V Ezagumendu sintaktikoaren erabilera erroreen detekzioan eta zuzenketan

V.1 Sarrera

Lengoaia naturalaren prozesamendurako sistemek hasieratik tratatu behar izan duten arazo bat erroreena izan da, testu idatzietan zein ahotsaren tratamenduan mota askotako erroreak aurki baitaitezke. Testu idatzietan, nahiz eta konturik handiena jarri, beti gertatzen dira mota desberdinetakoak diren eta kausa desberdinak dituzten erroreak. Honen ondorioa testuak orrazten denbora handia pasatzea da eta, gainera, azkenean testua argitaratzen den momentua iritsibezain pronto erroreen bilketaren fasea hasi behar izatea. Hau egia bada euskara neurri batean menperatzen dugunontzat, zer esanik ez euskara menperatzen ez duten edo ikasten ari direnentzat. Euskararen kasuan pentsa daiteke erroreen agerpena beste hizkuntza batzuetan baino maiztasun altuagoarekin gertatuko dela, bere egoera bereziagatik (hau da, hizkuntzaren normalizazio-prozesua, hiztun askoren ezagutza partziala edo inguruko erdaren eraginaren ondorioz). Honengatik, hain zuzen ere, euskararen morfologiaren tratamendu automatikoaren lehen aplikazioa zuzentzaile ortografikoa izan da (Aduriz *et al*. 1997, Aldezabal *et al*. 1999a). Erroreen tratamenduan informazio morfologiko hutsetik informazio sintaktikora pasatzeak asko zabaltzen du aplikaziorako eremua eta, horregatik, kapitulu honetan testu idatzietako erroreen tratamenduan ezagumendu sintaktikoaren ekarpena aztertzeko lehen saioak aurkeztuko ditugu. Asmo nagusia garatu diren tresnak erabiliz egin daitezkeen tratamenduen bideragarritasuna frogatzea izan da. Dena dela, badakigu, sintaxiaren tratamenduan aipatu dugun bezala, eremu hau neurri handi batean landugabea dela, eta etorkizunerako ikerlerroa izaten jarraituko duela zenbait urtetan.

Erroreen tratamenduan bi aspektu nagusi bereizten dira: *detekzioa* eta *zuzenketa* (Kukich 1992). Alde batetik erroreen detekzioak hitz batean edo esaldi batean akats bat gertatu den ala ez emango du. Zailagoa izaten da askotan zuzenketaren arazoa, errore bat dagoela jakinda ere proposamenen zerrenda batetik zuzena dena aukeratzea lan konplexua baita. Detekzioan nahiz zuzenketan, bereiztu behar dira hitz isolatuen prozesua eta testuingurua aztertzen duena. Hitz isolatuen kasuan, arrakastatsu suertatu dira hainbat zuzentzaile ortografiko komertzial. Beraien tratamendua hitzen zerrenda batean edo lexikoian oinarritzen denez mugak dauzkate, kasu askotan beste motetako ezagumenduak, sintaxia edo semantika, behar izaten direlako detekzioa edo zuzenketa egiteko. Testuingurua kontuan hartuz gero, beste errore mota batzuetara zabal daiteke esparrua. Adibidez, hitz batean egindako erroreak beste hitz zuzena eman dezake (hauei *real-word errors* edo benetako hitzen erroreak deituko diegu), eta ezinezkoa izango da hitz isolatuen tratamendu hutsarekin detektatu. Berdin gertatuko da beste errore batzuetan (komunztadurakoak adibidez), hitzez hitz hartuta zuzenak direlako. Hauei errore sintaktikoak edo semantikoak deitzen zaie.

Kapitulua hasteko, erroreak sailkatzeko irizpideak deskribatuko dira (§ V.1.1), ondoren erroreen tratamenduaren munduan egin diren zenbait sistemaren azterketarekin jarraitzeko (§ V.1.2). § V.2n erroreen tratamenduaren inguruan egin ditugun lanak azalduko ditugu. Bertan erroreen detekzio eta zuzenketarako proposamen ezberdinak aurkeztuko dira, beren ebaluazioarekin batera. Egindako lanaren gehiena euskarari aplikatu zaion arren, ingeleserako errore ortografikoen zuzenketa automatikorako esperimentua ere aurkeztuko da. Kapitulua bukatuko da erroreen detekzio eta zuzenketari buruzko lanen ondorioekin eta geratzen zaizkigun eginkizunen deskribapenarekin.

V.1.1 Errore motak

Erroreen sailkapen bat egiteko momentuan lehen arazoa hori egiteko irizpidea aukeratzean datza, horren araberako sailkapen ezberdinak agertuko direlako (Douglas eta Dale 1991, Gómez Guinovart 1996, 1999). Hona hemen erabili diren irizpide batzuk:

?? Irizpide deskriptiboa. Honek erroreen forma aztertzen du, errore ortografikoen tratamenduan egiten den antzera (Alegria 1995). Horrela aldaketa, sorrera, desagerpena eta trukearen ondorioz sortutako erroreak bereizten dira. Hitz-mailan eragiketa horiek karaktereen gainean egiten dira, eta beste mailetan (sintaxia, semantika), ezaugarri sintaktiko edo semantikoen gainean. Inplementazioaren ikuspuntutik, sailkapen hau egokiagoa da errore ortografikoen tratamenduan, hitz batetik karaktereak aldatuz sortzen diren proposamenak mugatuak direlako, eta esaldi bateko osagaien ezaugarri sintaktiko eta semantikoen aldaketek
proposamen-kopuru ikaragarria sor ditzaketelako.

?? Irizpide linguistikoa. Erroreen detekzio edo zuzenketarako behar den informazioaren arabera. Horrela, errore ortografikoak, sintaktikoak, semantikoak eta pragmatikoak bereiz daitezke. Mota honetako beste sailkapen sinpleagoa Kukich-ek (1992) egindakoa da: hitz isolatuen erroreak (ortografikoak) eta testuingurua behar dutenak (testuinguru sintaktiko, semantiko edo pragmatikoa). Kasu honetan, zuzenketa egin behar duenaren ikuspuntutik sailkatzen dira erroreak, jatorri erabat ezberdinetakoak izan daitezkeen erroreak multzo berean sartuz, baldin eta zuzenketan behar den ezagumendua antzekoa bada. Adibidez, askotan tekleatze-erroreek hitz zuzenak ematen dituzte (Peterson 1986), eta hauek errore sintaktiko edo semantikoetan sailkatuko dira, ezagumendu mota horiek beharko baitira tratamendu egokia egiteko, jatorriz errore ortografikoak diren arren.

?? Irizpide etiologikoa. Aurreko irizpidearen beste aldea kontuan hartuz gero, beste sailkapen mota bat erroreak jatorriaren arabera multzokatzearena da. Hemen enfasia ez da zuzenketarako tresnan jartzen, baizik eta errorea egiten duen pertsonaren gainean. Adibidez, pertsona batek *astakeritik* hitza idazten badu, arrazoia
tekleatze-errorea edo *a* organikoduna dela ez jakitearen ondorioa izan daiteke. Honek psikolinguistikarekin eta hizkuntzen irakaskuntzarekin ditu loturak (Díaz de Ilarraza *et al*. 1997, Maritxalar 1999), eta konplexutasuna gehitzen dio erroreen tratamenduari, detekzio-moduaz gain jatorriari buruzko informazioa erabili behar baita. Abantaila erroreen tratamenduaren egokitasunean dago, erabiltzailearen aldetik.

Gure kasuan, erroreetan dugun interesa egindako tresna informatiko eta linguistikoen erabileran datza eta, beraz, bigarren motako erroreen sailkapenean, irizpide linguistikoan, oinarrituko gara. Gainera, tresnek morfologia eta sintaxia lantzen dutenez, ezagumendu mota horiek aplikatuz lor daitezkeen emaitzak azalduko ditugu, semantika[32] edo pragmatika bigarren maila batean utziz. Azter ditzagun, beraz, errore mota ezberdinak:

?? Errore lexikal edo ortografikoak. Lehen esan den bezala, multzo honetan lexikoi bat eta hitzen formaziorako erregelen bidez detekta daitezkeen erroreak sartuko ditugu, eta hitz zuzenak sortzen dituzten erroreak sintaktiko edo semantikoen multzoan kokatuko ditugu.

[32] Erroreen tratamendurako semantikaren aplikazioa (Agirre, 1999) tesian azaltzen da.

Errore lexikalen detekzioa morfologiarekin lotuta dago; bere tratamendua 80ko hamarkadan egiten hasi zen (Peterson 1980, 1986, Damerau eta Mays 1989), eta gaur egun emaitza onak lortzen dituzten zenbait sistema garatu dira (Ispell[33]; Office 97rako tresnak[34]; Aldezabal *et al*. 1999a). Arazo nagusietako bat hiztegian ez dauden baina zilegi diren hitzekin gertatzen da, hiztegiaren hutsunea edo neologismoa dela eta. Tesi honetan sintaxiaren erabilera aztertu nahi dugunez, ez dugu errore hauen detekzioari buruzko gehiago esango.

Detekzioaren arazoa neurri handi batean ebatzita egon arren, zuzenketak beste problema batzuk dauzka. Hori argitzeko, ikus dezagun *probem* hitza sartuta zein diren ingeleserako zuzentzaile ortografiko baten proposamenak:

problem probe proem probed probes

Proposamen horien artean bereizteko beharrezkoa izango da testuinguruari buruzko informazioa, sintaxia (kategoria sintaktiko ezberdinetako proposamenak daude) edo semantika erabili beharko delako.

