork ni denean), NR_HI (nor hi denean), NI_HI (nori hi denean), ...
  - modua-denbora: A1 (indikatibozko orainaldia), B1(indikatibozko lehenaldia), ...
  - aspektua: burutua (BURU), ezburutua (EZBURU) eta etorkizuna (GERO).
  - modalitatea: egiatasuna (bai, ba, ez) eta ziurtasuna (ote, omen, ei, ...).
- Sintagma kateekin batera jaso diren ezaugarriak:

- kategoria: izen (IZE), adjektiboa (ADJ), aditza (ADI), adberbioa (ADB), determinatzailea (DET), izenordaina (IOR), loturazko elementua (LOT), partikulak (PRT), esapideak (ESP), esklamazioak (ESK), bestelakoak (BST).
- azpikategoria: hemen bakarrik izenari dagozkionak jaso ditugu, hau da, arrunta (ARR), izen berezia (IZB) eta leku-izen berezia (LIB).
- kasua: absolutiboa (ABS), ergatiboa (ERG), datiboa (DAT), etab. Hau da, deklinabideko kasuak.
- mugatasuna: mugagabea (MG) eta mugatua (MUGM).
- numeroa: plurala (NUMP), singularra (NUMS) eta plural hurbila (PH).
- funtzio sintaktiko nagusiak: @SUBJ, @OBJ, @ZOBJ, @ADLG, @PRED, etab. Funtzio sintaktiko nagusiak jasoko ditugu bakarrik, sintagma-bukaerako elementutik jasotzen baititugu ezaugarri hauek guztiak.

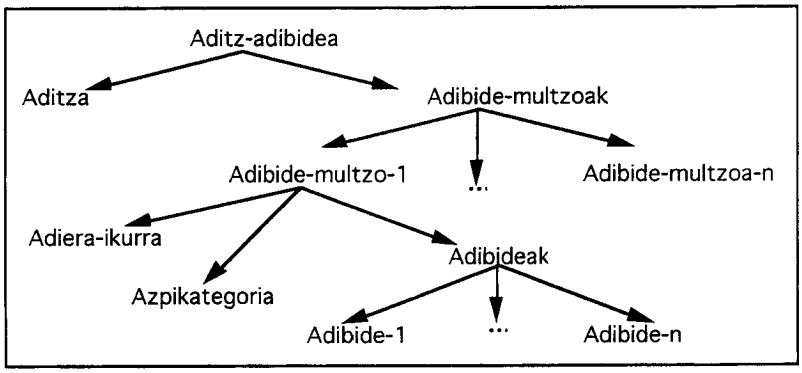
Ikus dezagun adibide baten bidez adibide bakoitzerako jaso den informazioa. Adibidez, *Neguko eguzkiak ez du berotzen*; *berotu*, *bero*, *berotzen* aditzaren adibidetik ondorengo laukian azaltzen den informazioa jasoko litzateke:

```
-Aditz_kate_adibidea
      -Aditza -> "berotu, bero, berotzen"
      -Adibide_multzoa
        -Adiera_ikurra -> "1"
        -Azpikategoria -> "DA-DU"
        -Adibideak
          -Adibide_katea -> "Neguko eguzkiak ez du berotzen"
          -Sarrerako_aditz_katea
            -Katea -> "ez du berotzen"
            -Adi_LAG
              -Katea -> "du"
              -Erroa -> "*edun"
            -Aditzen-funtzio-sintaktikoak
              -Aditzen-funtzio-sintaktikoa-1
                -Katea -> "du"
                -Funtzioa-> "@+JADLAG"
              -Aditzen-funtzio-sintaktikoa-2
                -Katea -> "berotzen"
                -Funtzioa-> "@-JADNAG"
            -Pertsona
              -Nor -> "NR_HU"
              -Nori
              -Nork -> "NK_HU"
            -Aspektua -> "EZBU"
            -Modalitatea -> "EGI"
            -Modua_denbora -> "A1"
          -Sintagmak
            -Katea -> "Neguko eguzkiak"
            -Kategoria -> "IZE"
            -Funtzioa -> "@SUBJ"
            -Kasua -> "ERG"
            -Numeroa -> "NUMS"
            -Mugatasuna -> "MUGM"
            -Azpikat_ize -> "ARR"
          -Bestelako_aditz_kateak
            -Katea ->
            -Adi_LAG
              -Katea ->
              -Erroa
            -Aditzen-funtzio-sintaktikoak
              -Aditzen-funtzio-sintaktikoa-1
                -Katea ->
                -Funtzioa->
              -Aditzen-funtzio-sintaktikoa-n
                -Katea ->
                -Funtzioa->
            -Pertsona
              -Nor ->
              -Nori
              -Nork ->
            -Aspektua ->
            -Modalitatea ->
            -Modua_denbora ->
```

**V.5 Irudia.-** Adibideetatik zer-nolako informazioa jasoko den adibide baten bidez eta modu eskematikoan adierazita.

Goiko adibidean ikus daitekeenez, jaso beharreko informazioa hautatzeaz gain, informazioa antolatu egiten dugu. Antolaketari dagokionean, azpimarratu behar da sarrerako aditz-kateari eta bestelako aditz-kateari dagokion informazioa bereizten dugula. Hau da, adibide batean aztertzen ari garen aditzarekin batera bestelako aditzik azaltzen baldin bada, bakoitzari dagokion informazioa bereiztuta geratzen da adibidearen errepresentaziorako aukeratu dugun errepresentazioan.

Adibideetarik jasoko ditugun ezaugarriak argitu ondorik, adibideak multzokatzeko irizpideak aurkeztuko ditugu. Horrela bada, eta errepresentazioari dagokionean, bide desberdinak aztertu ondoren (horien artean TEIko gomendioetan azaltzen dena errepresentazio sintaktikorako), gure adibideei lotutako programa baten bidez, adibideen analisiak, testu huts direnak, egitura aberatsago batera egokitzeko xedearekin zuhaitz-egitura batean gordeko ditugu. 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 B eranskinean adibideen analisien DTD definizioa). Bestalde, adibideak multzokatzeko irizpide nagusia adiera-ikurrarena izango dugu, eta adiera-ikurra esaterakoan adiera-ikurra bera, adiera-multzoa adierazten duen ikurra, adiera-ñabardura, etab. izendatzen ditugu. 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:




**V.6 Irudia.-** Adibideen antolamendua erakusten duen zuhaitza.

Adibideen zuhaitzaren gainean egingo diren galderen aurkezpena egingo dugu hurrengo puntuan. Hain zuzen ere, galdeketa-sistema honi esker lortu ditugu VI. kapituluan aurkeztuko ditugun emaitzak.

## V.1.8 Adibideen ezaugarri-egituren gainean lan egiteko galdeketa-sistema.

Aditzen adibideek EUSMGko analisiari esker eta zatien gramatikei esker aditz-katea eta sintagmak bereiziak dituzte. Horrez gain, aurreko puntuan zehaztu dugu zer-nolako informazioa jasoko den zati horietatik. Informazio hori guztia ustiatzeko ezaugarri-egituren errepresentazioa baliatu da, eta honen gainean garatuko dira oinarrizko argumentu-egitura jasotzearen ikuspegitik interesgarri izan daitezkeen zenbait galdera.

Galdera horiek azaldu aurretik adibide bat emango dugu aditzen adibideen ezaugarri-egitura (SGMLz adierazia) honek jasotzen duen informazioa irudikatzeko.

```
<Aditz-adibidea>
  <Aditza> jaso, jasotzen. </Aditza>
  <Adibide-multzoak>
    <Adibide-multzoa>
     <Adiera-ikurra>A1.</Adiera-ikurra>
     <Azpikategoria>DU</Azpikategoria>
     <Adibideak>
      <Adibidea>
       <Adibide-katea>Harria zortzi aldiz jaso du minutu batean.
       </Adibide-katea>
       <Sarrerako-aditz-katea>
          <Katea>jaso du</Katea>
          <Posizioa>3</Posizioa>
          <Aditz-laguntzailea>
             <Erroa>*edun</Erroa>
             <Katea>du</Katea>
          </Aditz-laguntzailea>
       <Aditzen-funtzio-sintaktikoak>
          <Aditzen-funtzio-sintaktikoa>
             <Funtzioa>@+JADLAG </Funtzioa>
             <Katea> du </Katea>
          </Aditzen-funtzio-sintaktikoa>
          <Aditzen-funtzio-sintaktikoa>
             <Funtzioa>@-JADNAG </Funtzioa>
             <Katea> jaso </Katea>
          </Aditzen-funtzio-sintaktikoa>
       </Aditzen-funtzio-sintaktikoak>
          <Pertsona>
             <Nor>NR-HU</Nor>
             <Nork>NK-HU</Nork>
          </Pertsona>
          <Modu-denbora>A1</Modu-denbora>
       </Sarrerako-aditz-katea>
       <Sintagmak>
          <Sintagma>
               <Katea>Harria</Katea>
               <Posizioa>1</Posizioa>
               <Kategoria>IZE</Kategoria>
               <Funtzio-sintaktikoa>@OBJ</Funtzio-sintaktikoa>
               <Kasua>ABS</Kasua>
               <Numeroa>S</Numeroa>
               <Mugatasuna>M</Mugatasuna>
          </Sintagma>
          <Sintagma>
               <Katea>zortzi aldiz</Katea>
               <Posizioa>2</Posizioa>
```

```
                <Kategoria>IZE</Kategoria>
                <Funtzio-sintaktikoa>@ADLG</Funtzio-sintaktikoa>
                <Kasua>INS</Kasua>
                <Mugatasuna>MG</Mugatasuna>
            </Sintagma>
            <Sintagma>
                <Katea>minutu batean</Katea>
                <Posizioa>4</Posizioa>
                <Kategoria>DET</Kategoria>
                <Funtzio-sintaktikoa>@ADLG</Funtzio-sintaktikoa>
                <Kasua>INE</Kasua>
                <Numeroa>S</Numeroa>
                <Mugatasuna>MUGM</Mugatasuna>
            </Sintagma>
        </Sintagmak>
      </Adibidea>
    </Adibideak>
  </Adibide-multzoa>
</Adibide-multzoak>
</Aditz-adibidea>
```

**V.1.8 Adibidea.** *jaso* aditzaren *Harria zortzi aldiz jaso du minutu batean* adibidearen analisiaren emaitzatik zer jasoko den SGMLz adierazita.

Aitzitik, argitu beharra dago aurreko puntuan adibideen informazio aberats horretatik zenbait informazio-mota jasotzea aukeratu bada ere, edozein dela ere galdera aztergai dugun aditz-katea (hots, hiztegi-sarrerakoa) eta adibide osoa jasoko direla. Horrela bada, beti jakingo dugu zein aditz den aztertzen ari garena eta dagokion adibide osoa izango genuke gordea. Horrez gain, beti jasoko da zati ororen posizioari dagokion informazioa ere.

Galdeketa-sistemaren bidez, azpikategorizazioaren gaia lantzeko bidea irekitzen dugu. Kontuan izanik informazio aberatsa jaso dugula, ikuspegi desberdinetatik heldu ahal zaio gai honi. Gure kasuan, IV.2 atalean aurkeztu genuen oinarrizko apikategorizazioaren ildotik planteatuko ditugu galderak.

Galdera hauek lau sail nagusitan bana ditzakegu:

1) Laguntzaile-motaren araberako multzoak osatzeko galdera.
   Galdera honen bidez bost multzotan banatuko dira aditzen adibideak: DA, DU, DA-DU, DIO eta ZAIO. Jakina, galdera honen arabera aditz-sarrera bera multzo batean baino gehiagotan ager daiteke.
   Adibidez: *alhatu, alha edo alhatu, alhatzen* aditzaren adibideak, DA, ZAIO eta DU multzoetan agertzen dira.

2) Ezaguturiko kateen artean bereizketa egiten duten galderak.
   Galdera hauen bidez aztergai dugun aditz bakoitza zenbat zatirekin azaltzen den jakin nahi dugu. Zati hauen artean bereizketa bat egiten da:

   2.1) Horietako zatiren bat perpausa den. Horretarako bestelako aditz-katerik dagoen begiratzen dugu, ea mendeko perpausen funtziorik —ikus IV.8.2.3.2 puntua—

agertzen den aztertuz. Azaltzen zaigun perpausa konplementua dela erabakiz gero, interesgarria litzateke konplementu horrentzako, sintagma konplementutik bereizten duen kodeketaren bat definitzea. Esate baterako, LDOCEn kodeketa sinple baten bidez bereizten zen bi konplementu horien artean. Aditz iragankorrentzat T1 kodea erabiltzen dute izen-sintagma objektua adierazteko eta T3 kodea infinitibozko perpausa objektu gisa dutenerako.

2.2) Aztergai dugun aditza forma jokatuan edo ezjokatuan dagoen.

Zati hauek balizko argumentutzat hartuko dira, ezin baitugu automatikoki erabaki benetako argumentuak diren ala ez. Erabaki hori lexikografo edo linguistak hartu behar izango du.

3) Zatien ezaugarriak konbinatzen direneko galderak.

Nagusiki funtzioa eta kasua konbinatzen dira. Esate baterako, galde daiteke subjektu ergatiboa eta objektu absolutiboa duten adibideak emateko.

4) Posizioa kontuan hartzen duten galderak.

Posizioaren[5] bidez, atzemandako kateak azaleko mailan non agertzen diren zehazten dugu. Zehaztapen hori zenbakitzearen bidez burutzen da, agerpen linealari erreparatuz, kate orori zenbaki bat esleitzen zaio. Zenbaki horiek aztertzen ari garen aditz-katearekiko erreferentzia gisa erabiliko dira. Posizioa erabiliko dugu aztergai dugun aditzaz gain bestelako aditz-kateak agertzen zaizkigunean, aditz-kate bakoitzari zein argumentu dagokion zehazten laguntzeko. Horrela bada, lehendabizi, aztertzen ari garen aditzetik hurbilen dauden balizko argumentuei erreparatu diegu. Hau da, aztergai dugun aditzetik hurbilen dauden elementuek aukera gehiago izango dute bere argumentu-egituraren parte izateko. Urrutiago daudenak, edota bestelako aditz-kate bat baino harantzagoko posizioetan dauden osagaiak aldiz, aukera urriagoak izango dituzte argumentu-egituraren parte gisa hartu ahal izateko.

Dena den, arazo hau ez da erraz konpontzen, eta lexikoian aurretik azpikategorizazioaren informazioa beharko genuke, jakiteko ea aditz batek mendeko perpausik hartzen duen ala ez. Jakina, aditz-sarrerarekin bat ez datorren aditza bestelako aditz-kate gisa ezagutzen dugu, eta, beraz, posizioaren nozioa baliatzerakoan ez dugu modurik bestelako aditz-kate hori aztertzen ari garen aditzak eskatzen duen ala ez jakiteko. Posizioaren nozioak balioko digu aztergai dugun aditzak, kasuan kasuko adibidean, mendeko perpausik ez duenean hartzen. Orduan, aztergai dugun aditzetik hurbilen dauden sintagmak, aditz horri dagozkionak lirateke, eta bestelako aditz-katetik harantzago daudenak ez genituzke

5 Gure kasuan, posizioak agerpen linealari egiten dio erreferentzia. Hala ere, jakina, badira posizioaren nozio konplexuagoak baliatzen dituztenak ere. Esate baterako, GENELEX proiektuaren baitan hiru elementuren arabera definitzen dute formalki: distribuzioa, funtzioa eta rol tematikoak.

kontuan hartuko. Hain zuzen ere, bestelako aditz-kate hori aztertzen ari garen aditzak eskatzen duen ala ez, ez dugu jakiterik. Horrexegatik adibidea mozterakoan, bestelako aditz-katea aztertu beharreko zatian gordetzen dugu.