?? Errore sintaktikoak. Esaldi bateko hitz guztiak lexikoaren bidez ondo analizatu eta gero, oraindik erroreak gerta daitezke esaldiak ez baditu betetzen gramatika baten arauak (gramatika bat zilegitzat eman daitezkeen egitura sintaktikoen deskribapen formala kontsideratuko dugu, gramatika nola adierazten den, eredu estatistikoak edo erregelak, zehaztu gabe). Definizio honekin ere errore ortografikoekin gertatzen zen arazo bera agertzen zaigu, zaila delako definitzea edozein testu ulertuko duen gramatika, eta horren ondorioz esaldi bat ezin denean ulertu, bereiztu egin beharko da errorea den edo gramatikatik at dagoen fenomenoa den.

Errore sintaktikoen artean komunztadura-erroreak (deklinabide-atzizkia eta izen-multzoa, subjektua eta aditza), aditzen azpikategorizazio-eredu okerren erabilera edo menpeko esaldiekin lotutako erroreak aipa ditzakegu. Nahiz eta oraindik emaitzak errore ortografikoentzat lortutakoak bezain onak ez izan, merkatuan errore sintaktikoen tratamendurako tresna ugari dago (Smith 1992, Rabinovitz 1993).

?? Errore semantikoak. Sintaktikoki zuzenak diren esaldiek errore semantikoak izan ditzakete. Hauen artean bereizteko hautapen-murriztapenak erabili ohi dira lengoaia naturalaren prozesamenduan. Errore semantikoen detekzioaz gain, informazio semantikoa errore ortografiko edo sintaktikoetatik sortutako proposamenak diskriminatzeko ere

33 http://fmg-www.cs.ucla.edu/geoff/ispell.html

34 http://www.proofing.com (*Proofing Tools for Microsoft Office 97*).

erabil daiteke. Gure lanean sintaxiaren erabilera aztertzen ari garenez, informazio semantiko sinpleak bakarrik erabiliko ditugu eta horregatik ez diogu arreta handirik eskainiko errore mota honi.

?? Diskurtsoaren erroreak edo errore pragmatikoak. Errore semantikoekin bereizketa egiteko esaten da pragmatikak hitzen esanahia testuinguruan hartzen duela. Dena dela, berbaldia eta pragmatikaren tratamendu konputazionala oraindik hasieran dagoenez ez dugu hauei buruz gehiago esango.

?? Puntuazio-erroreak. Puntuazioak (Nunberg 1990) beste osagai linguistikoek bezala, bere arauak ditu, arau horien artean sintaxia, semantika eta berbaldia markatzeko erregelak daudela.

Errore mota bakoitzaren maiztasunen estimazio gutxi egin dira, batez ere hitz zuzenak sortzen dituzten erroreak ezin direlako automatikoki kalkulatu. Mitton-ek (1987), ikasleen 40.000 hitzeko testuak aztertu ondoren, erroreen %40a hitz zuzenak direla ondorioztatzen du. Honek sintaxia eta semantikaren erabilera eskatzen du, are gehiago hitz okerren proposamenen diskriminazioa ere kontuan hartzen bada, guztira erroreen erdiak baino gehiago tratatzeko sintaxia, semantika edo maila altuagoko ezagumendua erabiltzea beharrezkoa dela ondorioztatzen delako.

Hitz zuzena ematen duten erroreak neurtzeko saio bat egin zen Atwell eta Elliot-en (1987) lanean:

?? Erroreen portzentaje handi bat (%52tik %4ra, testu motaren arabera) hitz okerrak dira

?? Besteetatik gehienak (%48tik %28rako tartean) errore sintaktiko lokalak dira, hau da, hitz gutxi batzuen testuingurua aztertuz gero detektatu edo zuzendu daitezkeenak. Beste multzo bat errore sintaktiko globalena da (%16-%8 tartean), esaldi osoari buruzko ezagumendua beharko lituzketenak. Bukatzeko, errore semantikok aipatzen dituzte, %36-%10 tartean.

V.1.2 Erroreen detekziorako zenbait sistemaren azterketa

Ezagumendu sintaktikoan oinarritutako erroreen tratamendua nagusiki 80ko hamarkadan hasi zen, gehienbat garai berean sintaxiaren formalizazioaren alorrean egindako aurrerapenei esker. Nahiz eta ordutik hona sistema asko diseinatu eta inplementatu diren (Douglas 1991, Kukich 1992, Stede 1992), puntu honetan sistema horietatik garrantzitsuenak edo eredu berriak ireki dituztenak aipatuko ditugu:

a) Analisi sintaktikoan oinarritutako sistemak.

Esperimentatu diren zenbait sistemak (Schank *et al*. 1980, Fass eta Wilks 1983, Carbonell eta Hayes 1983) ezagutza semantikoa erabiltzen du erroreei ekiteko, baina horrek eremu jakin batzuetara mugatzen ditu hurbilpen hauek, edozein testu tratatzeko behar den ezagumendu semantiko aberats hori eskuragarri ez dagoelako, eta horregatik metodoen orokortzea ezinezkoa da une honetan.

Ezagumendu gutxiago behar duten beste hurbilpen batzuek gramatika sintaktikoa dute oinarritzat (Heidorn *et al*. 1982, Weischedel eta Sondheimer 1983, Rodríguez 1991, Genthial *et al*. 1991, Jensen *et al*. 1993, Vosse 1992, 1994, Douglas eta Dale 1992, Menezo *et al*. 1996, Ramírez eta Sánchez 1996), *erlaxazioaren* teknika erabiliz: esaldi baten analisirik lortzen ez denean sistema huts egin duen erregelaren bilaketan ahalegintzen da, erregela horren murriztapen baten erlaxazioak analisi bat ematen duen egiaztatzeko. Honela komunztadura-erroreak, deklinazio okerrak edo izena-determinatzailearen arteko desadostasunak detektatu ahal dira. EPISTLE/CRITIQUE (Jensen *et al*. 1993) mota honetako sistema da. 300 erregela sintaktikorekin eta lexikoi zabal batekin (100.000 hitz) 100 errore gramatikal eta estilokoak detekta litezke. Tresna hau 2.254 esaldiko corpus batean probatu zen, eta emaitza nagusia erroreen zuzenketa (edo zuzenketarako laguntza, zuzenketa asmatzen ez denean) %82 eta %41 artekoa da, testu motaren arabera. Erlaxazioaren arazo bat testu errealei aplikatu ondoren ikusten da, oraindik ez delako garatu edozein esaldi (egunkari batean agertzen direnak kasu) analizatzeko gai den sistema, gramatiken mugak alde batetik eta testu errealen lengoaiaren aberastasuna eta aldakortasuna bestetik (honi buruz, ikusi analisi sintaktikoari buruzko III. kapitulua). Horrela, esaldi batek analisirik ez badu, oso zaila da jakitea hori errore gramatikal baten edo gramatikaren hutsune baten ondorioa den. Bestalde, teknika honek aurkitzen duen beste oztopo bat erlaxazioak dakarren eraginkortasun-galera da, aukera-kopurua modu esponentzialean biderkatu egin daitekeelako.

Analisi sintaktikoan oinarria duten beste hurbilpen motak ere badaude. Sistema batzuek (Mellish 1989, Min eta Wilson 1998), *chart*-a erabiltzen saiatzen dira esaldi baten analisirik lortzen ez denean, bertan esaldiaren zatien analisiak daudelako, eta beraiekin zenbait konbinazio egin ondoren, batzuetan esaldi osoaren analisia lor daitekeelako.

Beste ideia bat gramatika oso bat erabili gabe erroreen testuinguruak deskribatzen dituzten patroiak erabiltzea da. Era honetan, gramatika oso baten gainean baino, gramatikaren zati txikiak deskribatzen dituzten erregelekin, ziurtasun handikoak, lan egingo da. Ideia honen adibidea konpilazio-errore tipikoak inplementatzeko erabili da (In-Sig *et al*. 1993) eta murriztapen-gramatikan oinarria duten sistemak ere garatu dira (Lingsoft, *http://www.lingsoft.fi/*), nahiz eta bere funtzionamenduari buruzko publikaziorik ezagutzen ez dugun. Erroreen tratamendu mota hau sintaxiaren munduan gertatu den fenomenoaren isla dela uste dugu, gramatika oso batetik gramatikaren zati batzuen deskribapenera pasa delako, era horretan testu errealak modu eraginkorrean tratatzeko gaitasuna lortuz.

Antzeko beste tratamendu bat (Holan *et al*. 1997, Schneider eta McCoy 1998) gramatika bati errore jakin batzuei buruzko erregelak (*mal-rules* edo *error productions*) txertatzea da. Horrela bi ezagumendu mota, zuzena eta okerra, kodetzen dira gramatikan.

Bigarren hizkuntzaren irakaskuntzarako sistemetan lan ugari egin da, gehienetan aipatutako tekniken konbinazioak erabiliz, baina beti erroreen detekzio hutsa baino harantzago joanez, helburu nagusia irakaskuntza delako. Adibidez, Menzel eta Schröder-en (1999) sisteman errore-produkzioak eta erlaxazioa erabiltzen dituzte, azken hau gramatika sintaktiko baten gainean, baina baita semantika edo munduko ezagutzaren gainean aplikatuz. Lortutako sistema oso sendoa da, baina testuinguru jakin batzuetarako bakarrik erabil daiteke, edozein testuren aplikaziotik urrun.

b) Metodo automatikoetan oinarritutako sistemak. Multzo honetan bi kategoria bereiztuko dugu: estatistikan oinarritutako sistemak eta ikasketa automatikoa erabiltzen dutenak. Biek komuna duten ezaugarri nagusia da eskuz edo automatikoki lortutako datu-multzo batetik ezagumendua ateratzen saiatzen direla.

Estatistikan oinarritutako sistemetan lengoaiaren modelizazio estatistikoa erabiltzen da erroreen detekzioa edo zuzenketa lantzeko, hitzen edo kategoria sintaktikoen bigrama/trigramak edo hitzen arteko agerkidetza-neurriak erabiliz. Atwell eta Elliott-en (1987) artikuluan 13.500 hitzeko corpus bateko erroreak detektatzen saiatzen dira, kategorien bigramen bidez, %62 detekzio-tasa eta %35eko doitasuna emaitza nagusia dela. Alarma faltsuen (doitasun baxua) kopuru altuegiak direnez emaitzak pobreak kontsideratu behar ditugu.

(Mays *et al*. 1991) lanak benetako hitzen erroreen (hitz zuzena sortzen dutenak) detekzio eta zuzenketaren problemari heldu nahi dio, 20.000 hitzeko lexikoiaren gaineko hitzen trigramen eredua erabiliz. 100 esalditik 8.628 esaldi oker sortu zituzten ausaz, hitz batetik hurbil dauden aukerak sortuz. Erroreen detekzioa %76an egiten da ondo, %74ko zuzenketa-tasa lortuz. Lan honetako zenbait alde oraindik ikertu behar dira, esaldi multzo itxi batekin egin zelako, ez testu errealekin, eta horrek eragina du kostu handiko datu estatistikoak aldez aurretik kalkulatu direlako.

Gale eta Church-ek (1990) hitz okerren zuzenketarako esperimentua deskribatzen dute, zuzenketarako proposamenak diskriminatzeko hitzen bigramen informazioa erabiliz. Era horretan ezagumendu sintaktiko eta semantiko primitibo baten erabilera egiten da. Emaitza testuingururik gabeko zuzenketarako beste hurbilpen batekin konparatzen da, %87tik %90erainoko igoera lortuz doitasunean. Emaitzek hobekuntza adierazten duten arren, oraindik esperimentatu beharko da bide hau.

Ikasketa automatikoaren arloan ere zenbait lan egin dira, ezagutza modu induktiboan ateratzeko. Golding eta Schabes-ek (1996) kategoria sintaktikoen trigramen eta testuinguruaren ezaugarriak konbinatzeko metodoa proposatzen dute benetako hitzen erroreak detektatu eta zuzentzeko. Bere esperimentuan agertzen den sistemak Microsoft Word-en zuzentzaile gramatikalak baino emaitza dezente hobeak lortzen ditu, baina

estaldura hamazortzi errore-multzotara mugatzen da, bakoitzean okerra egiteko probabilitate altua duten antzeko bi edo hiru hitz daudela (adibidez: *weather*, *whether*). Mangu eta Brill-ek (1997) esperimentu bera egiten dute, ikasketa-metodo desberdina erabiliz. Antzeko emaitzak lortzen dituzte, baina beraien alde ikasitako erregela-kopuru txikia eta irakurgarritasuna aipatzen dituzte.

V.2 Erroreen detekzioan eta zuzenketan egindako esperimentuak

Errore motak eta erroreen tratamendurako sistemen aipamena egin eta gero, puntu honetan guk egindako lanak azalduko ditugu. Hasteko, euskarazko testuetan identifikatu ditugun erroreen sailkapena emango da (§ V.2.1). Ondoren, berorien detekzio eta zuzenketarako aukera ematen duten metodoen inplementazioa eta ebaluazioa azalduko dugu. Hiru esperimentu burutu ditugu. Hasieran, murriztapen sintaktikoen erlaxazioa aplikatuz lortutako emaitzak aipatuko dira (§ V.2.2), erroreen patroien detekzioaren hurbilpenaren lehen ebaluazioarekin jarraitzeko (§ V.2.3). § V.2.4en ingeleserako aplikatu den errore ortografikoen zuzenketarako sistema baten diseinua eta emaitzak aurkeztuko dira.

V.2.1 Euskarazko testuetako erroreen sailkapena

Ondorengo lerroetan euskarazko testuetan aurkitutako erroreen lehen sailkapen bat egingo dugu, hurrengo puntuetako aplikazioetan errore mota batzuk kokatzeko balioko duena. Sailkapen hau egiteko hainbat oinarri hartu ditugu: beste hizkuntzetarako egin diren saioak (Douglas eta Dale 1991, Gómez Guinovart 1996), eta euskara bereziki tratatu dutenak (Walz 1985, HABE 1985, Egunkaria 1992, Zubimendi eta Esnal 1992, Maritxalar 1999). Azken hauen artean badira argitaratu gabeko zenbait barne-dokumentu (euskararen irakaskuntzan diharduten instituzioetatik sortuak gehienbat). Erroreen deskribapen hauez gain, euskara ikasten ari diren ikasleen testu-sorta ere zuzenean aztertu ahal izan dugu (IRALE, AEK edo ILAZKI euskaltegiko ikasleen testuak), erroreen benetako agerpenen iturri aberatsa emanik. Ondorengo lerroetan azalduko diren erroreen adibide gehienak ikasleek egindakoak diren arren, gure ustez errore horiek euskaldunek orokorrean egiten dituzten benetako erroreen adierazgarri dira, euskara-irakasleen iritzia jaso dugulako edo beste hizkuntzetan egindako erroreekin konparatu ditugulako.

Sailkapen honek ezagumendu sintaktikoa erabiliz erroreen detekzio eta zuzenketarako sistema baten definizioaren lehenengo pausoa emateko balioko du. Erroreak sailkatzearekin batera, erroreen kausa zein den aztertuko da, eta beren detekziorako eta zuzenketarako inplementagarriak diren metodoak aztertuko dira:

?? Errore ortografikoak. Hemen hitz solteetan egindako akatsak, hitz okerra ematen dutenak, kontsideratuko ditugu. Errore hauek hainbat arrazoirengatik sor daitezke, V.1 taulako adibideetan ikus daitekeenez.

Adibide horietatik ikusten da erroreen jatorriak mota askotakoak izan daitezkeela, eta jatorri hori jakitea zuzenketarako proposamenak sortzeko edo sailkatzeko orduan erabil daitekeela. Adibidez, errore fonologikoak tratatzeko, hitz okerretatik fonologikoki hurbil dauden proposamenak atera daitezke, eta hurbiltasun-neurri hori zuzenketa bat aukeratzeko erabili (Yannakoudakis eta Fawthrop 1983, Berkel eta DeSmedt 1988). Berdin gertatuko da tekleatze-erroreekin, kasurik gehienetan karaktere baten distantzian egoten direlako hitz zuzenak (Pollock eta Zamora 1984), edo OCR (*Optical Character Recognition*) sistemek egindako erroreekin. Dena dela, informazio horiek lagungarriak

izan arren, zuzenketa guztiz automatikoa egiteko doitasun handiaz sintaxiaren eta semantikaren erabilera ezinbestekoa izango da.

Errorea	Errorearen arrazoia
araso, iztilua, naskagarri, herbestean, majikoen	errore fonologikoak
eskarrizketa	tekleatze-errorea
Itziarreri	deklinabidearen aplikazio okerra
bait da	idazkera-arauen aldaketa
inglesak, errelazio	inguruko erdaren eragina
dirazenean	aldaera dialektala
mexikatarrak, aurkeztzaileari	eratorpen okerrak

V.1 taula. Errore ortografikoen adibideak.

?? Errore sintaktikoak. Hauen jatorria mota askotakoa izan daiteke: errore ortografiko baten ondorioz hitz zuzen bat sortzen denean, komunztadura-erroreak, deklinabide-erroreak, edo menpeko esaldien erabilera okerretik sortutako erroreak. V.2 taulan horien adibideak daude.

Adibide horietatik ikus daiteke sintaxiaren tratamendua behar duten erroreen aniztasuna errore ortografikoena baino askoz handiagoa dela. 1) adibideak erakusten du nola tekleatze-errore batek hitz zuzen bat ematen duen, eta hori detektatzeko ezagumendu sintaktikoa gutxienez edo seguruenik semantikoa ere beharko litzatekeela. Beste errore batzuk (2, 3, 4 eta 5) izen-sintagma baten edo testuinguru mugatu baten azterketarekin detektagarriak izan daitezkeela pentsa daiteke. Bukatzeko, badaude esaldi osoaren azterketa sintaktikoa beharko duten errore motak (6, 7, 8 eta 9).

Zuzenketarako proposamenak lortzeko, errore mota bakoitzak bere metodo propioen definizioa beharko duela uste dugu. Errore ortografikoekin bezala, proposamen horien artean zuzena automatikoki aukeratzeko informazio sintaktiko eta semantikoak sartu beharko dira.

	Errorea	**Errorearen arrazoia**
1)	*flotatze dute*	tekleatze-errorea
2.a)	*gure herrialdetan lan egiteko*	mugagabearen erabilera okerra
2.b)	*itsasorentzat txarra da*	
2.c)	*edozein momentuan jasango du*	
2.d)	*zenbat gonbidatuak*	
2.e)	*hainbeste urtean*	
3.a)	*nahiz eta animali basati izan*	deklinabidearen aplikazio okerra
3.b)	*Lurra osoa*	
4)	*Donostian, 1999ko urtarrilak 1ean*	dataren errorea (zenbakia ezin da deklinatu hila ergatiboan badoa)
5)	*ez ziren elkar bizi etxe kanpo*	komunztadura-errorea eta postposizio okerra
6.a)	*Hori eztiarekin zerikusirik ez dauka*	komunztadura-erroreak
6.b)	*nik ez naiz izan*	
6.c)	*hau ez du esan nahi hori*	
6.d)	*bueltak ematen nuen*	
6.e)	*Patxik kontatu zidan kuadrilako planak*	
6.f)	*biztanle guztiak goitizena dute*	
7.a)	*lan bat neretzat*	esaldiaren osagaien ordena
7.b)	*proposatzen dio plan bat*	
7.c)	*gauza batzuk gertatutakoak kontatuko dizut*	
8)	*prest dago guri zirkora eramateko*	kasuaren asignazio okerra edo aditzaren azpikategorizazioaren ezagutza-eza
9.a)	*lagunei molestatzen zebilen*	azpikategorizazioaren ezagutza-eza
9.b)	*Donostiara joatea behar nuela*	
9.c)	*nik ez dakit nola da Bretaña*	
9.d)	*alarma hautsi lortu zuten*	

V.2 taula. Errore sintaktikoen adibideak.