Ikusteko posizioa aplikatzerakoan jatorrizko adibidea nola mozten dugun, har dezagun *segitu* aditz-sarrerako adibide bat *du* azpikategoriaren ondoren datorrena:

a) jatorrizko moduan:

*Eta gero haren hatzari jarraikirik hura **segi** dezagun.*

b) posizioa aplikatu ondoren, hots, adibidea moztu ondoren.

*jarraikirik hura **segi** dezagun.*

Moztu ezik adibide honetatik datibozko konplementua hartzen duela *segi* aditzak azalduko zitzaigun analisiaren emaitza gisa. Mozketa dela medio, konplementu gisa absolutiboa hartzen duen aditz gisa ageri zaigu.

Bukatzeko, esan beharra dago galdeketa-sistemak aditzak sailkatzeko eskaintzen dituen aukera horietarik, kasua eta funtzio sintaktikoa aintzat hartzen dituzten galderak baliatu ditugula. Horrela bada, ondorengo kapituluan aurkezten dugun aditzen sailkapena kasuan eta funtzio sintaktikoan oinarritzen da.

## V.2 TACAT-en eta gure azaleko sintaxiaren arteko konparazioa.

Atal honetan, gure asmoa ez da bi ikuspegiak, oso azaletik bada ere, konparatzea baino. Horretarako, TACAT erabiliz landu dugun gramatikatxoa aplikatuz sortzen den irteera ikusi ondorik, esaldi bera gure azaleko sintaxirako landu ditugun gramatikak aplikatu ondoren aurkeztuko dugu. Hona hemen adibide bat:

Gramatika:

```
P ==> P2,pj(eta), P11.
P ==> IS , AS.
P ==> IS,IS, AM.
P1 ==> IS, IS, ISIZLG.
P11 ==> IS, ISIZLG.
P2 ==> P1, AM, IS.
IS ==> adlg.
IS ==> atrib.
IS ==> km, izlg, subj.
IS ==> IZLGM , subj.
IS ==> subj.
IZLGM ==> izlg.
IS ==> izlg, km, adlg.
IS ==> id, obj.
ISIZLG ==> AMIZLG, IS.
AM ==> jadnag.
AM ==> jadnag, jadlag.
```

```
AMIZLG ==> jadnag, jadlagizlg.
AMIZLG ==> jadnagizlg.
```

Tacaten irteera:

```
{{Hurrengo_izlg  postal_km  batean_adlg}_IS  {beste_id  hauxe_obj}_IS
 {idatzi_jadnag zenuen_jadlag }_AM }_P
```

Kate-osaketarako gramatikak aplikatu ondoren lorturiko emaitza:

```
  "<Hurrengo>" S:1258, 4163
     <Correct!> "hurrengo" ADJ IZL HAS_MAI  S:955 @IZLG> %SIH
  "<postal>" S:2193
     <Correct!> "postal" IZE ARR ZERO  S:1007 @KM>
  "<batean>" S:2533
     <Correct!> "bat" DET DZH NUMS DEK NUMS MUGM DEK INE @ADLG %SIB
  "<,>"
     PUNT_KOMA
  "<beste>" S:3677, 3675
     <Correct!> "beste" DET DZG ZERO  S:919 @ID> %SIH
  "<hauxe>" S:5501
     <Correct!> "hau+xe" DET ERKARR ABS NUMS MUGM GRA IND @OBJ %SIB
  "<idatzi>" S:1866
     <Correct!> "idatzi" ADI SIN AMM PART ASP BURU NOTDEK  S:812 @-JADNAG
S:104  %ADIKATHAS
  "<zenuen>" S:1643, 4565
     <Correct!> "*edun" ADL B1 NOR_NORK NR_HU NK_ZU  S:796 @+JADLAG S:55
%ADIKATBU
  "<$.>"
     PUNT_PUNT
```

Aditzen adibideak aztertzeko garatu dugun metodologiako alderik irekienetarikoa dugu aditz-kateak eta sintagmak ezagutzeko landu dugun partea. Batetik, atal honetan aipatu ditugun beste tresna horiek atzematen dituzten kateen parekoak ezagutzen ditugula kontuan harturik; lan horietan aipatzen diren aplikazioetarako gure lanak zer-nolako aukerak edota emaitzak izango lituzkeen etorkizuneko ikerbidea baita.

Bestetik, atal hau hobetzeko posibilitateak sakon aztertzeari eta aplikatzeari ere interesgarri deritzogu. Horrez gain, kontuan izan behar dugu, IXA taldean bertan, badugula azaleko sintaxiaren arloan, eta aplikazioen artean azpikategorizazioa helburu duen beste lan bat. Lan honetan (Aldezabal eta beste, 1999), analizatzaile sintaktiko partziala garatzen da analisirako teknika desberdinak konbinatuz, eta morfemak hartzen ditu oinarri gisa. Gure hurbilpenean, aldiz, CGn oinarritu garenez, hitza hartu dugu oinarri gisa.

Teknika desberdinak konbinatzetik datorkio ikuspegi horri bere ahalmena, sintagmak ezagutzeaz gain zenbait mendeko perpausa ere ezagutzeko gauza baita. Desanbiguatze morfosintaktikorako EUSMG baliatzen du, hori aplikatu ezik analisi-kopurua trataezina izango bailitzateke.

Bestalde, baterakuntzan oinarritutako erregelak (PATR-II) lantzen dituzte osagai sintaktiko partzialen egituraketarako. Horiez gain, egoera finituko parserren teknologia (Xerox Finite

State Tool, XFST) ere erabiltzen dute zenbait eginkizunetarako, esate baterako: EUSMG-k utzitako anbiguotasuna tratatzeko (CGrekiko abantaila gisa, hitza baino unitate sintaktiko handiagoei egin dakieke erreferentzia), analisiko zenbait tasun iragazteko aplikazioaren arabera eta esaldien zatiak eskuratzeko.

Kontuan izanik, euskara beste zenbait hizkuntzaren aldean pobrea dela analisi sintaktiko automatikorako baliabideetan, taldean bi ikuspegi lantzea aberasgarri dela uste dugu. Eta bide horiek jorratzean erdietsitako emaitzek eta esperientziek etorkizunean egin behar diren urratsak sendotzen lagunduko digutelakoan gaude.

# VI. EHko aditzen adibideen azterketatik ateratako emaitzak

Aditzen adibideen azterketarako garatu dugun metodologiaren bidez lor daitezkeen emaitzetariko batzuk aurkeztuko ditugu kapitulu honetan. Aurreko kapituluaren bukaeran esan bezala, kasua eta funtzio sintaktikoak baliatu ditugu aditzak multzokatzeko. Emaitzak aurkeztu aurretik, lehenbizi azaleko emaitza horien inguruan jardungo dugu —VI.1 puntuan—, azpimarratuz batetik, emaitzak berak baino metodologia garatzea izan dela helburu nagusia, eta, bestetik, azaleko sintaxiak duen garrantzia emaitza horiek erdiesteko. Eta horrez gain, garbi utziz gauza bat dela aditz bat azaltzen deneko azaleko patroi sintaktikoak jasotzea; eta oso bestelakoa argumentu-egitura finkatzea. Ondoren, —VI.2 puntuan— aditzak multzokatzeko irizpideak aurkeztuko ditugu. Irizpideak azaldu ondorik, —VI.3 puntuan— emaitzak emango ditugu, hau da, atzemandako azaleko patroiak eta hauei dagozkien adibideak azalduko ditugu. Ondoren, —VI.4 puntuan— azaleko patroi horiek automatikoki erdiesterakoan izandako zailtasun nagusiak deskribatu, eta emaitzak ebaluatuko ditugu. Bukaeran, garatu dugun azaleko sintaxiaren, eta honen ondorioz egin dugun aditzen sailkapenaren inguruan zenbait ondorio nabarmenduko ditugu.

## VI.1 Emaitzen inguruan.

Aditzen adibideetatik azpikategorizazioari dagokion informazioa erdiesteko garatu dugun metodologiaren bidez zer-nolako emaitzak jaso ditugun aurkeztu baino lehen, emaitza horien inguruan mintzatuko gara. Lehendabizi, esan beharra dago batez ere aditz bakarreko adibideetan oinarrituko garela azterketa honen emaitzak aurkezterakoan. Horren arrazoi nagusia, aditz bat baino gehiagoko adibideak aztertzerakoan azaleko sintaxiak aurre egin ezin dion arazo batekin topo egitean datza, hots, zein aditzi dagokion konplementu bakoitza erabaki beharra. Horrez gain, aditz bakarrekoetatik erdiesten den informazioa eskuz egiaztatzeak lan gutxiago emango du. Aditz bat baino gehiagokoetatik erdietsitako emaitzek, aldiz, eskuzko lan handia eskatuko lukete automatikoki atera denetik patroi zuzena erabakitzeko. Beraz, aditz bakarreko adibide guztiak aztertu ditugu. Emaitza hauek azaleko patroi sintaktikoak dira, hau da, aditz bat zer-nolako egitura sintaktikoetan agertzen den hiztegiko adibideetan. Aztergai dugun aditzaren inguruan dauden sintagmak eta aditz-kateak ezagutzeko azaleko sintaxia garatu dugu —ikus § V.1.6—. Horrela, azaleko sintaxiari esker ezagutu dugun kate bakoitzetik azpikategorizazio-lanetarako baliagarria den informazioa jaso dugu —ikus § V.1.7—; zehazki, sintagmen kasua eta funtzioa jaso ditugu, beti ere azaleko sintaxiaren eremura mugatuz.

Bestalde, azterketa errazteko bi multzo nagusi bereizi genituen adibideetan:

- Aditz bakarrekoak. Lehenengo multzo horretan bi azpimultzo egin genituen era berean: jokatu eta ezjokatuen artean.
- Aditz bat baino gehiagokoak. Adibide horietatik zenbait aztertu ditugu posizioaren nozioaren baliagarritasuna aztertzeko —ikus § VI.4.2—.

Bukatzeko, kontuan izan behar da, beti ere jasoriko emaitzak linguista edota lexikografo batek egiaztatu beharko dituela eskuz, eta egiten den azpikategorizazio-proposamena azalekoa izango dela. Horrez gain, garbi dago gauza bat dela aditz hauek adibide hauetan zer-nolako sintagmak eta aditz-kateak dituzten jasotzea; eta oso bestelakoa dela argumentu-egitura finkatzearena. Hots, elementu horiek azaltzeak ez du esan nahi argumentu direnik. Argumentu-egitura finkatzeko azterketa sakonagoa egin beharko litzateke eta. Ildo horretatik, azterketa etorkizunean corpus handiago batera zabaltzea ezinbestekotzat jotzen dugu —ikus § VII.1.1—.

## VI.2 Aditzak multzokatzeko irizpideak.

Lehenago aipatu dugun bezala, azaleko analisiari esker adibide bakoitzaren analisia erdietsi dugu, eta analisi horretatik azpikategorizazioa lantzeko erabilgarriak izan daitezkeen zenbait tasun jaso ditugu. Informazio aberats hori guztia kontuan izanik era askotara multzoka daitezke adibideak. Baina, gure kasuan, kasuari eta funtzio sintaktikoari erreparatu diegu. Eta erdietsitako egitura sintaktiko horiek multzokatzerakoan, funtzio sintaktiko/kasu hauetan oinarritu gara: @SUBJ-ERG, @SUBJ-ABS, @OBJ-ABS, @ZOBJ-DAT. Hau da, funtzio/kasu bikote horien agerpenaren arabera egin ditugu multzoak. Beraz, hiztegiko adibideetan azaltzen dituzten elementuei erreparatuko diegu, hots, lexikalki gauzatzen direnei. Izan ere, jakin badakigu euskaraz aditzarekin komunztatzen duten elementuak azalean ez gauzatzea oso arrunta dela, eta hortaz, eliditzen diren elementuak ez ditugu kontuan hartu sailkapenean.

Esate baterako, aztertu ditugun aditz jokatuen kasuan, 2.700 adibideetatik 500 adibidetan ez da ez subjektu ergatiborik, ez subjektu absolutiborik, ez objektu absolutiborik ez zehar objektu datiborik agertzen. Horretaz gain, maiz gertatuko da gainontzeko multzoetan horietako funtzioren bat eliditzea, batez ere subjektu ergatiboa dugu gehien eliditzen den kasua. Datu hori nahiko esanguratsua dela esango genuke, eta eskatuko luke azalean agertzen diren bestelako kasuak, hots, gure azaleko patroietatik kanpo geratzen direnak ere aintzat hartzea azpikategorizazioa aztertzean. Beraz, uste dugu azaleko egitura sintaktiko hauetatik kanpo geratzen diren zenbait kasu/funtzio ere kontuan hartu behar liratekeela argumentu-egituraren parte diren ala ez erabakitzeko orduan. Esate baterako, leku-kasuak zenbait aditzetan aditz horien argumentu-egituraren parte izan daitezke. Hona hemen, ez subjektu ergatiborik, ez

objektu absolutiborik, ez zehar objekturik (ZERO-@SUBJ_ERG-@OBJ_ABS-@ZOBJ_DAT) ez duten multzo horretan azaldu zaizkigun zenbait aditz eta zer-nolako kasuak azaltzen dituzten:

atera: 8 adibidetan leku-kasuekin: ABL eta INE (32 adibideetatik)

igo: 4 aldiz ALA eta 1 INE (22 adibideetatik)

iritsi: 2 ALA, 2 INS, 1 INE, 1 ABL (17 adibideetatik)

itzuli: 5 ALA eta 1 ABL (32 adibideetatik)

hurbildu: 2 ALA eta 1 INE (14 adibideetatik)

dudatu: 3 INS (6 adibideetatik)

Aipaturiko aditz horietan (mugimenduzko aditzak, batik bat) batez ere leku-kasuak agertzen dira azalean. Beste aditz batzuetan kasu instrumentala ageri da, esate baterako, *aldatu*, *baliatu*, *begiratu*, *burlatu* aditzen adibideen kasuan. Kasu horiek guztiak aditzen argumentu-egituraren azterketatik kanpo geratu dira askotan, baina, hala ere, lehenago aipatu dugun bezala, kontuan hartu behar liratekeela uste dugu.

Zergatik ez ditugu horiek kontuan hartu? Ez direlako argumentu-egituraren parte izan daitezkeen kasu/funtzio ohikoenak. Horrelako kasu/funtzioek, oro har, gutxiagotan parte hartu izan dute aditz baten argumentu-egitura zehazterakoan, eta, beraz, horregatik ez ditugu multzoak eraikitzeko irizpide gisa hartu. Hala ere, galdeketa-sistemaren bidez azterketa zuzenduago bat egin dezakegu, esate baterako, leku-kasuak/funtzioak hartzen dituzten aditzen adibideak aztertuz. Guk adibide horien analisi osoak erdietsi ditugu, eta, beraz, badugu aditz batek inguruan dituen sintagmen kasu/funtzioen berri. Hau da, aditz bakoitzeko dugu, zer adibide dituen —adibide bakoitza dagokion adiera eta azpikategoriaren arabera sailkaturik— eta adibide bakoitzari dagokion azaleko analisia. Azaleko analisi horren aurretik aditzaren partizipioa, adibide hori zein adierari dagokion, azpikategoria eta adibide-zenbakia ditugu. Elementu horien bidez aditz-gakoa definitzen dugu. Ondoren, aditz-gako bakoitzari dagokion azaleko analisia dugu, azaleko analisi horretan lehenbizi hiztegian duen laguntzaile-mota azaltzen da, eta jarraian funtzio sintaktiko/kasu bikoteak[1] ; eta adibide horretan bestelako aditz-katerik azaltzen baldin bada, MP ikurraren bidez adierazten dugu (+ mendekoa / - ez-mendekoa izan den ala ez). 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-
```

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

```
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 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.*

Eranskinean jasotzen ditugu aditzen adibide guztiak adizka sailkaturik, horrela aditz bakoitzak zer-nolako azaleko egitura sintaktikoak aurkezten dituen ikus daiteke (ikus C eranskina). Hala ere, ondorengo puntuan —ikus § VI.3.— funtzioa eta kasua bakarrik kontuan hartuko ditugu aditzak multzokatzerakoan. Irizpide horren arabera bereizi dugun multzo bakoitzerako zenbait adibide emango ditugu (adibide guztiak C eranskinean jasoko dira).

Azkenik, esan beharra dago definituriko funtzio sintaktikoen multzoak ere baduela eraginik egitura posibleak jasotzerakoan —ikus § IV.8.2—. Jakina, jasotzen ditugun azaleko egitura horiek gure funtzio sintaktikoen multzoan definiturikoak dira. Analisi sintaktikoaren ikuspegitik funtzio sintaktiko egokiak ditugula esan genezake. Baina, adibideetan aplikatzerakoan funtzio horietan egindako zenbait bereizketa kaltegarri suertatzen dira. Hots, bereizketa horiek automatikoki ebazten oso zailak dira. Hori dela eta, zenbaitetan okerreko egitura sintaktikoak jasoko ditugu. Horietako funtzio sintaktiko batzuk hemen aipatzearren, @PRED (izenki-predikatua) funtzioa bereizteak okerreko egiturak jasotzera eramaten gaitu gehienetan. Berez, @PRED funtzioa bereiztea beharrezkoa ikusten dugu azpikategorizazioari begira, baina horrelako funtzioa ongi esleitzea zaila suertatzen da, hain zuzen ere azpikategorizazioaren informaziorik ez dugulako lexikoian. Beraz, hastapeneko analisietan funtzio sintaktiko hori bereizi gabe jardutea egokiagoa iruditzen zaigu.

Bestalde, mendeko aditzen funtzio sintaktikoetan, mendeko horrek nagusian zer funtzio betetzen duen zehazteak ere egitura okerrak atzematera eramaten gaitu. Funtzio sintaktikoen multzoan bereizketa horiek izatea linguistikoki motibatuta egon arren, garatu dugun metodologiaren helbururako ez da egokia. Hots, esate baterako, aditz batek bere inguruan mendeko ezjokatu bat duela, mendeko hori jasotzera irits gaitezke, baina, ezjokatu horrek nagusiarekiko zer funtzio betetzen duen ez gara kapaz erabakitzeko. Horretarako, azpikategorizazioaren informazioaren laguntza beharko genuke. Azken finean, funtzio sintaktiko xeheagoak erabiltzeak desanbiguatze-urratsa zailtzen du, eta, beraz, jasoriko informazioan akats gehiago egoteko arriskua izango dugu.

Beraz, etiketa sintaktikoen multzoari dagokionez, saiakera honetatik ondoriozta daiteke funtzio sintaktikoen multzoan mendekoek nagusiarekiko zer-nolako funtzioa betetzen duten zehaztea azpikategorizazioa zehazten denerako utzi beharko litzatekeela. Eta izenki-predikatuaren (@PRED) funtzioa ere azpikategorizazioa lantzean zehaztu behar litzateke. Orduan izango genuke aukera aditz bakoitzak zer-nolako mendekoak eskatzen dituen eta mendeko horien funtzioak zehazteko.

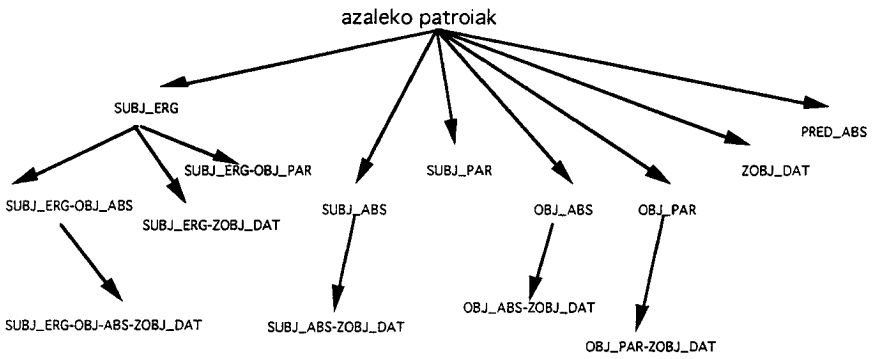
## VI.3 Jasoriko azaleko patroien multzoa.

Lehenago aipatu dugun bezala, funtzio sintaktikoak eta kasuak hartzen ditugu kontuan adibideak multzokatzerakoan. Modu honetara, aditz desberdinak, agertzen dituzten azaleko funtzio sintaktiko/kasu horien arabera multzo berean ikus ditzakegu. Jakina, azaleko funtzio sintaktiko/kasu horietan bat etorri arren, aditz horiek multzo xeheagoetan multzokatzeko irizpideak landu beharko lirateke. Multzokatze hau, hasierako urrats xumea baino ez da. Honen gainean, beste lan batzuen artean, has daitezke rol tematikoak zehazteko lanak.

Patroi hauek aditz bakoitzak onar ditzakeen egiturak dira besterik gabe. Esan bezala, aditz hauek derrigorrezko dituzten argumentuak finkatzeak azterketa sakonagoa eskatzen du. Beste irizpide batzuen artean, semantika ere kontuan hartu beharko litzateke autore batzuen ustez; esate baterako, Levin (1993)-n esaten denaren arabera, jokaera sintaktikoa aditzaren semantikak mugatzen du. Horrelako azterketei bidea emateko, aditzen adibideak antolatzerakoan adibide bakoitzean aditza zein adieratan erabilia den jasotzen dugu. Dena dela, azterketa hori etorkizunerako lan gisa utzi dugu beste batzuekin batera.

## VI.3.1 Azaleko patroiak eta adibideak.

Aurrerago esan bezala, atal honetan erdietsi ditugun azaleko patroiak aurkeztuko ditugu, horietako bakoitza zenbait adibiderekin[2] ilustratuz. Ikus dezagun eskema baten bidez zernolako patroiak jaso ditugun:




**VI.1 Irudia.-** Adibideen azaleko patroiak.

### @SUBJ_ERG

Multzo honetan aditz iragankor objekturik gabeak ditugu. Beste batzuen artean aditz hauek ditugu: kurritu (*kurritu-A0.-DU-4*), zirkulatu (*zirkulatu-A0.-DU-1*), distiratu (*distiratu-A0.-DU-2*).

Adibidez:

```
|   |   adibidea-1:
|   |   |   identifikadorea: distiratu-A0.-DU-1
|   |   |   adibide-katea: Mila eguzkik bezala distiratuz.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: distiratuz
|   |   |       posizioa: 3
|   |   |       f-adi-sint-1:  distiratuz  @-JADNAG-MP
|   |   |   sintagma-1:  Mila eguzkik  1  IZE   ERG  MG  ARR  @SUBJ
|   |   |   sintagma-2:  bezala  2  ADB  @ADLG
```

### @SUBJ_ERG-@OBJ_ABS

Multzo honetan aditz iragankor objektudunak ditugu. Multzo honetan gehienak DU eta DA-DU laguntzaile-motakoak dira, baina bada DIO eta DA laguntzaile-motako adibideren bat ere. Adibide ugariko multzoa dugu. Hona hemen zenbait adibide:

```
|   adibide-multzoa-2:
```

---

[2] Adibibideen analisiak SGMLz adierazi ditugun arren, tesi-txostenean modu honetara paratu ditugu toki gutxiago hartzeagatik. Esan beharra dago, aurkezpenerako modu hau SGML-z adierazita egoteari esker erdietsi dugula.

```
|   |   adiera-ikur: A1.N1.
|   |   azpikat: DA-DU
|   |   adibidea-1:
|   |   |   identifikadorea: hondatu-A1.N1.-DA-DU-1
|   |   |   adibide-katea: Zure seme horrek , berriz , emagalduekin hondatu
                          ditu ondasun guztiak.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: hondatu ditu
|   |   |       posizioa: 3
|   |   |       pertsona  NR-HK  NK-HU
|   |   |       adi-lag-1:  *edun  ditu  @+JADLAG
|   |   |       f-adi-sint-1:  hondatu  @-JADNAG
|   |   |       aspektua: BURU
|   |   |       modua-denbora: A1
|   |   |   sintagma-1:  Zure seme horrek  1  DET  ERG  NUMS  MUGM  @SUBJ
|   |   |   sintagma-2:  emagalduekin  2  IZE  SOZ  NUMP  MUGM  ARR  @ADLG
|   |   |   sintagma-3:  ondasun guztiak  4  DET  ABS  NUMP  MUGM  @OBJ
```

Patroi bereko aditz ezjokatuarekin aurkituriko adibide bat:

```
|   |   |   identifikadorea: harrapatu-A1.N1.-DU-1
|   |   |   adibide-katea: Animalia batek beste bat harrapatu.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: harrapatu
|   |   |       posizioa: 3
|   |   |       f-adi-sint-1:  harrapatu  @-JADNAG
|   |   |   sintagma-1:  Animalia batek  1  DET  ERG  NUMS  MG  @SUBJ
|   |   |   sintagma-2:  beste bat  2  DET  ABS  NUMS  MG  @OBJ
```

Objektu zuzena partitiboan egiten dutenak ere hemen sartu ditugu.

```
|   |   adibidea-7:
|   |   |   identifikadorea: ediren-A0.-DA-DU-7
|   |   |   adibide-katea: Gaiztoak ez du inon ere segurantzarik edireten.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: ez du * edireten
|   |   |       posizioa: 2
|   |   |       pertsona  NR-HU  NK-HU
|   |   |       adi-lag-1:  *edun  du  @+JADLAG
|   |   |       f-adi-sint-1:  edireten  @-JADNAG
|   |   |       modalitatea: EGI
|   |   |       modua-denbora: A1
```

```
|   |   |   sintagma-1:  Gaiztoak  1  ADJ  ERG  NUMS  MUGM  @SUBJ
|   |   |   sintagma-2:  inon  3  ADB  @ADLG
|   |   |   sintagma-3:  segurantzarik  4  IZE  PAR  MG  ARR  @OBJ
```

### @SUBJ_ABS-@ZOBJ_DAT

Subjektu absolutiboa eta datibozko objektua hartzen dituzten adibideak. Multzo honetan ZAIO laguntzaile-mota agertzen da gehien, baina DA-DU eta DA motakoak ere ageri zaizkigu nahiko maiz. Espero zitekeen bezala, gutxien agertzen direnak DU eta DIO motako adibideak dira. Adib:

```
|   |   adibidea-2:
|   |   |   identifikadorea: garaitu-A2.-ZAIO-2
|   |   |   adibide-katea: Infernuko ateak ez zaizkio hari garaituko.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: ez zaizkio * garaituko
|   |   |       posizioa: 2
|   |   |       pertsona  NR-HK  NI-HU
|   |   |       adi-lag-1:  izan  zaizkio  @+JADLAG
|   |   |       f-adi-sint-1:  garaituko  @-JADNAG
|   |   |       aspektua: GERO
|   |   |       modalitatea: EGI
|   |   |       modua-denbora: A1
|   |   |   sintagma-1:  Infernuko ateak  1  IZE  ABS  NUMP  MUGM  ARR
                          @SUBJ
|   |   |   sintagma-2:  hari  3  DET  DAT  NUMS  MUGM  @ZOBJ
```

Patroi bereko aditz ezjokatuarekin aurkituriko adibide bat:

```
|   |   adibidea-1:
|   |   |   identifikadorea: itsatsi-N1.-ZAIO-1
|   |   |   adibide-katea: Bideko hautsa oinetakoei itsatsi.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: itsatsi
|   |   |       posizioa: 3
|   |   |       f-adi-sint-1:  itsatsi  @-JADNAG
|   |   |   sintagma-1:  Bideko hautsa  1  IZE  ABS  NUMS  MUGM  ARR  @SUBJ
|   |   |   sintagma-2:  oinetakoei  2  IZE  DAT  NUMP  MUGM  ARR  @ZOBJ
```

### @SUBJ_ABS

Aditz iragangaitzak ditugu multzo honetan. DA laguntzaile-motakoak dira gehien azaltzen diren adibideak. Hala ere, gainontzeko motakoak ere agertzen zaizkigu. Horietatik, DA-DU motakoak dira ugarienak. Bestalde, DU motakoak direnetan, laguntzaile intrantsitiboa agertzen

denetan aditzak iragankor izaten jarraitzen duela esan genezake, baina egitura sintaktikoa inpertsonala izango da. Hori fenomeno ezaguna da, DU motakoak egitura inpertsonalaren bidez DA laguntzaile-motarekin azaltzea. Adibidez: *Ezkontza **urra** daiteke?*

Ikus dezagun orain multzo honetako aditz baten adibide osoa analizaturik:

```
|   aditza: adostu, ados edo adostu, adosten.
|   adibide-multzoa-1:
|   |   adiera-ikur: A1.
|   |   azpikat: da
|   |   adibidea-1:
|   |   |   identifikadorea: adostu-A1.-DA-1
|   |   |   adibide-katea: Bi taldeak 37ra adostu dira , eta handik aurrera
                           beste hiruzpalau aldiz , azkenik 44ra.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: adostu dira
|   |   |       posizioa: 3
|   |   |       pertsona: NR-HK
|   |   |       adi-lag-1:  izan  dira  @+JADLAG
|   |   |       f-adi-sint-1:  adostu  @-JADNAG
|   |   |       aspektua: BURU
|   |   |       modua-denbora: A1
|   |   |   sintagma-1:  Bi taldeak  1  IZE  ABS  NUMP  MUGM  ARR  @SUBJ
|   |   |   sintagma-2:  37ra  2  IZE  ALA  NUMS  MUGM  ARR  @ADLG
|   |   |   sintagma-3:  handik  4  ADB  ABL  @ADLG
|   |   |   sintagma-4:  aurrera  5  IZE  ALA  NUMS  MUGM  ARR  @ADLG
|   |   |   sintagma-5:  beste hiruzpalau aldiz  6  LOT  INS  MG  ARR
                           @ADLG
|   |   |   sintagma-6:  azkenik  7  ADB  @ADLG
|   |   |   sintagma-7:  44ra  8  IZE  ALA  NUMS  MUGM  ARR  @ADLG
```

Patroi bereko aditz ezjokatuarekin aurkituriko adibide bat:

```
|   |   adibidea-3:
|   |   |   identifikadorea: abiatu-A1.N1.-DA-3
|   |   |   adibide-katea: Txakurra erbiaren atzetik abiatu.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: abiatu
|   |   |       posizioa: 3
|   |   |       f-adi-sint-1:  @-JADNAG  abiatu
|   |   |   sintagma-1:  Txakurra  1  IZE  ABS  NUMS  MUGM  ARR  @SUBJ
```

```
|   |   |   sintagma-2:  erbiaren atzetik  2  IZE  ABL  NUMS  MUGM  ARR
                          @ADLG
```

### @SUBJ_ERG-@ZOBJ_DAT

Multzo honetan subjektu ergatiboa eta datibozko objektua dutenak biltzen ditugu. Adibidez:

```
|   |   adibidea-2:
|   |   |   identifikadorea: eragin-H1.A1.N1.-DIO-2
|   |   |   adibide-katea: Platonek ere sakonki eragin zion Fray Luisen
                           gogamenari.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: eragin zion
|   |   |       posizioa: 3
|   |   |       pertsona  NR-HU  NI-HU  NK-HU
|   |   |       adi-lag-1:  *edun  zion  @+JADLAG
|   |   |       f-adi-sint-1:  eragin  @-JADNAG
|   |   |       aspektua: BURU
|   |   |       modua-denbora: B1
|   |   |   sintagma-1:  Platonek  1  IZE  ERG  MG  IZB  @SUBJ
|   |   |   sintagma-2:  sakonki  2  ADB  @ADLG
|   |   |   sintagma-3:  Fray Luisen gogamenari  4  IZE  DAT  NUMS  MUGM
                         ARR  @ZOBJ
```