?? Errore semantikoak eta pragmatikoak. Tesi-lan honetan sintaxiari eskainiko diogunez indar handiena, ez diegu semantikarekin lotura dituzten erroreei garrantzi handia eman. V.3 taulan zenbait adibide ditugu, sintaktikoki zuzenak diren esaldiak baina arazo semantikoak dituztenak (honi buruz ikusi Agirre *et al*. 1994, Agirre 1999). Taulako lehen esaldian, adibidez, komunztadura-errore bat agertzen da, interpretazio sintaktiko zuzena onartzen duena (*ekintza* subjektutzat hartuz gero).

Errorea	Errorearen arrazoia
Edozein ekintzak ez duela onartuko	komunztadura-errorea
haizkorarekin ebaki dugu *irratitik zabaldu dute berria* *horrena hitzegingo dut*	instrumentala (haizkoraz, irratiz, horretaz) erabili behar da
badaukazu ura?	partitiboa erabili behar da
beroa egiten du	azpikategorizazioaren ezagutza-eza

V.3 taula. Errore semantikoen adibideak.

V.2.2 Murriztapen sintaktikoen erlaxazioa

V.2.2.1 Metodoaren azalpen laburra

Puntu honetan errore sintaktiko batzuei murriztapenen erlaxazioaren (*constraint relaxation*) metodoaren aplikazioaren emaitzak azalduko dira, (Gojenola eta Sarasola 1994) lanean eta ondorengo beste esperimentu batzuetan egin direnak. Metodo hori zenbait sistematan erabili izan da (Heidorn *et al*. 82, Douglas eta Dale 92, Tomabechi 93, Ramírez eta Sánchez 96, Mitjushin 1996), eta bere oinarria gramatika bat da. Analizatzaile sintaktikoak, esaldi oso baten analisia lortzen ez duenean, esaldiaren zenbait osagairi buruzko murriztapenak kentzen ditu (erroreen sorburua izan daitezkeenak, adibidez, komunztadura numeroan) eta, analisi bat lortzen bada, orduan errorearen kausa erlaxatutako murriztapenekin erlaziona daiteke. Era honetan, sistema gai izango da gramatikaren estalduratik kanpo dauden esaldiak ulertzeko ere.

Murriztapenen erlaxazioaren metodoaren aurkezpena egin dugunean, esaldi bat analizatzerik ez dagoenean bi arrazoirengatik izan daitekeela esan dugu: errore sintaktikoa dagoela edo esaldia zuzena dela baina gramatikaren estalduratik kanpo dagoela. Garatu dugun gramatika partziala denez, esan dugu testu errealetako esaldi asko ezin izango direla osorik
analizatu (§ III.2). Horregatik, bi aukera horien arteko banaketa sinplifikatzeko, erlaxazioaren probak gramatikaren estalduraren barruan dauden esaldien gainean egingo ditugu, eta horrela esaldi batekin analisia ez lortzea errore baten agerpentzat ulertuko da. Sinplifikazio honen desabantaila handiena testu errealetako esaldiak saihestea da (edo testu errealetako esaldiak erabiltzen badira berauek gramatikaren barruan egon daitezen moldaketak egin beharko dira), baina daukan abantaila nagusia da erlaxazioan oinarritutako metodoaren ebaluaziorako egokia dela, gramatikaren estalduraren arazoaren abstrakzioa egiten delako. Gainera, hau interesgarria izan daiteke beste testuinguru batzuetan, euskara irakasteko sistema batean adibidez, gramatika sinplifikatu baten gaineko ariketak emateko. Era honetan, gramatika ideal bat bageneuka bezala joka dezakegu (kasu hoberena), erlaxazioaren ekarpena eta bideragarritasuna modu isolatuan neurtuz. Pauso hori eman eta gero, testu errealekin lan egiteko behar diren egokitzapenak diseina daitezke hurrengo fase batean. Horregatik, puntu honetan hartuko dugun hasierako hipotesia esaldi guztiak gramatikarekin analiza daitezkeela izango da, esperimentuak horren arabera eginez. Etorkizuneko lanen aipamenean (§ V.3), erlaxazioaren tratamendua testu errealetako esaldiekin nola lot daitekeen azalduko dugu.

Metodo horren euskararako aplikazioa baterakuntzan oinarritutako gramatikaren gainean egingo dugu. Baterakuntza-formalismoak egokiak dira murriztapenen erlaxazioaren aplikaziorako (Douglas eta Dale 1992), beraietan ezagumendu sintaktikoa murriztapenak deskribatzen dituzten ekuazioen bidez adierazten delako, eta ekuazio horietako batzuk kenduz inplementa daitekeelako. Kendutako ekuazio horiek emango dute akatsaren iturria, beraiek gabe esaldiaren analisia lortzea balego.

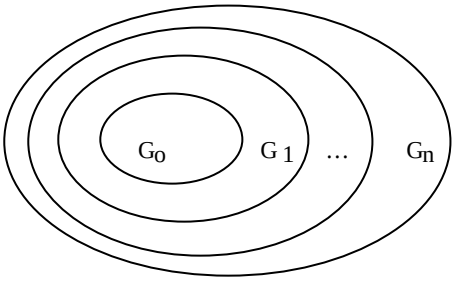
V.1 adibidean erregela bat azaltzen da, izen-sintagma eta aditzaren arteko numeroaren komunztadura eskatzen duena (4garren ekuazioa). Horrela, ez dira onartuko '*txakurra datoz*' bezalako esaldiak, baina laugarren ekuazioa askatuz gero lortuko da esaldi horren analisia. Ekuazioak numeroaren komunztadura egiaztatzen duenez, errorearen arrazoia numeroaren desadostasun batetik datorrela izan liteke zuzenketarako hipotesi bat. Hemen ikusten da erlaxazioaren beste abantaila bat, gramatika bera erabiltzen dela esaldi zuzenak eta okerrak ezagutzeko. Baterakuntzan oinarritutako gramatikak egokienak dira metodo honen aplikaziorako, murriztapenak modu erazagutzailean agertzen direlako, horrela murriztapen jakin bat askatzeko aukera emanez. PATR formalismoa, alde

honetatik, LFG edo HPSG formalismoak baino errazago da metodo honekin erabiltzeko, formalismo konplexuagoetan murriztapen asko printzipio orokorren bidez adierazten direlako (adibidez, komunztaduraren printzipioa) eta zailagoa da esatea erregela jakin batean murriztapen bat (adibidez, bi osagaien arteko komunztadura numeroan) aska daitekeela.

```
X0 ---> X1, X2
      1)  X0/kat              <=> esaldia
      2)  X1/kat              <=> izen-sintagma
      3)  X2/kat              <=> aditz-sintagma
      4)  X1/num              <=> X2/num
```

V.1 adibidea. Numeroaren komunztadura egiaztatzen duen erregela baten adibidea.

Murriztapen sintaktikoen erlaxazioa mota ezberdinetako lengoaietara aplikatu da, gaztelera edo ingelesa kasu (Heidorn 1982, Rodriguez 1993, Tomabechi 1993). Dena dela, hizkuntza eranskari edo malgukarietan (errusiera, suomiera edo euskara (Miller 1986)), metodoaren aplikazioak analisirako aukera-kopuru trataezinak sortzen ditu. Horregatik, erlaxazioa modu gradualean aplikatu beharko da, aukera-kopurua tratagarria dela ziurtatzeko (Douglas eta Dale 1992). Era horretan, gramatika baten estaldura zabaldu egin daiteke, V.1 irudian agertzen den moduan. G_0 hasierako gramatika bada, zenbait pausotan G_1, G_2, eta G_3 gramatika zabalagoak lor daitezke, murriztapen sintaktikoak kenduz, eta horrela esaldi ez gramatikalak ulertzeko gaitasuna lortuko da.

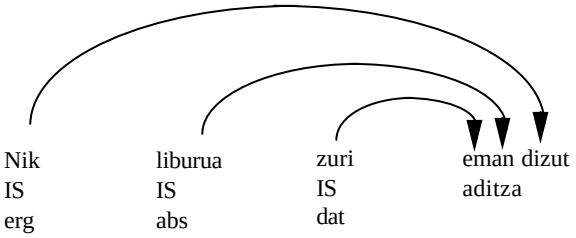


V.1 irudia. Murriztapen sintaktikoen erlaxazioaren bidezko gramatikaren zabalpena.

V.2.2.2 Egindako esperimentuak

Esperimentuetarako, 119 esaldi okerrez osatutako multzoa aukeratu genuen, euskara ikasten ari diren ikasleek idatzitako testuak erabiliz. Testu hauek datu-iturri aberatsa osatzen dute, errore-kopuru altua baitute eta gainera egindako akatsak adierazgarriak dira jatorrizko euskaldunak egiten dituzten erroreekin alderatuz gero, euskara-irakasleen iritziz. Testuen corpusak euskararen morfologiaren irakaskuntzarako IDAZKIDE proiektuan lortu ziren (Díaz *et al*. 1997), eta 100 testu inguru ditu, bakoitza 150-200 hitzekin. Erlaxazioaren proba egiteko, aukeratutako erroreak maiztasun gehienekoak izan ziren, § V.2.1en aurkeztutako errore mota guztietatik:

?? Aditza eta esaldiko osagaien arteko komunztadura kasu, numero eta pertsonan. Komunztadura mota hau ergatibo, absolutibo eta datibo kasuetan doazen izen-sintagmekin gertatzen da (ikus V.2 irudia). Osagaien ordena librea eta elipsia (osagai bat ez agertzea) gehituz gero, ondorioa da esaldi okerren kasuan errore-hipotesi kopuru handia aztertu beharko dela.