Patroi bereko aditz ezjokatuarekin aurkituriko adibide bat:

```
|   |   |   identifikadorea: burutu-A1.N2.-DU-2
|   |   |   adibide-katea: Nork burutu zuri jokoan?
|   |   |   sarrerako-aditz-katea
|   |   |       katea: burutu
|   |   |       posizioa: 2
|   |   |       f-adi-sint-1:  @-JADNAG  burutu
|   |   |   sintagma-1:  Nork  1  IOR  ERG  MG  @SUBJ
|   |   |   sintagma-2:  zuri  3  IOR  DAT  NUMS  MG  @ZOBJ
```

### @SUBJ_ERG-@OBJ_ABS-@ZOBJ_DAT

Multzo honetan subjektu ergatiboa, objektu absolutiboa eta datibozko objektua dituzten adibideak biltzen ditugu. Adibidez:

```
|   |   adibidea-3:
|   |   |   identifikadorea: jaurti-A0.-DU-3
|   |   |   adibide-katea: Atezainak aurrelari bati jaurti dio pilota.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: jaurti dio
```

```
|   |   |       posizioa: 3
|   |   |       pertsona  NR-HU  NI-HU  NK-HU
|   |   |       adi-lag-1:  *edun  dio  @+JADLAG
|   |   |       f-adi-sint-1:  jaurti  @-JADNAG
|   |   |       aspektua: BURU
|   |   |       modua-denbora: A1
|   |   |   sintagma-1:  Atezainak  1  IZE  ERG  NUMS  MUGM  ARR  @SUBJ
|   |   |   sintagma-2:  aurrelari bati  2  DET  DAT  NUMS  MG  @ZOBJ
|   |   |   sintagma-3:  pilota  4  IZE  ABS  NUMS  MUGM  ARR  @OBJ
```

Patroi bereko aditz ezjokatuarekin aurkituriko adibide bat:

```
|   |   adibidea-1:
|   |   |   identifikadorea: errefusatu-A0.-DU-1
|   |   |   adibide-katea: Gurasoek haurrei janaria errefusatu.
|   |   |   sarrerako-aditz-katea
|   |   |       katea: errefusatu
|   |   |       posizioa: 4
|   |   |       f-adi-sint-1:  errefusatu  @-JADNAG
|   |   |   sintagma-1:  Gurasoek  1  IZE  ERG  NUMP  MUGM  ARR  @SUBJ
|   |   |   sintagma-2:  haurrei  2  IZE  DAT  NUMP  MUGM  ARR  @ZOBJ
|   |   |   sintagma-3:  janaria  3  IZE  ABS  NUMS  MUGM  ARR  @OBJ
```

**@PRED_ABS**

Azkenerako utzi dugun multzo honetan, izenki-predikatuaren funtzioa (@PRED) hartzen dugu kontuan azaleko argumentu-egituraren parte gisa, nahiz eta sailkapen automatikorako irizpideen artean ez kontuan izan. Hona hemen, multzo honetarako hautatu dugun pare bat adibide:

*Leial agertu zaitez* (agertu-H1.MA.N1.-DA-DU-1)

*Ordezkari hautatu dutenari* (hautatu-A1.-DU-3)

## VI.4 Automatikoki erdietsitako azaleko patroiak: zailtasunak eta ebaluazioa.

Atal honetan, garatutako metodologiaren bidez aditzen sailkapena egiterakoan izandako zailtasun nagusiak, eta sailkatze hori zenbateraino izan daitekeen fidagarria izango ditugu mintzagai. Zailtasunei dagokienean, azaleko sintaxiaren mugak, posizioaren erabilgarritasun mugatua eta aditzen adibideon zenbait ezaugarri aipatuko ditugu. Ondoren, erdietsitako sailkapenaren ebaluazioa egingo dugu. Hau da, patroi bakoitzaren fidagarritasunaren neurri batzuk emango ditugu, lagin baten gainean egindako azterketaren ondorioz.

## VI.4.1 Azaleko sintaxiaren mugak.

Azaleko sintaxiaren atala garatzerakoan, 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 —ikus § V.1.6—. Beraz, kontuan hartu behar da zer ezagutzeko kapaz garen (hau da, zer kate) eta zer ez —ikus § V.1.6—. Horrez gain, atal horren bukaeran —ikus § V.1.6.3— ebaluatu dugu kate-osatze urrats hori, ezagutu ahal dugun horretan zer-nolako arazoak edota akatsak ditugun jasoaz. Ezagutzeko kapaz garenaren %79a ongi etiketatua azaltzen zaigu, hau da, %79an zatiak ongi eratuta daude. Horrez gain, kontuan hartu behar da ongi markatzen diren zati horietatik jasotzen diren funtzioak eta kasuak zer-nolako fidagarritasuna duten —ikus § V.1.6.4—.

Aipatu ditugun atal horietan esandakoak kontuan hartzeaz gain, azpimarratu beharra dago azaleko sintaxiaren bidez ezin dugula perpaus nagusi eta mendekoaren arteko harremanik adierazi. Eta beraz, aditz bat baino gehiagoko adibideetako informazioa ezin izan dugu baliatu sailkapen automatikorako. Ikus dezagun adibide bat:

*Liburu askoz **baliatu** dira idazlan hori prestatzeko.*

Adibide hau *baliatu* aditzaren erabilera erakusteko landu du lexikografoak. Eta gure metodologia aplikatuz, ezin izango genuke bereiztu adibide horretan *baliatu*-k ez duela objekturik (*idazlan hori*) hartzen eta *prestatu*-k aldiz bai. Hau da, ez gara gai adibidea honela ikusteko:

*Liburu askoz **baliatu** dira [idazlan hori prestatzeko.]*

Horrelako esaldien analisi sakonagoa egiteko, azpikategorizazioaren informazioa lexikoian zehaztuta egon behar litzateke. Baina, lehenago aipatu dugun legez azaleko sintaxia lantzerakoan ez dugu horrelako informaziorik baliatzerik izan.

Ditugun baliabideekin, garatu ahal izan dugun analisi sintaktikoaren bidez, aditz bat baino gehiago dituzten adibideetarik oso zaila da automatikoki erabakitzea argumentu (edota argumentu izateko gai) bakoitza zein aditzi dagokion. Honelako adibideetatik automatikoki jasoriko informazioak akats gehiago izango lituzke. Hau horrela gertatzen da garatu dugun azaleko sintaxiak ez duelako halako arazoei aurre egiteko aukerarik ematen. Emaitza horiek eskuzko lan handia eskatuko lukete automatikoki atera denetik patroi zuzena erabakitzeko. Horren ordez, nahiago izan dugu automatikoki jasoriko informazioa fidagarriagoa, eta eskuz egiaztatzerakoan lan gutxiago emango duena izatea. Horrek eraman gaitu aditz bakarreko adibide guztiak aztertzera, eta aldiz, aditz bat baino gehiagoko adibideetatik zenbait hautatzera posizioaren erabilgarritasuna aztertzeko.

## VI.4.2 Posizioaren erabilgarritasunaz.

Posizioa baliatu dugu aditz bat baino gehiago dituzten adibideetan aditz bakoitzari zein sintagma (edota mendeko perpausa) dagokion zehazteko lanean laguntzeko. Horrela bada, ezaguturiko zati sintagmatikoei nahiz aditz-kateei zenbaki bat erantsi zaie zein ordenatan agertzen diren markatuz. Jakin badakigu, euskaraz ordena librea dela eta ordenak ez duela argumentuek zein funtzio beteko duten erabakitzen (fokalizazio kasuak salbu, orduan justu aditzaren aurrean jarriko baita fokalizatutako elementua). Dena den, gure ideia da, argumentugai diren sintagmak edota aditz-kateak ez direla nonahi azaltzen, baizik eta seguruenik azpikategorizatzen dituen aditzarengandik hurbil azalduko direla. Ideia horri jarraituz, aditz bat baino gehiago zituzten adibideak moztu egin ditugu, honako kasu hauek aurreikusiz:

- Aztertzen ari garen aditza bestelako aditz-katea baino lehenago azaltzen zaigunean, bestelako aditz-kate horren ondoren datozen elementuak ez ditugu jasoko.

- Bestelako aditz-katea aztergai dugun aditz-katea baino lehenago azaltzen baldin bada, bestelako aditz-katearen aurretik dauden elementuak ez ditugu jasoko, hots, bestelako aditz-kateak duen posizio-zenbakia baino txikiagoa dituzten elementuak baztertu egingo dira.

Horietako lehenengo kasuari erreparatuz, hau da, aztertzen ari garen aditzaren ondoren bestelako aditz-kate bat azaltzen baldin bada, bestelako aditz-kate horren ondoren azaltzen diren sintagmak ez jasotzea erabaki dugu. Hau da, hain zuzen ere bestelako aditz-kate horretan moztuko genuke adibidea, baina bestelako aditz-katea bera jaso egingo genuke, ez baitakigu aditz hori aztertzen ari garenok azpikategorizatua izan daitekeen ala ez. Adibidez:

- Jatorrizko adibidea:

*Zure okerrak **tapatu** nahirik egin dituzu pausuak, zer enganio egin didazun jakitun daude auzoak.*

- Jatorrizko adibidea posizioaren nozioa baliatu ondorik, hots, jatorrizko adibidea moztuta:

*Zure okerrak **tapatu** nahirik egin dituzu ...*

Egin duguna da hiztegian dagoen adibidea moztu, gure azterketa moztutako zati honetara mugatzeko. Hipotesia da, esaldia moztu ondoren lortu dugun adibidean aztergai dugun aditzari dagokion informazio pertinentea aurkitzen dela, eta, aldiz, mozketaren ondorioz baztertu dugun adibide puska horretan ez dela sarrerako aditzari dagokion informaziorik aurkitzen. Mozketa egiteko baliatu dugun nozio honek, alabaina, zenbait kasutan huts egiten du, esate baterako bi aditz koordinatu ditugunean. Honelakoetan esan daiteke biek konpartitzen dituztela osagai berberak, baina gure kasuan azterketatik kanpo utziko ditugu. Adibidez:

-Jatorriko adibidea: *Edanak eragiten ditu eta* ***erasaten*** *gauza lotsagarriak*

-Moztutako adibidea: *...eragiten ditu eta* ***erasaten*** *gauza lotsagarriak*

Ikus daitekeen bezala *edanak* aztergai dugun aditzaren aztereremutik kanpo utzi dugu, baina *erasaten*-en subjektua ere bada.

Gure kasuan, azaleko sintaxia landu dugunez, azaleko posizioa hartu dugu irizpide lagungarritzat; nahiz eta jabetzen garen oso konplexua dela gaia, adibideak mozteko egokia dela iruditzen zaigu ikuspegi praktikotik. Posizioaren aplikagarritasuna handiagotzeko, oso garrantzitsua litzateke juntagailu, lokailu eta puntuazio-ikurrak baliatu ahal izatea. Hau da, elementu horiei posizioa esleitzea eta adibidea mozteko erreferentzia gisa erabili ahal izatea. Dena dela, elementu horien erabilpena azterketa honetatik kanpo geratu da.

## VI.4.3 Adibideen bestelako arazoez.

Atal honetan aditzen sailkapenari eragiten dioten aditzen adibideon zenbait ezaugarri aipatuko ditugu:

A) Adibideen egokitasun eza. Zenbait adibide baztergarritzat jo behar ditugu, azpikategorizazioa automatikoki lantzeko jasotzen dugun informazioaren ikuspegitik. Hau da, zenbait adibide esaldi ezosoak ditugu, eta hauetatik automatikoki jasotzen den informazioak, aztergai dugun aditzari ez dagozkion egitura sintaktikoak izango ditu. Okerreko informazio hori jasotzearen arrazoia, kasu honetan adibide horiek esaldi ezosoak izatean datza. Adibidez: *Zaldiak alhatzen diren soroa* (alhatu-A1.Xa.-DA-1)

Hau da, erlatibozko bat dugu adibide gisa *alhatu*-ren erabilera jasotzeko, eta azaleko analisiari erreparatuz subjektu absolutiboa (*zaldiak*) jasoko genuke, baina, *soroa*-k dituen funtzioetatik bat bera ere ez dagokio *alhatu* aditzari. Hau da, *soroa*-k objektu eta subjektuaren funtzioak izango ditu baina horietatik ezin dugu bat bera ere aukeratu. Horietako bat hautatzeko, erlatibozko perpausak zein funtzio betetzen duen zehaztu beharko genuke, beste aditz batek horrelaxe eskatuz gero. Baina, erlatibozkoa azaltzen denez soilik, hautapen hori ezin da egin. Beraz, horrelako adibideetatik jasotzen den informazioak errore-kopuru handiagoa izango du.

B) Errore gramatikalak dituzten adibideak. Zenbait kasutan, halako erroredun adibideak aurkitu ditugu, esaterako, komunztadura-akatsak aurki daitezke, adibidez:

*Zein gogorki ferekatzen gintuzten eguzkiaren urrezko izpiak.*

Jakina, honelako adibideetatik jasotako informazioa okerra izango da, eta, beraz, baztertu beharrekoa.

C) Adibideetako akats tipografikoen eragina. Badira zenbait adibide formatuan dituzten karaktere bereziengatik edota adibideetan azaltzen diren errore tipografikoengatik erroreak ematen dituztenak. Hala ere, honelako erroreak gutxitan gertatzen dira. Adibidez:

***Auspoa-k** sorberritu dizkigun atzoko idazleak* (formatu kontuengatik Auspoa eta -k bereiz azaltzen zaizkigu)

*Niri **faltza** zaidana zuk osa ezazu* (errore tipografikoa)

D) Aditzak adjektibo gisa:

Ezjokatuen kasuan gertatzen den fenomeno bat da aditza adjektibo gisa erabilia izatea. Adibidez: *Arantza **bihurkatuz** eginikako koroa bat.* Hona hemen adibide gehiago:

amorratu-A1.N2.-DA-DU-3: *Jokalari **amorratuen** artean.*

asmatu-A.N1.-DU-1: *Hitz **asmatua**.*

atxilotu-N1.-DU.1: *Txori **atxilotuak**.*

azpildu-A1.-DU-3: *Zapi **azpilduak**.*

bordatu-A0.-DU-3: *Soineko **bordatua**.*

Interesgarria litzateke aztertzea, ea adjektibo gisa erabiliak agertzen diren aditz horiek elkarren artean tasun antzekorik duten. Automatikoki lorturiko laginetan 98 aldiz agertu dira aditzak adjektibo gisa erabiliak.

E) Aditza hitz-elkarketaren parte gisa.

Zenbait kasutan aditza hitz-elkarketaren parte gisa ere agertu da. Adibidez:

aspertu-A5.-DU-2: *Negar-**aspertu** bat nahi eta malkorik ezin lehertu.*

Aditzak hitz-elkarketan parte hartzearen gaia gure azterketatik at geratzen da.

## VI.4.4 Patroien ebaluazioa.

Aditzen sailkapenean eman ditugun azaleko patroi horien ebaluazioa egitea oso garrantzitsua da patroi horien fidagarritasuna neurtzeko. Horretarako, V.2.7.4 atalean eginiko azterketan oinarritu gara. Hau da, atal horretan azaldu irizpideei jarraituz, patroi bakoitzeko aztertu duguna da ea, funtzio sintaktiko/kasuei erreparatuz, zenbat kasutan asmatzen den eta zenbatetan huts egin den patroia osatzen duen funtzio sintaktiko/kasua. Ondoren, aztertu laginean (1211 adibide, horietatik 646 aditz bakarrekoak) agertu den patroi bakoitzeko, adibide-kopurua eta ebaluazioaren emaitzak jasotzen dituen taula emango dugu. Taula horietan jasotzen diren emaitzak aditz bakarrekoei dagozkienak dira. Gainontzeko adibideak ebaluazio

honetatik at geratzen dira: aditz bat baino gehiago dituzten adibideak (406) eta aditzen sailkapenerako erabili dugun funtzio sintaktiko/kasurik azaltzen ez duten adibideak (159).

Ebaluazioaren emaitzak, sailkatze automatikoa eta eskuzkoa erkatzearen ondorio dira. Patroi bakoitzeko, aditzak sailkatzeko aintzat hartu diren funtzio sintaktiko/kasuak hartu dira kontuan. Esan bezala, funtzio/kasu bakoitzeko, asmatu den ala ez neurtzen dugu, emaitza horiek taula batean jasoaz. Baina, horretaz gain eskuz analizatutako laginean agertzen ez diren, eta analisi automatikoaren bidez markatzen diren funtzio sintaktikoen berri ere ematen dugu. Hona hemen erdietsitako emaitzak:

1. Patroia: OBJ_ABS; kopurua: 293

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| OBJEKTU GISA MARKATUAK | 293 | 200 | 93 |
| ZEHAR OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |

Esan beharra dago patroi honetako adibideen artean, eskuzko analisian ez dauden funtziodun sintagmak aurkitu direla markaketa automatikoan: 93 subjektu eta 5 zehar objektu.

2. Patroia: SUBJ_ERG; kopurua: 38

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 38 | 32 | 6 |
| OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| ZEHAR OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |

Patroi honetako adibideetan eskuzko analisiarekin bat ez datozen 7 objektu aurkitu dira.

3. Patroia: SUBJ_ABS-ZOBJ_DAT; kopurua: 5

| SINTAGMAK | GUZTIRA | ASMATU | *HUTS EGIN* |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 5 | 3 | 2 |
| OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| ZEHAR OBJEKTU GISA MARKATUAK | 5 | 5 | 0 |

Patroi honetan, analisi automatikoaren emaitzan eskuzkoan azaltzen ez den objektu bat dugu.

4. Patroia: SUBJ_ERG-OBJ_ABS-ZOBJ_DAT; kopurua: 1

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 1 | 1 | 0 |
| OBJEKTU GISA MARKATUAK | 1 | 0 | 1 |
| ZEHAR OBJEKTU GISA MARKATUAK | 1 | 1 | 0 |

Analisi automatikoan eskuzkoan ez ditugun lau subjektu agertzen zaizkigu.

5. Patroia: SUBJ_ABS; kopurua: 79

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 79 | 48 | 31 |
| OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| ZEHAR OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |

Analisi automatikoan, eskuzkoan ez dauden honako funtzio hauek aurkitu ditugu: 14 objektu eta 2 zehar objektu.

6. Patroia: OBJ_ABS-ZOBJ_DAT; kopurua: 8

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| OBJEKTU GISA MARKATUAK | 8 | 7 | 1 |
| ZEHAR OBJEKTU GISA MARKATUAK | 8 | 8 | 0 |

Analisi automatikoan, eskuzkoan ez dauden 3 subjektu daude markaturik.

7. Patroia: ZOBJ_DAT; kopurua: 6

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| ZEHAR OBJEKTU GISA MARKATUAK | 6 | 5 | 1 |

Patroi honetan, analisi automatikoak eskuzko analisian ez dauden 2 subjektu eta objektu bat markatu ditu.

8. Patroia: OBJ_PAR; kopurua: 2

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 0 | 0 | 0 |
| OBJEKTU GISA MARKATUAK | 2 | 0 | 2 |
| ZEHAR OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |

Patroi honetan, analisi automatikoak eskuzko analisian ez dauden 2 subjektu markatu ditu.

9. Patroia: SUBJ_ERG-OBJ_ABS; kopurua: 54

| SINTAGMAK | GUZTIRA | ASMATU | HUTS EGIN |
|---|---|---|---|
| SUBJEKTU GISA MARKATUAK | 54 | 42 | 12 |
| OBJEKTU GISA MARKATUAK | 44 | 10 | 34 |
| ZEHAR OBJEKTU GISA MARKATUAK | 0 | 0 | 0 |

Azken patroi honen kasuan ere, analisi automatikoak eskuzko analisian ez dauden 5 subjektu markatu ditu.

Oro har, azaldu ditugun emaitzetan ikus daiteke patroian subjektuak edo objektuak bakarrik hartzen duenean parte, asmatze-tasa txikiagoa izaten dela beste funtzio batekin konbinatzen direnean baino. Esaterako, lehenengo patroiarekin lortutako emaitzak okerragoak dira seigarrenarekin edo bederatzigarrenarekin lortutakoak baino. Hain zuzen ere, funtzio horiek ongi etiketatzeak ematen ditu arazorik handienak. Hala ere, bigarren patroiaren emaitzak nahiko onak dira. Fidagarritasunari begira, hirugarrena eta zortzigarrena ez dira oso fidagarriak; seigarrena eta zazpigarrena, aldiz, fidagarrienak ditugu.

## VI.5 Ondorioak.

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 metodologia bera sendotzeko. Horretarako, lorturiko informazioa lexikoian integratu beharko litzateke gure tresna sintaktikoetan aplikatu ahal izateko. Esan beharra dago erdiesten den azpikategorizazioari buruzko informazioa nola txertatu lexikoian edota parserrean analisi sintaktikoan baliagarri suertatzeko, azterketa sakonago baten premian dagoela.

Horrekin batera, azpikategorizazioa aztertzeko bidea erraztearen xedea bete dugulakoan gaude. Hau da, gure hurbileko helburu praktikoetatik aparte ere, esan beharra dago eginiko multzo horiek azterketa sakonagoetarako material baliagarria direla.

Hiztegiko adibideak aditzaren portaera aztertzeko iturri egokiak direla uste bagenuen, oraingo urratsen ondorioz areagotu egin da egokitasun hori. Adibideak sintaktikoki etiketatuak baitaude, eta oinarrizko zatiak detektatuak baitituzte. Eta gainera, interesgarri deritzogun hainbat tasun jaso dira. Horrez gain, informazio hori guztia ustiatu ahal izateko SGMLz

adieraztean, testu huts izatetik formatu aberatsago batera aldatu ditugu. Hain zuzen ere, formatu horretan adierazteak galdeketa-sistemaren garapena ahalbideratzen du; aztertzeko bideak eta aukerak ugalduz. Tasun ugari jaso direnez gero, horien arabera egokiera ematen da aditzen portaera modu desberdinetara azaltzeko. Gure kasuan, erakutsi dugun legez —ikus VI.3 atala— kasua eta funtzio sintaktikoaren informazioa baliatu ditugu aditzak multzokatzeko.

Aditzen sailkapena erdiesteko, azaleko sintaxia landu dugu, aditz-kateak eta hauen inguruan dauden sintagmak ezagutuz. Baina, unitate sintaktiko horien analisi sintaktikotik areago joan nahi izanez gero, parsing sakonagoa beharko genuke. Hau da, aditzei buruzko azpikategorizazioa zehazteak sintagma edota aditz-kateen analisitik esaldi konplexuagoen analisira jauzia egiteko aukera emango luke. Jakina, parsing sakonago bat garatzeko azpikategorizazioaren informazioa behar-beharrezkoa dugu. Hain zuzen ere, arazo honi aurre egiteko azpikategorizazioa zehaztua beharko genuke izan lexikoian. Baina, gogora dezagun, gure kasuan funtsezkoa den informazio hori gabe abiatu garela ahal den heinean aditz baten inguruan azaltzen diren sintagmak eta aditz-kateak atzematera. 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.

Hala eta guztiz ere, lehenago esan bezala, uste dugu eginiko azaleko analisia baliagarria dela azpikategorizazioaren alorra lantzen laguntzeko.

# AURRERA BEGIRAKOAK ETA ONDORIOAK

# VII. Etorkizuneko lanak eta ikerlerroak.

Tesi-proiektu honetan ikergai geratutakoak eta, gure ustez, aurrera egiteko bide interesgarrienak diratekeenak azalduko ditugu kapitulu honetan, labur-labur.

## VII.1 Hiztegigintzarako bideak urratu: informazio lexikalaren eskuratzearen zein errepresentatzearen aldetik.

Ikerlan honetan, LNPko teknikak garatu ditugu LNPrako baliabide lexikalak osatzeko helburuarekin. Uste dugu, gainera, teknika horiek hiztegigintzarako ere baliagarriak izan daitezkeela. Hau da, bai LNPrako bai giza erabiltzaileari zuzenduriko hiztegiak, lexikografoen iritzietan oinarritzen dira gehiago testuetan baino. Horren ordez, interesgarria litzateke LNPko teknikez baliatzea hiztegiak eraikitzerakoan, esate baterako, informazioa corpusen azterketatik eskuratuz. Ildo horretatik, landu dugun metodologia aditzen azpikategorizazioari dagokion informazioa eskuratzen laguntzeko baliagarria izan daitekeelakoan gaude. Eta lorturiko azaleko patroien ondorioz, azpikategorizazioa erdiautomatikoki erdiesteko bidea jorratzeari interesgarri deritzogu.

### VII.1.1 Azpikategorizazioa erdiesteko bideak hedatu.

Landu dugun metodologiak erakusten du azpikategorizazioaren informazioa eskuratzeko bide berriak urra daitezkeela. Eta, gainera, bide horiek azpikategorizazioaren azterketa sakonagoetarako lagungarriak izango direlakoan gaude. Ildo horretatik, eginkizun interesgarri eta erronka gisa geratzen dira, batetik, taldean baterakuntza-formalismoarekin corpus

orokorren gainean egiten ari den azpikategorizazio-lana[1] eta hiztegiko adibideen gainean egin dugun hau uztartzea linguistari informazio osatuagoa eskaintzeko, eta bestetik, *bootstrapping*-ari begira lehen mailako azpikategorizazio-emaitza hauek erabiliz analizatzaile hobeak eraikitzea informazio zehatzagoa lortzeko.

Uste dugu azterketan jarraituriko urratsen emaitzek azterketa sakonagoak egiteko material aproposa osatzen dutela. Eta, beraz, multzo horien gainean azterketa sakonagoak bultzatzearen alde gaude. Galdeketa-sistemaren bitartez, definituriko galdera baten arabera adibide multzo jakin bat eskuratzen dugu. Baina, galderak bestelako irizpideak erabiliz ere defini daitezke; galdeketa-sistema irekia baita. Horrela, bada, aurrera begira eman daitezkeen urratsen artean, galdera berriak definitzeari interesgarri deritzogu. Eta, horrez gain, erdietsi ditugun multzoen gainean egin daitezkeen azterketa sakonagoen artean, aditzak multzokatzeko zein tasunez balia gaitezkeen erabaki behar litzateke. Hau da, aditz bakoitza kokatzen deneko multzoa zein ezaugarrik definitzen eta bereizten duen gainontzekoetatik.

Bestalde, egin ditugun urratsak corpus zabalago batean aplikatzea oso garrantzitsua da. Batetik, jarraitutako bidearen sendotasuna frogatzeko, eta, bestetik, hiztegiko corpus "bitxi" horretatik eskuratu ditugun azaleko patroi horiek corpusetan ea zenbatetan agertzen diren egiaztatzeko eta hiztegietatik jaso ez diren patroi berriekin aberasteko. Horrez gain, corpusaren azterketa zabala eta sailkapen zurrunago bat konbinatuz gero, azpikategorizazio-patroien maiztasuna baliatu ahal izango genuke. Informazio hori ere garrantzitsua dugu, esaterako (Carroll eta Minnen, 1998)-n parserren doitasuna hobetzeko baliagarria dela frogatzen baitute.

## VII.1.2 Errepresentazioaren alorra landu.

Informazioaren eskurapenaren ikuspegiaz aparte, errepresentazio formalak ere asko du irabazteko, tresna konputazionalak erabiliz gero. Beraz, uste dugu interesgarria litzatekeela hiztegigintzan LNPko teknikez baliatzea eta formatu estandarrak kontuan hartzea.

Horrela bada, hiztegiak eraikitzerakoan interesgarri litzateke informazioa eskuratzeko LNPko tresnez baliatzea, eta horrez gain informazioa jasotzeko kodeteka estandarra jarraitzea. Ildo honetatik, aurrera begira EHrako berrantolaketa dugu helburu, jada, dagoen informazio oro TEIko gidalerroetara egokituz eta etorkizunean egin daitezkeen aldaketak zein aberasketak TEI gidalerroei jarrituz egin behar liratekeela uste dugu. TEIra egokitze sendoa burutzeko, automatikoki erdietsi dugun TEIko bertsioa hartu, eta horren gainean orrazketa-lana eskuz burutu beharra dago. Behin, TEIko bertsio sendo bat erdietsiz gero, etorkizunean egin

[1] Euskal aditzaren azpikategorizazioa. Azterketa sistematiko-automatikoa", 1996ko Eusko Jaurlaritzaren beka-deialdian onartutako proiektuaren inguruan Izaskun Aldezabal garatzen ari den tesia. Interesgarri litzateke aztertzea ea, taldean azpikategorizazioaren bidetik egin diren zein egiten ari diren lanetan, zenbateraino lagun dezaketen adibideetatik erdietsitako emaitzek eta urraturiko bideek.

daitezkeen aldaketa, aberasketa zein edizio-mota desberdinak modu azkarrago eta ziurrago batean burutuko lirateke.

Bukatzeko, esan beharra dago LNPko komunitatearen, lexikografoen eta argitaratzaileen elkarlana bultzatzea oso garrantzitsua dela. Halaxe diote (Ide, N. & Veronis, J., 1994:284)-en ere:

> "The most promising avenue of activity, however, involves collaboration between the NLP community and lexicographers and electronic publishers."

# VII.2 Aurrera begira sintaxia nola landu.

Sintaxiaren arloan nondik jo daitekeen erabakitzea saihestezina dugu taldean. Ondokoan, gure ustez, egindako bideari jarraiki egin ahal denari lotuko gatzaizkio.

## VII.2.1 Landutako azaleko sintaxiaren bidea jorratu.

Ireki diren bideen artetik ondorengo hauek urratzearen interesa azpimarratu nahi genuke:

- *Bootstrapping* lexikala: hau da, erdietsitako azaleko emaitzak landu dugun azaleko gramatikan berrerabiltzea litzateke gakoa. Hartara, bi alderdi garatuko genituzke: batetik, analisi sintaktikoa bera, eta, bestetik, informazioa eskuratzeko metodologia. Beraz, bi alderdi horiek aberastean azpikategorizazioa lantzeko laguntza handiagoa eskaini ahal izango genuke.

- Corpusera hedatu landutako azaleko sintaxia. Hartara, landu dugun sintaxiaren sendotasuna zenbaterainokoa den azter daiteke. Horrez gain, corpus orokor bati aplikatzeak corpus horren azterketarako ateak irekiko lituzke. Hau da, corpus hori sintaktikoki etiketatu ondoren, horren gainean galdeketa-sistema erabiliz modu azkarragoan azter ditzakegu adibide kopuru handiago bat (eskuz egitearen aldean). Etiketaturiko corpusari eta galdeketa-sistemari esker, azkarrago eta errazago eskura zein azter daiteke corpusean gorderik dagoen informazio linguistikoa. Bestalde, corpusaren gainean lan egiteak landuriko metodologia ebaluatzeko balioko liguke, ikusteko zer-nolako emaitzak lortzen diren hiztegiko adibideetatik kanpo.

- Zatien osaketarako teknika berriak baliatzearen interesa: zatien osaketarako garatu ditugun gramatikez aparte, zatiak atzemateko beste teknika batzuk azter daitezke gramatika horietan adierazten den informazioa aprobetxatuz. Esaterako, egoera finituko formalismoek eskaintzen duten tresneria baliatzearen aukera da azter daitekeenetariko bat.

- Azpiesaldiak mugatzeko irizpideak aztertu: hau da, esaldian aztergai dugun aditzaren inguruan dauden elementuak zein diren mugatzeko zereginean lagungarri izan daitezkeen elementuen azterketa sakondu. Hartara, baliatu dugun posizioaren nozioaz gain beste elementu batzuk baliatu ahal izanez gero, metodologia sendotu egingo litzateke.

- Posposizio-egituren lanketa: zenbait posposizio-egitura ezagutzen ditugun arren, zabaldu egin behar litzateke hauek ezagutzeko dugun ahalmena. Egitura horiek ezagutzeak badu zerikusirik aditz bakoitzarentzat jasotzen diren azaleko patroi horiekin. Beraz, azaleko sintaxiko urrats hori indartu ahala erdietsiko litzatekeen informazioak gero eta eskuzko lan gutxiago eskatuko luke.