V.2 irudia. Esaldiko osagaien arteko komunztadura.

?? Izen-sintagmaren barruko erroreak, hauen artean determinatzaile edota numeroaren gehitzea, kenketa edo ordezkapena daudela.

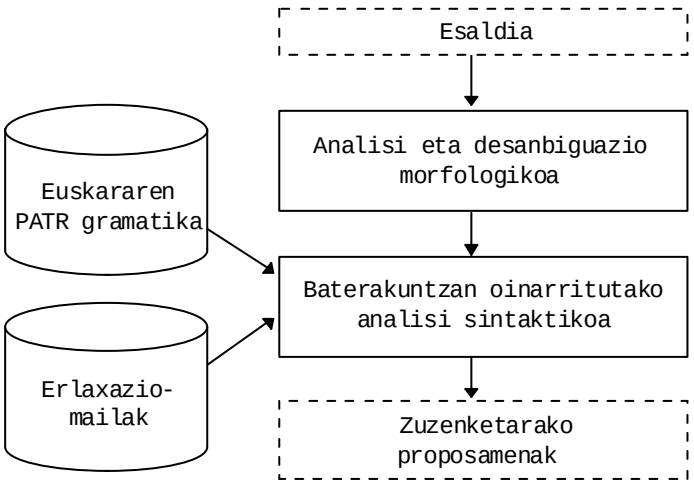
?? Menpeko esaldiekin egindako erroreak, menpekoaren atzizkia edo markaren faltarengatik edo ordezkapenarengatik. Errore hauek komunztadura mota berezi gisa ere ikus daitezke, edo azpikategorizazio-murriztapenen akats moduan.

V.4 taulan errore horien adibideak agertzen dira, dagozkien zuzenketekin batera. Lehen esan denez, esaldiak gramatikaren estalduraren barruan daude, hau da, esaldi erroredunen zuzenketek analisi bat (edo gehiago) izango dute. Horrela egiteko lana gramatika zabaltzea izango da, murriztapen sintaktikoen ezabaketen bidez esaldi oker guztiak analizatu arte.

	Errorea	**Zuzenketa**
Komunztadura-erroreak	*bat-batean gizon bat hamabost zakurrekin agertu ziren .*	*bat-batean gizon bat hamabost zakurrekin agertu zen .*
	notizia horrek gauza onak ekarriko du .	*notizia horrek gauza onak ekarriko ditu .*
	Zenbait ikerlari eman nion .	*Zenbait ikerlariri eman nion .*
	Pezetaren debaluazioa alde onak ekarriko dizkigu .	*Pezetaren debaluazioak alde onak ekarriko dizkigu .*
Izen-sintagmaren barruko erroreak	*Zenbait probintzietan hondartza dago .*	*Zenbait probintziatan hondartza dago.*
	atezaina batekin egon zen .	*atezain batekin egon zen .*
	kotxeren historiak aldatu du .	*kotxearen historiak aldatu du .*
Menpeko esaldien erroreak	*esanda dago inkestak datoz .*	*esanda dago inkestak datozela .*
	gizon horrek esan du nik istilua izan dut .	*gizon horrek esan du nik istilua izan dudala .*
	Nik eman dizut liburua polita da	*Nik eman dizudan liburua polita da*

V.4 taula. Tratatu diren errore sintaktikoen adibideak.

V.3 irudian era honetako erroreen tratamendurako egin dugun sistemaren arkitektura azaltzen da. Bertan gramatika eta erlaxazio-mailei buruzko informazioa modu erazagutzailean kodetuko da. Analisi morfologikoaren ondoren, analizatzaile sintaktikoa esaldia analizatzen saiatuko da. Emaitzarik ez bada ateratzen, orduan lehen mailako murriztapenen askatzea onartuko da. Analisia lortzen bada, askatutako murriztapenek errorearen iturria markatuko dute. Analisirik ez balego, orduan bigarren, hirugarren, ... mailak probatzen jarraituko du, analisi bat aurkitu arte edo maila guztiak aplikatu arte.



V.3 irudia. Erroreak tratatzeko sistemaren arkitektura.

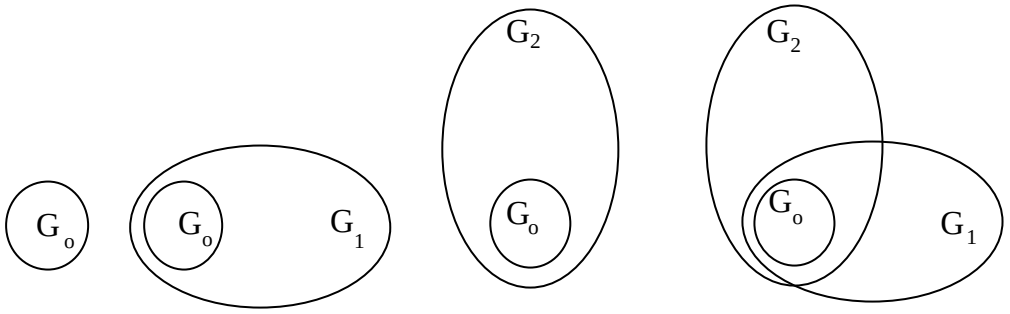
```
% izenlaguna --> is3(mugagabe) + knmdek(gen/gel)
% adibideak: (zenbait gizon) + en
%            (gizon bi) + ren
%            (Peio)  + ren
% gaizki:   *(zenbait gizon) + aren
X0 ---> X1, X2
          1)  X1/kat        <=>    is3
          2)  X2/kat        <=>    knmdek
          3)  X2/kom/kas    badago [gen, gel]
          4)  X0/kat        <=>    izlg
          5)  X0/izlg_mota  <=>    en_ko
          6)  X0/info       <=>    X1
          7)  X1/kom/mug    <=>    X2/kom/mug
          8)  X1/kom/num    <=>    X2/kom/num
          9)  edo[X1/osgk/det/gune/azp    badago [dzh, dzg],
                  X1/gune/azp             badago [izb, lib]]
         10)  X2/kom/mug    badago [mg]
         ...
Beharrezko murriztapenak: [1,2,3,4,5,6, 8,  ...],
  erlaxazio-posiblea([7,10], "izen-sintagmaren atzizkiak mugagabe izan behar du")
  erlaxazio-posiblea ([9], "izen-sintagma hau ez da mugagabea")
  Maila: 1


% as --> is(abs) + as
% adibideak: gizonak  + (etorri dira)
%            gozokiak + (nik eman dizkiot gizonari)
% gaizki:   *gozokia  + (nik eman dizkiot gizonari)
X0 ---> X1, X2
          1) X0/kat       <=> as
          2) X1/kat       <=> isk
          3) X2/kat       <=> as
          4) X2/ordena    <=> ezkerra
          5) X1/kom/kas   <=> X2/azpikat/abs/kom/kas
          6) X1/kom/num   <=> X2/azpikat/abs/kom/num
          7) X1/kom/per   <=> X2/azpikat/abs/kom/per
          8) X1/kom/mug   <=> X2/azpikat/abs/kom/mug
          ...
Beharrezko murriztapenak: [1,2,3,4,8, ...],
  erlaxazio-posiblea([5], "objektua-aditza komunztadura errorea (kasua)")
  erlaxazio-posiblea([6], "objektua-aditza komunztadura errorea (numeroa)")
  erlaxazio-posiblea([7], "objektua-aditza komunztadura errorea (pertsona)")
  Maila: 1
```

V.2 adibidea. Erlaxazioaren aplikazioa bi erregeletan.

V.2 adibidean bi erregela agertzen dira, bakoitzean erlaxatu daitezkeen murriztapenekin. Erregela bakoitzeko, beharrezko murriztapenak (erlaxatzen ez direnak) zehazten dira alde batetik, eta bestetik errorea eman dezakeen murriztapena(k). Horietako bat kenduta analisi bat lortzen bada, dagokion mezuak errorearen kausa emango du. Adibidez, lehenengo erregelaren bi murriztapenek (7 eta 10) kasu-marka mugagabea dela egiaztatzen dute, izen-sintagma mugagabeak osatzeko. Akats arrunta mugatua jartzea izaten denez (‘**zenbait etxe + ak*’), murriztapen horiek kenduz gero izen-sintagma horiei analisi bat emango zaie. Antzeko gauza egiten da bigarren erregelarekin, aditz-sintagma bat absolutiboan doan izen-sintagma bat lotzeko erabiltzen dena. Seigarren ekuazioaren askatzeak, adibidez, numeroaren desadostasuna detektatuko du (‘**gizonak + ikusi dut*’).

Gramatika eta erroreen corpus osoak aztertu eta gero, 75 murriztapen aukeratu ziren, beraien ezabaketaren bidez erroreen analisia lortzen lagungarriak izan zitezkeenak. Lehen esperimentu batean metodoa modurik sinpleenean aplikatu zen, errorea detektatzen lagun zezaketen murriztapen guztiak aldi berean askatuz. Horrela,

analizatzaileak bi pausotan funtzionatuko zuen: lehen maila batean murriztapen guztiak bete behar ziren, esaldi zuzenak analizatuz; bigarren maila bat aplikatzen zen esaldi baten analisia lortzerik ez zegoenean, murriztapen guztiak aska zitezkeela. 63 esaldi okerrekin probatu ondoren, analizatzaileak 58ren analisia lortu zuen, metodoaren erabilgarritasuna frogatuz eta analizatzaileen sendotasuna lortzeko metodoaren baliagarritasuna ere bai. Geroago esango dugunez, ezagututako esaldi bakoitzeko zuzenketarako proposamen bat edo gehiago lortzen ziren. Dena dela, aukera honek eraginkortasun-arazoak zeuzkan, errore bat dagoenean 75 murriztapen guztiak batera askatzeak analisi-aukeren ugaltze ikaragarria eman baitezake, anbiguotasunaren betiko arazoari gehitzen zaiona. Horregatik bost esaldik ez zuten analisirik hartu, denbora eta espazioaren leherketa ekarri baitzuten. Esaldi okerren analisirako batez besteko denbora 29,41 segundo[35] izan ziren, zegozkien esaldi zuzenduen batez besteko 2,5 segundoekin konparatuz.