Eta, azkenik, lan-taldeko ikerkuntza-bideak, urratuak zein urratzen ari direnak[2], kontuan hartu behar dira, aurkeztu dugun lan hau, batetik, horietan integratzeko, eta bestetik, gauzatzen laguntzeko.

## VII.2.2 Dependentzia-egituren bidetik jarraitu.

Landu dugun azaleko sintaxian CG formalismoa izan dugu oinarri, eta formalismo horrek eskaintzen digun tresneria baliatuz zatiak ezagutzeko hastapeneko urratsak egin ditugu. Baina, oraingoan, gure kezka zati horiek nola lot daitezkeen da. Hau da, sintaxia sakonago lantzeko egin beharreko pausuek arduratzen gaituzte. Ardura horri erantzuteko, irtenbideetariko bat dependentzia-gramatiken bideari jarraitzea dela uste dugu. Esate baterako, ildo horretatik egiten dute lan *A Dependency Parser of English* lanean (Järvinen eta Tapanainen, 97). Lan horren ideia nagusia da erlazio sintaktikoak esplizitu egitea, hau da, dependentzia funtzionalak loturen (*link*) bidez adieraztea.

Hauen ideia nagusiari jarraituz gero, etorkizuneko lanen artean orain arte egin dugun analisian inplizitu dauden guneak azaleratzea, oinarrizko elementu sintaktiko gisa hartuz, eta dependentzia funtzionalak loturen bidez adieraztea lirateke.

Bestalde, guneak esplizituki adierazteak hautapen-murriztapenak aztertzeko bidea erraztuko liguke. Hau da, adibidez, aditzen kasuan aditz bakoitzarekin azaltzen diren guneak biltzerakoan ikus daiteke zer-nolako tasun (bizidun, bizigabe, gizakia, etab.) semantikodun guneak eskatzen dituen aditz bakoitzak.

Horrez gain, esan beharra dago, zenbait autoreren iritziz hobeak direla dependentzia-egiturak ordena libreko hizkuntzetarako (Skut eta beste, 97; Tapanainen, 98; Oflazer 99).

---

[2] Arlo honetan bukatzear dauden bi tesi-lan aipatu behar dira: (i) "Morfologiatik Sintaxira Murriztapen Gramatika baliatuz" (Itziar Aduriz), eta (ii) "Ezagumendu sintaktiko eta semantikoen azterketa errore ortografikoak zuzentzeko" (Koldo Gojenola).

Beraz, etorkizuneko lanen artean, sintaxia aurrera begira lantzeko dependentzia-egituren bidetik jarraitzeari ongi deritzogu.

### VII.2.3 Sintaktikoki etiketaturiko corpusak sortu beharra.

Sintaxiaren bidean aurrera egiteko, oso garrantzitsua litzateke baita ere, *treebank* edota sintaktikoki etiketaturiko corpusak sortzea. Hauen garapena oso interesgarria ikusten dugu, informazioa erauzteko zein tresnen ebaluaziorako.

Jakina, arazo konplexuagoa da corpus sintaktikoki analizatuak eskala handian sortzea, testu soilak biltzea edota kategoria gramatikalarekin etiketatzea baino (Garside & Leech, 1987; Garside, 1993).

Tesi honetan landutako azaleko sintaxia abiapuntu gisa hartuta, tresna erdiautomatikoak definitu beharko dira sintaxi-etiketatze handiak egin ahal izateko.

### VII.2.4 Ebaluazio-sistema landu beharra.

Ebaluazioak berebiziko garrantzia du LNPko aplikazioen alorrean. Gure kasuan egin duguna lehenengo saioa izan da, azaleko sintaxiko etiketatzea eta eskuratzen ziren azaleko patroiak ebaluatuz. Saio horren ondorioz, ebaluaziorako sistema orokorragoak erdiesteko bideak urratzeari oso interesgarri deritzogu, besteak beste, urratzen ari garen sintaxiaren alorrean emaitzak neurtzeko eta hobetzeko balio dezakeelako. Beraz, badira aztertu beharreko hainbat ikerlan, aurrera egiteko ebaluazio erabilgarriena zein den erabakitzen lagunduko dutenak. Hortxe dugu, esaterako, (Carroll eta beste, 99)-n azaltzen den ebaluazio-sistema, egokia izan daitekeena sintaxia ebaluatzeko. Lan horretan, ebaluazio-sistema dependentzia-egituren antzeko erlazio gramatikaletan oinarritzen dute.

# VIII. Ondorioak.

Kontuan harturik gogoan izan ditugun motibazio nagusiak (1) EH berrerabiltzea, EDBLren aberasketarako eta (2) aditzen argumentu-egitura lantzen laguntzeko bideak eskaintzea izan direla, esan dezakegu aipatu motibazioei erantzun diegula eginiko lanaren bidez.

Lehendabizi, EHren egitura definitzen duen gramatika zehaztu dugu eta hiztegia bera analizatu. Lan horiek gauzatzean lexikografoak hiztegia egiterakoan buruan duen gramatika azaleratu dugu. Eta, gainera, hiztegiko artikuluak eta artikulu bakoitzaren atalak egituratzeko testu-prozesadore batez baliatzeak dakartzan gabeziak eta akatsak nabarmendu ditugu. Horrek guztiak erakusten du formalizazio zorrotzago baten premia nabaria dela.

Horrez gain, hiztegia TEIko gidalerroen arabera kodetu dugu. Hau da, formatu estandar batez baliatu gara hiztegia errepresentatzeko. Eta, hiztegigintzari begira, aurkeztu dugun TEI ekimeneko gidalerroez baliatzea izan daiteke hiztegien kontsistentzia ziurtatzeko jarrai daitekeen bideetariko bat. Adibidez, TEIko gidalerroak, egokiak dira oso artikuluak idazten diharduen lexikografoarentzat, datuen zuzentasuna, osotasuna eta abar egiaztatzearen aldetik.

Bestalde, TEIra egokitze horrek hiztegiaren berrerabilgarritasunari irekitzen dizkio ateak. Hau da, batetik EDBLren aberasketarako informazioa jartzen du eskuragarri, eta, bestetik, EH aztergai edo lantresna duen edonorentzat baliagarri izango da. Aipatu gabe hiztegiaren etorkizuneko eguneratzeetarako eskaintzen dituen abantailak.

Berrerabileraren ildotik, gure kasuan aditzen adibideak izan ditugu aztergai. Azterketa horren helburu nagusia adibide horietan aditz bakoitzaren inguruan azaltzen diren sintagmak eta aditz-kateak jasotzea izan da. Eta helburu horrek eraman gaitu adibideen azaleko analisi sintaktikoa egitera.

EHko aditzen adibideak analizatu ditugu. Adibideok analizatzeko euskararen sintaxiaren parte bat konputazionalki deskribatu eta erabili dugu. Eta ondorioz, sailkapen bat (etorkizuneko azterketa sakonagoen euskarri nahi litzatekeena) erdietsi dugu, aditzen argumentu-egiturari erreparatuz.

Landu dugun sintaxiaren partea azaleko sintaxiaren arloan kokatzen da. Azaleko sintaxiak berebiziko garrantzia du adibideetatik jaso den informazioaren zuzentasunerako. Izan ere, aditz bakoitzaren argumentu posible gisa jasotzen direnak, azaleko sintaxiaren bidez ezagutzen diren sintagmak eta aditz-kateak baititugu.

Ez dugu zalantzarik esateko aditzei buruzko azpikategorizazioa zehazteak sintagma edota aditz-kateen analisitik esaldi konplexuagoen analisira jauzia egiteko aukera emango duela.

Uste dugu eginiko azaleko analisia baliagarria dela azpikategorizazioaren alorra lantzen laguntzeko, hain zuzen ere, proiektu honen bigarren motibazio nagusia izan denari erantzunez. Analisiaren emaitza errepresentatzeko SGML baliatu dugu analisi sintaktikoa errepresentatzeko DTDa definituz. Errepresentazio-modu horrek analisia testu huts izatetik errepresentazio aberatsago batera moldatzean, ustiapena errazteko bideak irekitzen ditu. Ustiapen hori burutzeko moduetariko bat, diseinatu eta inplementatu dugun galdeketa-sistema dugu. Galdeketa-sistemaren bidez, analisietatik eskuratu nahi dugun informazioa jasotzeko eta ikerketa errazteko bidea landu dugu. Horren bidez defini daitezkeen galderek analisiaren emaitzak aztertzen lagundu digute, lorturiko emaitzak modu desberdinetara antola daitezkeela.

Azkenik, egindako lanen ondorioz hiru ekarpen nagusi egin ditugula azpimarratu nahi genuke:

1. Hiztegiaren kodeketarako formatu estandar batez baliatuz, EH TEIko gidalerroen arabera kodetzea.

2. Azaleko sintaxiaren alorra urratzea: murriztapen-gramatika landu eta EHko adibideen gainean aplikatu dugu.

3. Aditzen argumentu-egitura lantzen laguntzeko metodologia jorratzea, emaitza gisa hiztegiko aditzen azaleko patroiak erdietsiz.

# BIBLIOGRAFIA


Abaitua J.. *Complex predicates in Basque: from lexical forms to functional structures* .Tesis doctoral, Univ. de Manchester. 1988.

Abaitua J., Aduriz I., Alegria I., Arregi X., Artola X., Arriola J.M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M.. *Estudio comparativo de diferentes formalismos sintácticos para su aplicación al euskera.* FISS EHU-UPV LSI- TR193. 1993.

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

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

Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M.. *MORFEUS: Euskararako analizatzaile morfosintaktikoa..* Eranskina UPV/EHU/LSI/TR 1-99. 1999.

Aduriz I., Agirre E., Alegría I., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Maritxalar M., Sarasola K. eta Urkia M.. A Morphological Analysis Based Method for Spelling Correction in *Proceedings of the E.A.C.L.*, Utrecht, The Netherlands, 1991.

Aduriz I., Aldezabal I., Arriola J.M., Artola X., Urkia M.. Euskararen normalizazioa eta Linguistika Konputazionala .*EUSKERA* XXXIX. 1994.

Aduriz I., Alegria I., Arriola J.M., Artola X., Gojenola K., Maritxalar M.. Different Issues In The Design of a Lemmatizer/Tagger For Basque . From texts to tags: Issues in multilingual language analysis. *Proceedings of the ACL SIGDAT Workshop. Seventh Conference of the Chapter of the Association for Computational Linguistics.*, University College Dublin. Belfield, Dublin, Ireland. 1995.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. eta Urkia M.. Euskararako murriztapen-gramatika lehen urratsak. Eranskina UPV/EHU/LSI/TR 2-96. 1996.

Aduriz I., Arriola J.M., Artola X., Díaz de Illarraza A., Gojenola K., Maritxalar M.. Morphosyntactic disambiguation for Basque based on the constraint grammar formalism. *In Proceedings of the Recent Advances in Natural Language Processing (RANLP)*, Tzigov Chark, Bulgaria, 282-287. 1997.

Ageno A., Castellón I., Ribas F., Rigau G., Rodríguez H., Smiotou A.. TGE: Tlink Generation Enviroment. *In Proceedings of the 16th International Conference on Computational Linguistics (Coling'94)*. Kyoto, Japan. 1994.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M.. XUXEN: A spelling checker/corrector for Basque based on Two-Level Morphology. *Proceedings of the Third Conference ANLP (ACL)* Trento, Italy., ( 119-125). 1992.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Urkia M.. Aplicación de la morfología de dos niveles al euskara. *Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN* 8. zb., 87-102 1989.

Agirre E., Ansa O., Arregi X., Arriola J.M., Díaz de Ilarraza, A., Lersundi M., Soroa A., Urizar R.. Extracción de relaciones semánticas mediante gramáticas de restricciones, Congreso SEPLN'98, Alicante, España, 1998.

Agirre E., Arregi X., Arriola J. M., Artola X., Díaz de Ilarraza A., Insausti J. M., Sarasola K.. Different issues in the design of a general-purpose Lexical Database for Basque, *in Proceedings of the First Workshop on Applications of Natural Language to Databases. (NLDB'95)* . Versailles, 299-313. 1995.

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

Agirre E., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Insausti J.M., Sarasola K.. *Different issues in the design of a general-purpose Lexical Database for Basque* First Workshop on Applications of Natural Language to Databases.(NLDB'95). Versailles. 1995.

Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. Aduriz I., Urkia M.. EUSLEM: Un lematizador/etiquetador de textos en euskara. *Actas del X.Congreso de la SEPLN, Córdoba..* 1994.

Aldezabal I., Gojenola K., Oronoz M.. Combining Chart-Parsing and Finite State Parsing. *in Proceedings of the ESSLLI Student Session*, Utrecht. 1999.

Aldezabal I., Gojenola K., Sarasola K., Goenaga P.. Subcategorización verbal vasca: propuesta inicial y herramienta de validación. SEPLN nº 23. 1998.

Aldezabal I., Gojenola K., Sarasola K.. A Bootstrapping Approach to Parser Development. *Proceedings of the International Workshop on Parsing Technologies (IWPT2000).* Trento. 2000.

Alegria I., Artola X., Sarasola K., Urkia M.. Automatic morphological analysis of Basque. *Literary & Linguistic Computing,* Vol. 11, Nº 4, 193-203. Oxford University Press. Oxford. 1996.

Alegria I., Artola X., Sarasola K.. Hizkuntzaren tratamendu automatikoa, JAKIN 102 zk., 61-82.1997.

Alonge A.. Analysing Dictionary Definitions of motion verbs. *Proc. of Colling-92*, Nantes, 23-28. 1992.

Alshawi H.. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *In: Computational Linguistics*, vol. 13. 1987.

Alshawi H.A.. Analysing dictionary definitions. *In B.Boguraev,T. Briscoe eds,., Computational Lexicography for Natural Language Processing*, New York: Longman, 153-169. 1989.

Allerton D. J.. *Valency and the English Verb*. New York: Academic Press. 1982.

Amsler R. A., and White J.S.. Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-readable Dictionaries. Technical Report MCS77-01315, NSF. 1979.

Amsler R. and Tompa F.. An SGML-based standard for English monolingual dictionaries. *Proceedings of the 4th Annual Conference of the UW Center for the New OED*, Waterloo, 61-79. 1988

Amsler R.A., White J.S.. Development of a Computational Methodology for Deriving Natural Language Semantic Structures via Analysis of Machine-Readable Dictionaries. LRC-79-7:1 Report. 6National Science Foundation. 1979.

Amsler R.A.. A Taxonomy for english Nouns and Verbs. *In Proceedings of the 19th annual Meeting of the Association for Computational Linguistics*, (ACL´81), Stanford, California, 133-138. 1981

Amsler R.A.. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph. D. Diss. Computer Science, TR-164. University of Texas, Austin. 1980.

Amsler R.A.. Third Generation Computational Lexicology. *Proceedings of The First International Lexical Acquisition Workshop*, Detroit, Michigan. 1989.

Arriola J.M., Artola X., Gojenola K., Soroa A.. TEI: testu-kodeketarako gidalerroak. *Ekaia, Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria* 7, 121-139. 1997.

Arriola J.M., Artola X., Maritxalar A., Soroa A.. A Methodology for the Analysis of Verb Usage Examples in a Context of Lexical Knowledge Acquisition from Dictionary Entries. *Proc. of LINC´99*, Bergen, Norvegia, 1-7. 1999.

Arriola J.M., Artola X., Soroa A.. Análisis automático del diccionario Hauta-Lanerako Euskal Hiztegia . *XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural,* Bilbao. 1995.

Arriola J.M., Artola X., Soroa A.. Hauta-Lanerako Euskal Hiztegiaren analisi erdiautomatikoa. *Anuario del Seminario de Filología Vasca "Julio Urquijo"*, XXX-2, 621-629. 1996.