V.4 irudia. 4 erlaxazio-mailaren erabilera.

Eraginkortasun-aldaketa estimatzeko, beste proba bat egin zen hamar erlaxazio-maila definituz; horietako zortzitan errore-mota baten tratamendua egiten zen, eta beste bi mailetan esaldi zuzenak eta murriztapen guztien erlaxazioa deskribatzen ziren. V.4 irudiak lau erlaxazio-mailaren erabilera azaltzen du. G_0 hasierako gramatika da, eta G_1 eta G_2 gramatikek bi errore moten murriztapenen multzoak osatzen dituzte. Azken aukeran murriztapen guztiak aska daitezke. Zortzi erlaxazio-mailak erroreen maiztasunak gogoan izanda definitu ziren, ideia nagusia maiztasun handiagoko erroreak detektatzeko erlaxazioak lehenago probatzea zela (ikus V.5 taula).

Errorea	**Errorearen adibidea**
Mugagabea jartzea mugatua denean	*erleren laguna han dago*
Komunztadura objektuaren (absolutibo kasuan doana) numeroan	*etxeak joan da* *nik etxeak ikusi dut*
Komunztadura objektuaren kasuan	*nik gizonek ikusi ditut*
Komunztadura subjektuaren kasuan	*ni gizonak ikusi ditut*
Konpletiboetan menpeko atzizkia ez agertzea	*esan du etorri da*
Mugagabea ez jartzea behar denean	*zenbait erlearen laguna ikusi dut*
Komunztadura subjektuaren numeroan	*gizonek egin du*
Komunztadura bigarren objektuaren numeroan	*emaitzei dagokiona*

V.5 taula. Erlaxazio-mailak eta erroreen adibideak, maiztasunez (altuenetik txikienera) ordenatuta.

Esaldien batez besteko analisiaren denbora 27 segundokoa izan zen. Emaitza hauek besteekin alderatuta alde gutxi dutela eman arren, faktore bi eduki behar dira gogoan:

?? Kasu honetan esaldi guztien analisia lortzen da, eta analizatzailea ez da sartzen analisi-aukera esponentzialaren tratamenduan. Aurreko kasuan, murriztapen guztiak batera askatzean, kasu batzuetan ezin zen aurkitu errorearen kausa. Oraingo aukeran banan-

35 Analizatzailearen inplementazioa ordutik hona hobetu denez, denborak oraingo analizatzailearenak baino txarragoak dira. Horregatik, emaitzak aztertzeko orduan denboren arteko erlazioa izan beharko da kontuan.

banan aztertzen dira erroreak, eta kasu guztietan analisi bat ematen da. Beraz, denbora irabazteaz gain doitasunean ere irabazten da.

?? Beste arazo bat analizatzailea optimizatu gabe egotea da, analisi-maila bakoitzean zerotik hasten delako, aurreko mailetan egindako lana berrerabili gabe. Adibidez, V.3 irudian G_0-ren analisi partzialak maila guztietan erabil daitezke. Soluziorik hoberena pausoz-pausoko analizatzailea erabiltzea izango litzateke (Wiren 1993), maila berri baten analisiari ekiteko momentuan erlaxazio berriek alda ditzaketen osagaiak bakarrik erabiliz, lan berria minimizatuz. Gainera, zortzi erlaxazio-mailak ez dira optimizatu, banan-banan probatu direlako, eta eraginkorragoa izan daitekeelako maila batzuk biltzea, beren maiztasunen arabera. Aldaketa horiek inplementatuz gero, espero dugu analisi-denborak ia antzekoak izatea esaldi oker eta zuzenentzat.

Emaitza horietatik zenbait ondorio atera ditugu. Batetik, erlaxazioaren metodoaren bideragarritasuna, esaldi okerrak ulertzeko ezagumendua gramatika bati lotzea ahalbidetzen duelako. Esperimentuetan esaldi guztientzako analisiak lortu ahal izan dira. Bestetik, erlaxazio hori modu eraginkorrean egiteko era gradualean egin behar dela, bestela denbora eta espazioarekin arazoak agertzen direlako, are gehiago erroreen multzoa zabaltzen doan heinean. Edozein kasutan, oraindik esperimentazio asko egin behar da arazo hau gainditzeko.

Laburbilduz, esango dugu aurreko puntuetan frogatu dugula erlaxazioak errore sintaktikoen tratamendua ahalbidetzen duela, esaldi okerrak analizatu ahal direlako murriztapen sintaktikoak erlaxatuz. Horrela, detekzioaren arazoa konpondu da gehienbat. Hori argitzeko, azter ditzagun 53 esaldi erroredunen gainean ateratako emaitzak (ikus V.6 taula). Bertan, esaldi bakoitzeko zuzenketa-proposamenen artean errore-kopuru minimoa daukatenak bakarrik aipatzen dira, kasu gehienetan horien artean egoten baita proposamen zuzena. Hasteko, 11 esaldik analisi bat hartzen dute, hau da, interpretazio sintaktiko zuzen bat daukate (adibidez, '*gizona eman dit*'). Honek adierazten du errore sintaktikoen proportzio handi batek (%20aren inguruan kasu honetan) esaldi zuzenak ematen dituela, eta beraien konponketarako ezagumendu semantikoa, gutxienez, beharko litzatekeela. Bestalde, 14 esaldik zuzenketa-proposamen bakarra dute, baina 28k bi edo gehiago dauzkate. Beraz, proposamenak sailkatzeko irizpideak definitzea garrantzitsua da, Menzel-ek (1990) egin duen bezala, murriztapen gutxien hautsi dituzten interpretazioak hautatuz edo heuristikoen arabera hobetsiz.

Analisi zuzena	Zuzenketa bat	Zuzenketa bi	Zuzenketa bi baino gehiago
11	14	17	11

V.6 taula. Analizatutako esaldi okerren banaketa zuzenketa-proposamenen arabera.

Proposamenak zuzentzeko, ezagutza sintaktiko eta semantikoa beharko da. V.7 taulan bi esaldi eta sistemak proposatutako zuzenketak agertzen dira. Adibidez, lehenengo esaldiaren zuzenketan bazter liteke *harreman sozialak* subjektutzat hartzen duen interpretazioa (jakinez gero *daukate*-ren subjektua normalean biziduna dela). Bigarren esaldian, kontua da aukeratzea zein diren *dauka*-ren subjektua eta objektua, *haitzuloa* edo *garrantzia*. Dena dela, zuzenketaren aspektu honek azterketa sakona beharko du, erabakitzeko zeintzuk diren bereizketa hori egiteko behar diren ezagumendu sintaktiko edo semantikoak.

Errorea	Zuzenketarako proposamenak
harreman sozialak daukate	subjektua-aditza komunztadura errorea (kasua): *harreman sozialek daukate*
	objektua-aditza komunztadura errorea (numeroa): *harreman sozialak dauzkate* edo *harreman soziala daukate*
	objektua-aditza komunztadura errorea (kasua): *harreman soziala daukate* (ergatibo -> absolutibo)
erreportaia honetan Ondarruko haitzuloa sekulako garrantzia dauka.	subjektua-aditza komunztadura errorea (kasua): *erreportaia honetan Ondarruko haitzuloa sekulako garrantziak dauka.*
	subjektua-aditza komunztadura errorea (kasua): *erreportaia honetan Ondarruko haitzuloak sekulako garrantzia dauka.*

V.7 taula. Esaldi erroredunak eta zuzenketarako proposamenak.

V.2.2.3 Ondorioak

Aurreko puntuan (§ V.2.2) errore sintaktikoen detekzio eta zuzenketarako erlaxazioaren metodoa euskarari aplikatu zaio. Euskarak ezaugarri bereziak dauzka erroreen tratamendurako (komunztadura aberatsa adibidez), ingelesa bezalako hizkuntzekin konparatuta. Baterakuntzan oinarritutako gramatikaren gainean inplementatu da sistema, bere erazagutzailetasunari esker murriztapenak adierazteko ekuazioen askapen erraza onartzen duelako. Erlaxazioak errore-multzo zabal baten analisia onartzen duela frogatu dugu, eta eraginkortasuna hobetzeko murriztapenen askapen graduala beharrezkoa dela. Emaitzek metodoaren bideragarritasuna frogatu duten arren, badira oraindik egin beharreko lanak:

?? Errore mota gehiago aztertu behar dira, eta dagozkien erlaxazio-mailak definitu. Oreka lortu behar da maila gehiago definitzeak dakartzan irabazien eta galeren artean. Mailakatzeak analisia posible egiten du, baina maila altuetako erroreak harrapatzeko kostu handiagoa izango du, aurreko maila guztiak probatu behar direlako. N errore posible definitu badira, batez beste N/2 aldiz analizatu beharko da esaldia errore sinple bat harrapatzeko (erroreen maiztasunak kontuan hartzen ez badira behintzat), esaldian bi errore badira (N*(N-1))/2 aldiz errore biak harrapatzeko, ... Adibidez, aditza eta izen-sintagmen arteko komunztadura egiaztatzeko, pertsona, numero eta kasuaren murriztapenak maila berean askatzeak errore-hipotesien kopuru altua aztertu behar izatea dakar, gainera gehienak baztertu egingo dira gero, errore asko dituztelako. Kasu honetan ideia hobea izan daiteke murriztapen horiek maila ezberdinetan askatzea. Maizen gertatzen diren komunztadura-erroreak, ergatiboaren kasu-marka ez jartzea adibidez, lehenengo mailetan jarri beharko dira.

?? Errore bat detektatu eta gero, zuzenketa aukeratzeko lana geratzen da, askotan proposamen bat baino gehiago agertuko delako. Une honetan, proposamen guztiak aurkezten zaizkio erabiltzaileari. Horretarako desanbiguaziorako teknikak garatu beharko dira. Bide interesgarri bat errore ortografikoen zuzenketa automatikorako egin diren lanak dira (Golding eta Schabes 1996, Mangu eta Brill 1997, Agirre *et al*. 1998). Zuzenketa aukeratzeko orduan kontrako justifikaziok eman dira, lan batzuetan errorearen testuinguru

hurbilaren azterketa gomendatzen delako baina beste batzuetan, aldiz, esaldiaren analisi orokorra egitea komenigarria dela aipatzen da. Zuzenketa bat egiteko, gehienetan hitz bat aldatu beharko da sortzaile morfologikoa erabiliz, baina beste kasu batzuetan osagai sintaktiko konplexuak aldatu behar dira, sorkuntza sintaktikoaren bidez. Ikusten denez, puntu honek ikertzeko aukera ugari ematen ditu.

?? Landu ez dugun beste ideia interesgarria sistema ingurune ezberdinetara egokitzea da (adibidez, ikasleen testuen zuzenketan, bigarren hizkuntzen irakaskuntzan, edo egunkari bateko testuak), maila-kopurua eta errore-motak testuaren ezaugarriei moldatuz.

V.2.3 Errore-patroien bidezko detekzioa

V.2.3.1 Sarrera

Aurreko puntuan ikusi dugu gramatika baten gainean murriztapen sintaktikoen askapenaren bidez errore batzuk, komunztadurak adibidez, detekta daitezkeela. Badaude beste errore batzuk, zenbait arrazoirengatik metodo horrekin trataezinak direnak. Egindako baterakuntzan oinarritutako gramatikak euskararen osagai nagusiak definitzen ditu: izen-sintagmak, adizlagunak, menpeko esaldiak eta esaldi sinpleak. Horien bidez lengoaiaren osagairik arruntenak edo maizen gertatzen direnak estaltzen dira. Ondorioz, gramatika hori erabiliz erlaxazioaren bidez detektatu ditugun errore gehienek esaldiaren analisi osoa beharko dute, hau da, lehenago egindako errore sintaktikoen bereizketan (§ V.1.1) aipatutako errore sintaktiko globalak dira. Bestalde, gramatikaren estaldura osoa ez denez, ezin dira gramatikaz kanpoko testuinguruetako erroreak tratatu, erlaxazioa esaldi zuzenen gramatika baten gainean oinarritzen delako. Adibidez, daten deskribapen sintaktikoa egin gabe dugunez, ezin dira daten barruko erroreak metodo horrekin tratatu. Antzeko gauza gertatzen da testuinguru lokalak aztertuz detekta daitekeen errore-multzoarekin, erroreak hitz edo sintagma txikien arteko korrespondentzien artekoak direlako, normalean gramatikan sartu gabe. Honen adibide bat *elkar* hitzarekin egindako erroreak dira (adibidez, '**elkar bizi ziren*'). Fenomeno horiek guztiak gramatika barruan sartzeko, lan handia egin beharko litzateke.

Aspektu hauek kontuan hartuta, puntu honetan errore lokalen patroien bidezko tratamendua (Douglas 1992, Ramírez *et al*. 1997) aztertuko dugu (ikus adibide batzuk V.8 taulan). Hurbilpen hau da merkatuko zuzentzaile gramatikal gehienek erabiltzen dutena. Gainera, gure kasuan ideia hori lortutako analisi partzialen egituraren (*chart*) gainean aplikatu ahal izango dugu, analizatzaileak emandako osagaiak berrerabiliz (Mellish 1989, Min eta Wilson 1998). Hurbilpen honek gramatika oso baten beharraren arazoa saihesten du, erregela lokalen bidez testuinguru batzuk identifikatzen direlako ziurtasun handiz, testu errealak aztertzeko sendotasuna lortuz. Azpitik analizatzaile partziala edukitzeak asko lagunduko du, egitura sintaktiko osoak (izen-sintagmak, menpekoak) erabili ahal izango baitira erroreen testuinguruen deskribapenean.

Errorea	**Errorearen kausa**
Donostia, 1995eko urtarrilak 15ean	hila ergatiboan badoa egunak deklinatu gabe joan behar du
Lagunak elkar bizi dira.	*elkarrekin bizi* idatzi behar da
Joseba mailuarekin baliatu zen	*mailuaz* instrumentala erabiliko da *baliatu* aditzarekin
Etxe guzti horiek egin ditut.	*horiek guztiak* idatzi behar da

V.8 taula. Patroien bidez tratagarriak diren erroreen adibideak.

Erlaxazioaren metodoaren desabantaila bat testu errealak tratatzeko zailtasuna zenez, atal honetan egingo dugun lana testu errealetako errore lokalen aplikaziora zuzenduko dugu, horrela etorkizunerako proposatuko dugun sistema konbinatu baten lehen pausoak emanez.

Oraingo testu-prozesadore gehienek gramatika-zuzentzaile komertzial bat duten arren, lan gutxi argitaratu da emaitzen ebaluazioaz, arrazoi ezberdinengatik:

?? Gramatiken estaldura partziala. Arazo hau esaldi zuzenekin gertatzen bada, zer esanik ez sintaktikoki okerrak diren esaldiekin, multzo ia infinitua osatzen dutelako. Beste alde batetik, sistema sendoek (analizatzaile estatistikoak, murriztapen-gramatika) askotan ezin dituzte bereiztu esaldi zuzenak eta okerrak. Maizen gertatzen diren errore-motak gramatikan sartu arren, oraindik erroreen kopuru osoaren portzentajea txikia izango litzateke. Beraz, esaldi baten analisirik lortzen ez denean, askotan ezin da jakin errore bat edo gramatikaren hutsunea den.

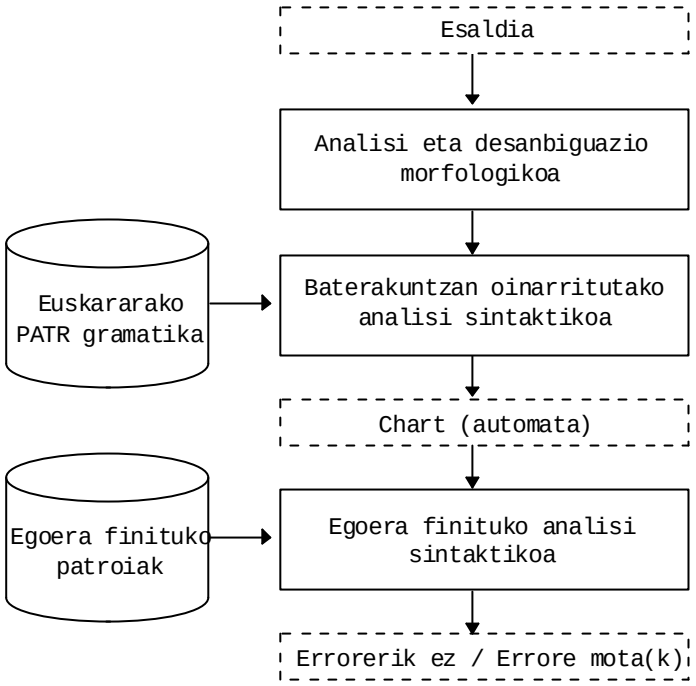
?? Alarma faltsuak. Erroreei buruz argitaratu diren lan gehienetan, errore sintaktikoak dagozkien esaldi zuzenen bidez definitu dira. Testu errealen erabilerak arazo desberdina dakar: esaldi zuzen bat okertzat ematea, errorearen testuinguruarekin antza duelako, nahiz eta askotan esaldi mota horrek erroredunekin inolako erlaziorik ez izan. Gertaera honek problemaren esparrua zabaltzen du: erroreak tratatzeko ez dira bakarrik esaldi okerren adibideak eta dagozkien zuzenketak aztertu behar, baizik eta corpus handi bateko esaldi guztiak.

?? Corpus handien erabilera. Errore sintaktiko bakoitza maiztasun txikiarekin gertatzen da eta, ondorioz, probarako corpus handiak behar dira. Corpus erraldoi horiek eskuragarri egongo balira ere, erroreen agerpenak ezagutzea lan handia da, ez dagoelako sintaktikoki etiketatutako erroreen corpusik ebaluazio edo probarako. Horregatik, erroreen benetako adibideak aurkitzeko, milaka testu aztertu eta markatu beharko dira eskuz eta automatikoki.

Arrazoi hauengatik, puntu honetako helburua corpusetan oinarritutako erroreen detekzioaren bideragarritasuna aztertzea da, zuzenketa-tasa altua lortzeko alarma faltsuen kopuru baxua mantentzen den bitartean. Honelako tresnak beharrezkoak dira euskara bezalako hizkuntzentzat, estandarizazio-arazoak direla eta, beste arazoen artean. Ondoren azalduko den esperimentua (Gojenola eta Oronoz 2000) artikuluan argitaratu dira.

V.2.3.2 Corpusetan oinarritutako patroien bidezko erroreen detekzioa

Erroreen tratamenduari ekiteko, azpikategorizazioari buruzko informazioarekin (IV. kapitulua) egin den modura, lehenengoz baterakuntzan oinarritutako analizatzailea aplikatuko da (hasieran analisi eta desanbiguazio morfologikoa eginda), eta ondoren egoera finituko patroiak definituko dira lortutako *chart*-aren gainean (V.5 irudia). Horrela, egoera finituko erroreen patroietan osagai sintaktiko konplexuak aipatu ahal izango dira, detekzio-ahalmena handituz, osagai lexikalen gaineko patroiak erabiltzen dituzten sistemekin konparatuz gero.