Arriola J.M., Soroa A.. Lexical Information Extraction for Basque. *Proceedings of CLIM´96*, Montreal, Kanada. 1996.

Artola X.. *HIZTSUA: Hiztegi-sistema urgazle adimenduaren sorkuntza eta eraikuntza.Hiztegi-ezagutzaren eskuratze eta errepresentazioa, dedukzio-mekanismoen ezarrera eta oinarrizko funtzionalitateen zehaztapena.* Ph. D. EHU. 1993.

Asker L., Gambäck B., Samuelsson Ch.. EBL: An Approach to Automatic Lexical Acquisition. *Proc. of COLING-92*, Nantes, 1172-1176. 1992 .

Atwell, E.. Constituent-likelihood grammar. In Garside, Leech & Samson (eds.), 57-65. 1987.

Barker K. and Szpkowicz S.. Interactive Semantic Analysis of Clause-Level Relationships. *Proc. PACLING 1995, Pacific Association for Computational Linguistics,* Brisbane, Australia, 22-30. 1995.

Barker K.. *Noun Modifier Relationship Analysis in the TANKA System.* Techincal Report TR-97-02, Dept. of Computer Science, Univ. of Ottawa, 1997.

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

Basilii R., Pazienza M.T. and Velardi P.. What Can Be Learned from Raw Texts? An Integrated Tool for the Acquisition of Case Roles, Taxonomic Relations and Disambiguation Criteria . *Machine Translation*, vol. 8, 147-173. 1993a.

Black A., van de Plassche J., Williams B.. Analysis of Unkown Words through Morphological Decomposition. *Proc. of 5th Conference of the EACL*, vol. 1, 101-106. 1991.

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

Bläsi C., Koch H.. Dictionary Entry Parsing Using Standard Methods. *Proc. of COMPLEX'92*, 61-70. 1992.

Bod, R.. Using a annotated corpus as a stochastic grammar. *Proc. of EACL'93.* ACL, Utrecht. 1993.

Boguraev B. & Levin B.. Machine Readable Dictionaries: a computational linguistics perspective . *Tutorial at the ACL Second Conference on Applied Natural Language Processing*, Austin, Texas. 1988.

Boguraev B. & Pustejovsky J.. Lexical Ambiguity and the Role of Knowledge Representation in Lexicon Design. *In Proceedings of the 13th International COLING*, Helsinki. 1990.

Boguraev B. et al. (1991), Database models for Computational lexicography. Research Report RC 17120, IBM Research Center, Yorktown Heights, NY. Byrd, R. J. et al. . Tools and Methods for Computational Lexicography, Computational Linguistics 13, ns. 3-4, 1987.

Boguraev B. eta Levin B.. Machine-Readable Dictionaries: A Computational Linguistics Perspective. *Presented as a Tutorial at the Second Conference on Applied Natural Language Processing*, Austin, Texas. 1988.

Boguraev B., Briscoe T., Carroll J. and Copestake A.. Database Models for Computational Lexicography. *Fourth International Congress on Lexicography ( Euralex Vox )*, Málaga, Spain. 1990.

Boguraev B., Briscoe T.. Large lexicons for natural language processing:utilising the grammar coding system of LDOCE . *Computational Linguistics*, Vol.13, Numbers 3-4. 203-218. 1987.

Boguraev B.. Building a Lexicon: The Contribution of Computers. *International Journal of Lexicography,* Vol. 4, Nº 3, 227-260. 1991.

Boguraev B.. Machine-readable dictionaries and Research in Computational Linguistics, Workshop "Automating the Lexicon" (Grosseto). 1986.

Boguraev B.K. and Briscoe E.J.. Computational Lexicography for Natural Language Processing, in Boguraev, B. and Briscoe, E. eds. Longman, London. 1989.

Boguraev B.K. and Briscoe E.J.. Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of the Longman Dictionary of Contemporary English. *Computational Linguistics* 13.4: 219-240. 1987.

Boguraev B.K., Briscoe E.J., Carroll J., Carter D. and Grover C.. The derivation of a grammatically-indexed lexicon from The Longman Dictionary of Contemporary English. *In Proceedings of the 25th Association for Computational Linguistics,* Stanford, Ca.. 1987.

Brent M. R. and Berwick R.C.. Automatic acquisition of subcategorization frames from Tagged Text. *In Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann. 1991a.

Brent M.R.. Automatic acquisition of subcategorization frames from untagged text. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, 1991 C.A. 193-200. 1991b.

Brent M.R.. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. Association for Computational Linguistics, vol. 19, nº 2. 242-263. 1993.

Briscoe E. eta Carroll J.. Automatic extraction of subcategorization from corpora. *Proceedings of the 5th ACL Conference on Applied Natural Language Processing.* Washington, DC. 1997.

Briscoe T. and Carroll J.. Towards automatic extraction of argument structure from corpora. Working Paper WP #46, ESPRIT BRA-7315. Acquilex-II, 1994.

Briscoe T., Copestake A., Boguraev B.. Enjoy the Paper. Lexical Semantics via Lexicology. In Proc. COLING. 1990.

Bruce R. and Guthrie L.. Genus Disambiguation : A Study in Weighted Preference. *Proc. of Colling-92*, Vol. IV, 1187-1191, Nantes, Aug. 23-28, 1992.

Byrd R. J.. Computational Lexicology for Building On-line Dictionaries: the Wordsmith Experience, in L. Fignoni and C. Peters eds., Vol. I, 117-137, *Computational Lexicology and Lexicography* (Special Issue dedicated to B. Quemada), Giardini:Pisa. 1990.

Byrd R.J., Calzolari N., Chodorov M. S., Klavans J.L., Neff M.S., Rizk O.A.. Tools and Methods for Computational Lexicography. *Computational Linguistics*, vol.13, ns.3-4, 219-240. 1987.

Calzolari N. and Picchi E.. A project for Bilingual Lexical Database System. *Proceedings of the Second annual Conf. of the Centre for the New OED*, University of Waterloo, Waterloo, Ontario, 79-82. 1986.

Calzolari N.. Detecting patterns in a lexical data base. *Proc. COLING'84* (Stanford Univ.),170-173. 1984b.

Calzolari N., Picchi E.. Acquisition of semantic information from an on-line dictionary. *Proc. COLING'84* (Budapest), 87-92. 1988.

Calzolari N.. An Overview of Written Language Resources in Europe: a few Reflectionws, Facts, and a Vision. *Proceedings of the First International Conference on Language Resources & Evaluation*, vol. I, 217-224. 1998.

Calzolari N.. Machine-readble dictionaries, lexical data bases and the lexical system. *Proc. COLING'84* (Standford Univ.), 460. 1984a.

Calzolari N.. Structure and access in a automated lexicon and related issues, *Workshop* "Automating the Lexicon" (Grosseto). 1986.

Calzolari N.. Structure and access in an automated lexicon and related issues, in L.Fignoni, C.Peters eds.,vol. I,139-161, *Computational Lexicology and Lexicography* (Special Issue dedicated to B. Quemada). Pisa: Giardini. 1990a.

Calzolari N.. The development of large mono- and bilingual lexical data bases, *contribution to the IBM Europe Institute Computer based Translation of Natural Language* (Garmisch-Partenkirchen). 1989.

Calzolari N.. Trends in Computational Lexicography and Natural Language Processing, lecture read at the X Reunión Anual SEPLN (Donostia). 1990b.

Carbonell J., Hayes P. J.. Recovery strategies for parsing extragrammatical language *American Journal of Computational Linguistics.* 1983.

Carpenter B. eta Penn G.. The Attribute Logic Engine, User's Guide, Version 2.0.1, CMU Technical Report. 1994.

Carroll J. and Grover C.. The derivation of a large computational lexicon for English from LDOCE, in Boguraev, B. and Briscoe, E. eds. *Computational Lexicography for Natural Language Processing*, Longman, London, 117-134. 1989.

Castellón I.. *Lexicografía Computacional: Adquisición Automática de Conocimiento Léxico*, Ph. D., Universitat de Barcelona, Barcelona. 1993.

Copestake A.. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary, Proceedings of the First International Workshop on Inheritance in Natural Language Processing, Tilburg, The Netherlands, 19-29. 1990.

Copestake A.. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. *Proc. of 1st Intl. Workshop on Inheritance in NLP* (Tilburg, Netherlands). 1990.

Cruse D.A.. *Lexical Semantics*. Cambridge: Crambidge University Press, 1986.

Chan C.H., Chen C.D.. HMM-based Part-of-Speech Tagging for Chinese Corpora. *Proceedings of the Second Conference on Applied Natural Language Processing, ACL.* 136-143. 1993.

Chanod J. P., Montemagni S., Segond F.. Multiple-pass parsing and dynamic relaxation: a text-driven approach to parsing *Proceedings of the International Conference on Expert Systems and Natural Language Processing*, Avignon, France. 1993.

Chanod J.P., Harrichausen B., Montemagni S.. A Two-State Algorithm to Parse Multi-Lingual Argument Structures. *Proceedings of International Conference on Currents Issues in Computational Linguistics*, 230-244, Penang, Malaysia. 1991.

Chodorov M. S., Byrd R.J., and Heidorn G. E.. Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the 23rd Annual Meeting of the ACL*, Chicago, IL, USA, 299-304. 1985.

Chomsky N.. *Remarks on Nominalization*, in Jacobs R. and Rosenbaum P. (eds.), Reading in English Transformational Grammar, Ginn, Watham, Mass, 1970.

Church K. W.. A Stochastic Parst Program and Noun Phrase Parser for Unrestricted Text. *In Proceedings and Conference on Applied Natural Language Processing* ( 136-143). 1988.

Delisle S. & Szpakowicz S.. Extraction of Predicate-Argument Structures from Texts. *In Proceeding of RANLP'97*, Tzigov Chark, Bulgaria, 318-323. 1997.

Delisle S., Barker K., Copeck T., Szpakowicz S.. Interactive Semantic Analysis of Technical Texts: Case Pattern Acquisition, *Computational Intelligence*, 12 (2), 273-306. 1996.

Delisle S.. *Text Processing without A-Priori Domain Knowldedge: Semi-Automatic Linguistic Analysis for Incremental Knowledge Acquisition*, Ph.D. thesis, TR-94-02, Dept. of Computer Science, Univ. of Otawa, 1994.

Dolan W., Vanderwende L. and Richardson S.. Automatically deriving structured knowledge bases from on-line dictionaries.*Proceedings of the first Conference of the Pacific Association for Computational Linguistics* (Pacling'93), Simon Fraser University, Vancouver, Canada. 1993.

Dolan W.. Word Sense Clustering: Clustering Related Senses. Proceedingsof the 15th International Conference on Computational Linguistics (COLING'94), 712-716. 1994.

Eckle J. and Heid U.. Extracting raw material for a German subcategorization lexicon from newspaper text. *In Proceedings of COMPLEX'96*, 39-51. 1996.

Evans R.. ProGram a development tool for GPSG grammars, in Linguistics 23: 213-243, 1985.

Fiedman J.. A computer system for transformational grammar, in Communications of the ACM 12: 341-348, 1969.

Fink S.R.. *Aspects of a Pedagogical Grammar Based on Case Grammar and Valence Theory.* Tübingen: Niemeyer, 1977.

Firzlaff B. and Haenelt K.. On the acquisition of Conceptual Definitions via textual modeling of meaning paraphrases. *Proc. of Coling-92*, Nantes, 23-28. 1992.

Fitzpatrick E. and Sager N.. The Lexical Subclasses of the LSP English Grammar, *Appendix 3 In N. Sager (ed.), Natural Language Processing, Addison Wesley, Reading*, Ma., 1981.

Fitzpatrick E. and Sager N.. The Lexical Subclasses of the LSP English Grammar. TR-9 Linguistic String Project, New York University, 1974.

Flickinger D., Pollard C., Wason T.. Structure-Sharing in Lexical Representation. *Proceedings of the 23rd. Annual Meeting of the Association for Computational Linguistics*, 262-267. 1985.

Flickinger D.. *Lexical Rules in the Hierarchical Lexicon.* Doctoral Disseertation. Standford University, Standford, California, 1987.

Fontenelle T.. Automatic extraction of lexical-semantic relations from dictionary definitions. *Proceedings of IV International Congress EURALEX'90,* Benalmádena (Málaga), 89-103. Barcelona: Bibliograf. 1990.

Fontenelle T.. Co-ocurrence knowledge, Support Verbs and Machine Readable Dictionaries. *Proc. of COMPLEX'92*, 137-145. 1992.

Friedman J.. Computational and theoretical studies in Montague grammar at the University of Michigan, in SISTM Quarterly 1: 62-66. 1978.

Gahl S.. Automatic extraction of subcorpora based on subcategorization frames from part-of-speech tagged corpus. *Proceedings of the COLING-ACL'98*, VOL. I, Montreal, Quebec, Canada. 1998.

Garside R., Leech G., Sampson G.. *The Computational Analysis of English: A corpus-based approach*. Longman. London, 1987.

Gellesrtam M.. Lexical Resources and Their Application. *Proceedings of the First European Seminar on Trans-European Language Resources Infrastructure (TELRI), Language Resoruces for Language Technology*, Tyhany, Hungary, 58-64. 1995.

GENELEX. Project Eureka GENELEX. Rapport sur la couche syntaxique, Technical Report Version 4.0, ASSTRIL, GSI-ERLI, IBM France an Sema Group, Paris, 1993.

Gomez F.. Learning Word Syntactic Subcategorizations Interactively. (This paper will appear as Chapter 18 in AI and Automation) (Bourbaki ed.), 413-430, October, 1995.

Granger R.H.. FOUL-UP: A program that figures out meanings of words from context. *The 5th International Joint Conference on Artificial Intelligence*, Cambridge, Massachusetts, 172-178. 1977.

Grimshsaw, A.. *Argument Structure*. MIT Press, Cambridge, Ma., 1990.

Grishman R. & Sterling J.. Acquisition of selectional patterns. *Proceedings of the Coling-92*, Nantes, France, 658-664. 1992.

Grishman R., MacLeod C., Meyers A.. Comlex Syntax: Building a Computational Lexicon, (New York: New York University), 1994.

Gross M.. *Methodes en Syntaxe*. Hermann, París, 1975.

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

Heid U., Heyn M., Christ O.. Extracting Linguistic information from machine-readable versions of traditional dictionaries- a metalexicographic method and some tools. *Proceedings of COMPLEX'92*. 1992.

Heid U.. Relating Lexicon and Corpus: Computational Support for Corpus-Based Lexicon Building in DELIS. *Proceedings of the EURALEX'94*, Amsterdam. 1994.

Heid U.. Towards Reusable Lexical Resources for Natural Language Processing. Some Proposals for Linguistic Knowledge Representation. *Proceedings of The Eleventh International Conference Expert Systems & Their Applications, Specialized Conference Natural Language & Its Applications*, Avignon, France, 89-101. 1991.

Ide N. and Véronis J.. Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time? *Knowlegde Building and Knowledge Sharing*, K.Fuchi and T. Yokoi (Eds.) Ohsmsha, Ltd. and IOS Press, 1994.

Ide N., Le Maître J., Véronis J.. Outline of a Model for Lexical Databases. *Information Processing and Management,* vol. 29, N° 2, 159-186. 1993.

Ide N., Véronis J.. *Text Encoding Initiative. Background and Context,* Ide N., Véronis J. eds., Kluwer Academic: Dordrecht. 1995.

Ingria R.. Lexical Information for Parsing Systems: points of Convergence and Divergence, Workshop "Automating the Lexicon" (Grosseto).1986.

Jackendoff R.. *Semantic Structures*, MIT Press, 1990.

Jacobs P. and U. Zernik.. Acquiring Lexical Knowledge form Text: A Case Study. *The 7th National Conference on Artificial Intelligence*, Saint Paul, Minnesota, 739-744. 1988.

Järvinen T. and Tapanainen P.. *A Dependency Parser for English.* Technical Report, n° TR-1, Department of General Linguistics. University of Helsinki. March 1997.

Jensen K., Heidorn G. and Richardson S.. *Natural language processing: The PLNLP approach*, Kluwer Academic Publishers, Ma., 1993.

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

Jensen K.. Why Computational Grammarians can be skeptical about existing linguistic theories. Proceedings of COLING´88, 48-449.1988.

Kaplan R. eta Bresnan J.. Lexical-Functional Grammar: A Formal System for Grammatical Representation. *In Bresnan, J. (ed.), The Mental Representation of Grammatical Relations.* Cambridge, Mass.: MIT Press, 173-281. 1982.

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

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

Karlsson F., Voutilainen A., Heikkilä J., Anttila A.. Constraint Grammar: a language-independent system for parsing unrestricted text français. 1992.

Karlsson F.. Constraint Grammar as a Framework for Parsing Running Text, in H. Karlgren (ed.), Papers presented to the 13th International Conference on Computational Linguistics, Vol. 3. Helsinki. 168-173. 1990.

Kay M.. The Dictionary of the future and the Future of the Dictionary, in A. Zampolli, A. Capelli eds. 1611-174, The possibilities and Limits of the Computer in Producing and Publishing Dictionaries. Pisa: Gardini, 1983.

Klavans J., Chodorov M., Wacholder N.. From Dictionary to Knowldedge Base Via Taxonomy. Proc. of the 6th Conference UW Center for the New OED, Waterloo, 110-132. 1990.

Knight K. and Luk S.. Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of the American Association for Artificial Inteligence.* 1994.

Kokkinakis D.. Towards Automatic Corpus-Based Argument Identification. *Proceedings of International Workshop on Predicative Forms in Natural Language and in Lexical Knowledge Bases*, Toulouse, France, 45-53. 1996.

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

Kucera H. F., Nelson W.. *Computational Analysis of Present-Day American English.* Providence, RHode Island: Brown University Press, 1967.

Kuhn J., Eckle-Kohler J. eta Rohrer C.. Lexicon Acquisition with — a Bootstrapping Approach. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, vol. I, 89-95. 1998.

Landau S.I.. *Dictionaries. The Art and Craft of Lexicography.* New York: Scriber, 1984.

Langer S., Maier P., Oesterle J.. CISLEX- An Electronic Dictionary for German: Its Structure and a Lexicographic Application. *Proceedings of COMPLEX'96*, 155-163. 1996.

Lenders W.. What's in a lexical entry? The contribution of computers to lexicography, in L. Fignoni & C. Peters eds. *Computational Lexicology and Lexicography* (Special Issue dedicated to B. Quemada), Vol. II, 45-63. Giardini: Pisa, 1990.

Levin B.. *English Verb Classes and Alternations.* The University of Chicago Press, 1993.

Li H. and Abe N.. Generalazing case frames using a thesaurus and the MDL principle. *Proceedings of International Conference on Recent Advances in Natural Language Processing*, 239-248. 1995.

Litkowsky K.C.. Requirements of text processing lexicons. *Proc. 18th Annual Conference of the A.C.L.* (Philadelphia, Pennsylvania), 153-154. 1980.

Longman Dictionary of Comtemporary English (LLDOCE). Harlow and London: Longrman Group Limited. 1981.

Manning C.. Automatic acquisition of a large subcategorisation dictionary from corpora. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 235-242. 1993.

Markowitz J., Ahlswede T., Evens M.. Semantically significant patterns in dictionary definitions. *Proc. 24th. Annual Meeting ACL* (New York), 112-119. 1986.

Martí M. A. and Castellón I.. Gramática para el análisis del diccionario VOX. *Boletín SEPLN*, 10, Donostia, 123-143. 1990.

Martin W., Demeersseman H., Vielgen M.. SNIC-Project, beschrijving, opzet en verantwoording, Amsterdam, 1992.

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

Mel'cuk I., Iordankaia L., Arbachewsky Y., Jumarie N.. Un nouveau type de dictionnaire: le Dictionnaire Explicatif et Combinatoire du francais contemporain, *Cahiers de lexicologie* v 38:1 3-34, París, 1981.

Meyers A., Macleod C. eta Grishman R.. Standardization of the complement adjunct distinction. *Proceedings of EUALEX'96.* 141-150. 1996

Miller G.A.. Dictionaries of the Mind. *Proc. 23rd. Annual Meeting of the ACL* (Chicago), 305-314. 1985.

Minaeva L.. Dictionary examples: friends of foes?. *Proceedings of EURALEX'92*, 77-80. 1992.

Monedero J., González J.C., Goñi J.M., Iglesias C.A., Nieto A. F.. Obtención automática de marcos de subcategorización verbal a partir de texto etiquetado.: EL SISTEMA SOAMAS. *Procesamiento del Lenguaje Natural*, boletín nº 17, septiembre de 1995.

Montemagni S. and Vanderwende L.. Structural Patterns vrs. String Patterns for Extracting Semantic Information from Dictionaries. *Proceedings of COLING'92*, 1992.

Nakamura J., Nagao M.. Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation. *Proc. COLING'88*, 459-464. 1988.

Neff M. S., Byrd R. J., Rizk O. A.. Creating and querying lexical data bases. *Proceedings of the 2nd Conference on Applied Natural Language Processing ACL*, Austin, Texas. 1988.

Neff M.S., Boguraev B.K.. Dictionaries, dictionary grammars and dictionary entry parsing. *Proc. 27th annual Meeting ACL* (Vancouver, British Columbia), 91-101. 1989

Oflazer K., Hakkani-Tür D.Z., Tür G.. Design for a Turkish Treebank. *Proceedings of a Workshop Sponsored by the ACL, LINC'99, EACL'99*, Bergen, Norway. 1999.

Oostdijk N.. *Corpus Linguistics and the Automatic Analysis of English*. Language and Computers: Studies in Practical Linguistics Nº 6, edited by Jan Arts and Willem Meijs, Amsterdam, Atlanta, GA 1991.

Pentheroudakis J., Vanderwende L.. Automatically Identifying Morphological Relations in Machine-Readable Dicgionaries. *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research*, 114-131. 1993.

Philips J. & Thompson H.. GPSGP -- a parser for generalized phrase structure grammars, *in Linguistics* 23: 245-261, 1985.

Pin-Ngern S. C., Evens M., Ahlswede T. and Strutz R.. Developing a large lexical Database for information retrieval, parsing, and text generation systems. *Inormation Processing and Management* Vol. 29, No. 4, 415-431. 1993.

Pin-Ngern S.. *A lexical database for English to support information retrieval, parsing, and text generation.* PHD Thesis, Illinois Institute of Technology. Chicago, 1990.

Pozanski V. & Sanfilippo A.. Detecting dependencies between semantic verb subclasses and subcategorization frames in text corpora. *Proceedings of the SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text*, Boguraev, B. & Pustejovsky, J. eds., 1993.

Procter P.. Longman dictionary of contemporary English, Longman, London, 1987.

Pugeault P., Saint-Dizier M.G.. Monteil. Knowledge Extractin from Texts: A Method for Extracting Predicate-argument Structures from Texts. *Proc COLING-94* (Kyoto, Japan), 1039-1043. 1994.

Pustejovsky J.. The generative Lexicon. *Association for Computational Linguistics* vol. 17, 4, 409-441. 1991.

Pustejovsky J.. The semantic representation of lexical knowledge, *in U. Zernik ed. Proc. First Int. Lexical Acquisition Workshop* (Detroit). 1989.

Pustejovsky J.. Type coercion and lexical selection, *in J. Pustejovsky eds. Semantics and the Lexicon*, Kluwer, Dordrecht, 1993.

Quirk R., Greenbaum S., Geoffrey L. and Svartvik J.. *A Comprehensive Grammar of the English Language*. Longman, Harcourt, 1985.

Raoul N., Smith.Edward Maxwel.. An english Dictionary for computerized syntactic and semantic processing.systems. *Proccedings of the Int. Conf. on Comp. Linguistics*- PISA 1973 A. Zampolli & N. Calzolari eds. Vol.36. 1980.

Resnik P.. *Selection and Information: a class-based approach to lexical relationships*. University of Pensilvanya, CIS Dept., Ph. D.. 1993.

Ribas F.. A experiment on learning appropriate selectional restrictions from a parsed corpus. *Proceedings of the Coling-94*, Kyoto, Japan, 769-774. 1994.

Richardson S. D.. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*, Tesia, The City University of New York. 1997.

Rigau G.. *Automatic Acqusition of Lexical Knowledge from MRDs*. Tesia, Bartzelona, 1998.

Rodríguez C.. CORRECTOR: un sistema de verificación sintáctica y estilística de textos. Proceedings of*VII congreso de la SEPLN*. 1991.

Rohen Wolff, S. Macleod C. & Meyers A.. COMLEX word classes, Mannual, Linguistic Data Consortium, UPenn, 1994.

Saint-Dizier P. and Viegas E.. *Computational Lexical Semantics*, (Cambridge: Cambridge University Press). 1995

Sammuelson C. & Voutilainen A.. Comparing a Linguistic and a Stocastic Tagger. *Proceedings of ACL-EACL'97*, Madrid. 1997.

Sarasola I.. *Euskal hiztegia.* Kutxa Fundazioa: Donostia, 1997.

Sarasola I.. *Hauta-lanerako euskal hiztegia.* Gipuzkoa Donostia Kutxa: Donostia, 1984-1995.

Shieber S.M.. *An Introduction to unification-based Approaches to grammar*. CSLI, Stanford University, 1986.

Simpson J.. The New OED Project. *Procs. First Conf. of the UW Centre for the New Oxford English Dictionary* (Waterloo, Canada), 1-6. 1985.

Sinclair J.M.. *Looking up. An Account of the COBUILD Project in Lexical Computing*. London: Collins, 1987.

Sinclair J.M.. The automatic analysis of corpora. *Trends in Linguistics*. Studies and Monographs Nº 65, 379-400. Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm. 1991.

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

Sparck J. K.. How much has information technology contributed to linguistics?. 1996.

Sperberg-McQueen C. M. and Burnard L.. *Guidelines for the encoding and interchange fo machine-readable texts*, ACH, ACL, and ALLC: Draft Version 0. 1990.

Sperberg-McQueen C.M., Burnard L.. *Guidelines for Electronic Text Encoding and Interchange.* Sperberg-McQueen C.M., Burnard L. arg. Chicago & Oxford. 1994.

Stawhowitz J.. Beyond feasibility study: lexicographic progres. *Proceedings of the Int. Conf. on Linguistics,* Pisa, 1973, A. Zampolli, N. Calzolari eds. Vol. 36. 1980.

Swanepoel P.. Problems, Theories and Methodologies in Current Lexicographic Semantic Research, *in: Willy Martin & Willem Meijs & Margreet Moerland & Elserniek ten Pas & Piet van Sterkenburg & Piek Vossen (eds), Euralex'94 Proceedings.* Amsterdam, 11-26. 1994.

Tapanainen P. and Järvinen T.. Syntactic Analysis of Natural Language Using Linguistic Rules and Corpu-based Patterns. *Proceeding of COLING-94*, Kyoto, 629-635. 1994.

Tapanainen P.. *The Constraint Grammar Parser CG-2.* University of Helsinki. Publications n. 27, 1996.

Ten Hacken P., Maas H. & Maegaard B.. Dictionaries in Eurotra, *in C. Copeland, J. Durand, S. Krauwer & B. Maegaard (eds.)*, The Eurotra linguistic specifications, The Commission of the European Communities, Brussels, 1991.

Tesniére L.. *Éléments de syntaxe structurale.* Klincksieck, 1959.

Urkia M.. *Euskal Morfologiaren Tratamendu Automatikorantz*, Ph. D. tesia, EHU, 1997.

Ushioda A., Evans D., Gibson T. and Waibel A.. The automatic acquisition of frecuencies of verb subcategorization frames from tagged corpora, *in B. Boguraev and J. Pustejovsky eds. The Acqusition fo Lexical Knowledge from Text. SIGLEX ACL Workshop*, Columbus, Ohio: 95-106. 1993.

Utsuro T. & Matsumoto Y.. Learning Probabilistic Subcategorization Preference by Identifying Case Dependencies and Optimal Noun Class Generalization Level. *Proceeding of the 5th ANLP*, 364-371. 1997.

Van den Hurk I. and Meijs, W.. The Dictionary as Corpus: Analyzing LDOCE'S Definition-Language. Corpus Linguistics II.

Van den Hurk I.,Meijs W.. *The dictionary as a corpus: analyzing LDOCE's definition-language,* Corpus Linguistics II, 99-125.

Van der Eijk P., Bloksma L., Van der Kraan M.. Towards Developing Reusable NLP Dictionaries. *Proc. of Colling-92*, Nantes. 1992.

Vanocchi M., Rosini R., Carenini M., Prodanof I. & Calzolari N.. *Italian verbs: Developing a neutral formalism for verbal representation*, Technical Report ILC-NLP-1994-1, ILC-CNR, Pisa, 1994.

Vater H.. Toward a generative dependency grammar. *Lingua* 36: 121-145, 1975.

Vossen P.. Polysemy and Vagueness of Meaning Descriptions in the Longman Dictionary of Contemporary English, *in J. Svatvik and H. Wekker (eds.), Topics in English Linguistics.* Mouton de Gruyter. 1991.

Vouitilainen A. eta Tapanainen P.. Ambiguity resolution in a reductionistic parser. *Proceedings of EACL'93*. Utrecht, Holland, 394-403. 1993.

Voutilainen A., Heikkilä J., Antttila A.. *Constraint Grammar for English. A Performance-Oriented Introduction*. 1992.

Voutilainen A.. NPtool, a detector of English noun phrases. *In Proceedings of the Workshop on Very Large Corpora*, Ohio State University, Ohio, USA, 48-57. 1993.

Voutilainen A.. *Three studies of grammar-based surface parsing of unrestricted English text*. Ph.D. thesis. University of Helsinki. Publications n° 24. 1994a.

Voutilainen A.. *Designing a Parsing Grammar*. University of Helsinki. Publications n° 22. 1994b.

Walker D., Zampolli A., Calzolari N.. Automating the Lexicon: Research and Practice in a Multilingual Enviroment. Walker D., Zampolli A., Calzolari N. (eds.): Cambridge University Press, Cambridge.1988.

Walker D.. Developing lexical resources. *V Annual Conference of the UW Centre for the NOED*, Oxford. 1989.

Walker D.. M.R. Dictionaries. *Proceedings of Coling'84*, (Panel Session) Standford University. 1984.

Weiner E.S.C.. Editing the OED in the Electronic Age. *Procs. Fifth Annual Conf. of the UW Centre for the New Oxford English Dictionary* (Oxford), 23-31. 1989.

Wilks Y., Fass D., Cheng-Ming G., McDonald J., Plate T., Slator B.. A Tractable Machine Dictionary as a Resource for Computational Semantics, in B. Boguraev, T. Briscoe eds., chap. 9, 193-228, *Computational Lexicography for Natural Language Processing*. New York: Logman, 1989.

Wilks Y., Fass D., Guo Ch., McDonal J., Plate T. and Slator B.. Providing Machine Tractable Dictionary Tools, *in Pustejovsky J. ed. Semantics and the Lexicon*, Dordrecht, Kluwer Academic Publishers, 341-401, 1993.

Zabala I.. *Predikazioaren Teoriak Gramatika Sortzailean (Euskararen Kasua)*, Ph. D. Tesia, EHU, 1993.

Zajac R.. The Habanera Lexical Knowledge Base Managment System. *Proceedings of the First International Conference on Language Resources & Evaluation*, vol. I, 263-268. 1998.

Zampolli A.. Perspectives for an Italian Multifunctional Lexical Database, *in A. Zampolli, ed. Studies in honor of Roberto Busa S.I.*, 301-341. Giardini: Pisa, 1987.

Zernik U.. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Zernik U. Ed. Lawrence Erlbaum Associates, publishers. Hillsade, New Jersey. 1991.