V.5 irudia. Errore-patroien bidezko detekziorako sistemaren arkitektura.

Sistemaren proba egiteko, euskarazko daten adierazpenak aukeratu genituen, bi arrazoi nagusirengatik:

?? Erraza zen probarako esaldiak aurkitzea, beste errore mota batzuekin konparatuz. Nahiz eta lan hau gehienbat eskuz egin behar, daten adierazpenek ezaugarri berezi batzuk dituzte (hilen izenak, urteen zenbakiak), adibideen bilaketa-prozesuan lagungarri direnak.

?? Aplikazio-eremua zabala da, hau da, daten adierazpenek fenomeno aberatsak dituzte morfologian eta sintaxian, errore mota desberdinak egiteko era ematen dituztenak. Errore hauek errore sintaktiko lokalen multzokoen adierazgarritzat ikus ditzakegu. V.9 taulak datak adierazteko formatu erabilienak erakusten ditu (Euskaltzaindia). Ikusten denez, hasieran aukerazko tokia doa, absolutibo edo inesiboan, ondoren urtea genitiboan, eta bukatzeko hila eta eguna doaz. Hila genitiboan badoa, orduan ondorengo eguna deklinatuta joan behar da, eta hila ergatiboan badoa (*martxoak*), egunaren zenbakia deklinatu gabe agertuko da. Horregatik, datetan agertzen diren elementu gehienak deklinatuta doaz, bakoitza bere numero eta kasuaren morfemekin. Gainera, konbinazio batzuk bakarrik dira onargarriak, eta hori erroreen iturburua da, errore gehienak zuzentzaile ortografikoentzat detektaezinak direlako, hitzez hitz hartuta zuzenak baitira.

Durango, 1999ko martxoaren 7a
Durangon, 1999ko martxoaren 7an
Durangon, 1999ko martxoaren 7tik ...
1999ko martxoaren 7an
Durangon, 1999ko martxoak 7

V.9 taula. Datak adierazteko formatu zuzenak.

Ikasleen 267 testu bildu genituen (errore maiztasun handia zuten), egunkaritik eta aldizkarietatik lortuak, guztira 500.000 hitz baino gehiago. Bertatik 658 esaldi hartu genituen (ikus V.10 taula), barruan data zuzenak, okerrak eta daten antzeko egiturak zeuzkaten esaldiak (azken hauek alarma faltsuak detektatzeko). Esaldiak multzo bitan banatu ziren, batean erroreen tratamendurako sistemaren garapenean erabiliko zirenak gordez, eta bigarrenean azken probarakoak utziz (hau da, azken proban aztertu gabeko esaldiak prozesatzeko).

	Garapenerako corpusa		**Probarako corpusa**	
Esaldi-kopurua	411		247	
Data zuzenak	65		39	
Daten *antzeko* esaldiak	255		171	
Data okerrak	91		37	
Data okerrak errore batekin	43	% 47	6	% 16
Data okerrak errore birekin	42	% 46	27	% 73
Data okerrak hiru errorekin	6	% 7	4	% 11

V.10 taula. Probarako esaldiak.

Garapenerako corpusean aurkitutako errore guztiak aztertu eta gero, maiztasun handieneko erroreak aukeratu genituen (ikus V.11 taula). Errore mota bakoitza detektatzeko patroi bat edo gehiago definitu ziren. Dena dela, laster ikusi genuen hau ez zela aukera onena, askotan bi edo hiru errore agertzen direlako adierazpen berean. Fenomeno honek patroien gainean erlaxazio gradualaren antzeko hurbilpena eskatzen du, gogoan izateko errore asko batera ager daitezkeela. Errore bakoitza modu independentean tratatu ordez, patroiak egiteko orduan adierazpen zuzena eta bere bertsio okerrak izan behar ziren gogoan. Bestalde, data zuzena kontsidera daitekeen ereduaren erlaxazioak alarma faltsuen kopurua gehitzeko arriskua dauka. Esan behar da ere *erlaxazio* deitzen dugun horrek aurreko atalean (§ V.2.2) egin dugunaren antza duela, kasu bietan esaldi zuzen baten esparrua zabaldu egiten delako, aldaera okerrak onartuz, baina ezberdintasun nagusia da orain patroi bakoitzeko erlaxazio hori eskuz aplikatu behar dela, eta lehengo ataleko inplementazioan, aldiz, behin askatzen diren ekuazioak zehaztu eta gero, automatikoki egiten zela. Erroreen arteko elkarrekintza modu esponentzialean hasten denez errore-kopuruarekin batera, gure errore-patroietan corpusean aurkitutako konbinazioak bakarrik landu genituen.

	Errorea	**Errorearen kausa**
1	*Donostian, 1995-eko martxoaren 22an*	Urtea zenbakiz idazten denean, ezin da marra jarri
2	*Donostian, 1995eko Martxoaren 22an*	Hila ezin da maiuskulaz jarri
3	*Donostian 1995eko martxoaren 22an*	Tokia idazten bada, ondoren koma jarri behar da
4	*Donostian, 1995eko martxoaren 22*	Hila genitiboan badoa eguna deklinatua egongo da
5	*Donostia, 1995eko urtarrilak 15ean*	Hila absolutiboan badoa eguna deklinatu gabe joan behar da
6	*Donostia, 1995eko martxoan 15ean*	Hila absolutibo edo genitiboan agertu behar da
	karrera bukatu nuenean 1997ko Ekainaren 30	2, 3, eta 4 erroreen konbinazioa

V.11 taula. Tratatu diren erroreak.

Bosgarren errore mota detektatzeko patroiak V.3 adibidean agertzen dira. Hasteko, errorearen testuingurua definitzen da (`ErroreMartxoak22an`), hau da, hila ergatibo edo absolutiboan (azken hau desanbiguazioaren balizko akatsak saihesteko) zenbaki deklinatu baten aurretik. Ondoren, transduktore batek (`MarkatuMartxoak22an`), errorearen hasieran eta bukaeran etiketa bi (`+ERRHAS5` eta `+ERRBUK5`) jarriko ditu. Erregelaren aplikazioa gehiago murrizteko, ezkerreko eta eskuineko testuinguruak zehaztuko dira, bi mailatako morfologiaren antzeko notazioa erabiliz, horrela alarma faltsuen agerpena gutxitzeko, baina aldi berean beste errore mota batzuk onartuz. Adibidean hori osagai ezberdinen bidez egiten da: `Urtea`-k urte zuzena eta okerrak onartuko ditu, `HilaAbsErg`-ek hila maiuskulaz idaztea ere onartuko du, eta `AukerazkoKoma`-k tokiaren atzetik koma ez jartzeko aukera emango du. Horregatik errore bi edo hiru dauden testuinguruetan ere bosgarren errorea detektatuko da (koma faltan eta eguna deklinatua: *Donostian 1999ko martxoak 13an*, edo hila maiuskulaz idazten bada).

```
define HilaAbsErg Hila & [$["+kas" "+erg"] | $["+kas" "+abs"]];

define ErroreMartxoak22an HilaAbsErg HitzMuga ZenbakiIzeDek; # martxoak 22an

define UrteaZuzena   OsagaiSintaktikoa &
                     $["+kat" "+izlg"]  &
                     $["+lema" ZenbakiIze];

define Urtea [UrteaZuzena | ErroreaMarraUrtean];

define MarkatuMartxoak22an [ErroreMartxoak22an] @-> %+ERRHAS5 ... %+ERRBUK5
                              || AukerazkoTokia AukerazkoKoma Urtea HitzMuga
                                 _ ;
```

V.3 adibidea. Bosgarren errorea (*1999ko martxoak 12an*) detektatzeko erregelak.

Errore-patroiak definitzeko orduan, arreta berdina jarri behar da data zuzen eta okerretan. Ondorioz, 60 patroi morfosintaktiko inguru definitu behar izan dira (bakoitza automata edo transduktore batez konpilatuta) sei errore motei dagozkien patroien definiziorako. Hauek murriztapen lokal txikienetatik hasita (45 automatek 100 egoera baino gutxiago daukate) patroi konplexuenetaraino doaz (transduktore bat 10.000 egoera eta 475.000 arkurekin).

V.12 taulak lortutako emaitzak erakusten ditu. Garapenerako corpusa egoera finituko erregelen prestaketan aztertu ahal zenez, bigarren eta hirugarren zutabeko emaitzak analizatzailearen (oraingo egoeran) kasurik hoberenaren adierazgarri kontsidera daitezke, %100eko doitasuna (alarma faltsurik ez) eta %91ren estaldurarekin. Aurretik ikusi gabeko 247 esaldien corpusaren gainean %84ko estaldura lortzen da. Doitasuna begiratzen badugu, 5 alarma faltsu daude, hau da, data zuzenak edo datak ez diren egiturak baina data okertzat hartzen direnak. Alarma faltsuak probarako corpuseko 247 esaldiekin zatitzen badira, alarma faltsuen tasaren estimazioa %2,02koa izango litzateke.

	Garapenerako corpusa		**Probarako corpusa**	
Esaldi-kopurua	411		247	

Detektatu ez diren data okerrak	7	9%	6	16%
Detektatutako data okerrak	84	91%	31	84%
Alarma faltsuak	0		5	

V.12 taula. Ebaluazioaren emaitzak.

V.13 taulak alarma faltsuak aurkezten ditu. Ikusten denez, horietako bi datak ez direnak baina datatzat hartu diren adierazpenak dira; beste bi egindako daten gramatikan kontuan hartu ez diren egiturak dira; eta azkena lexikoiaren hutsune baten ondorioz dator (*Primakov* pertsona-izen bereziaren interpretazioa eman ordez, hitz ezezaguntzat hartzen da). Emaitza onak diren arren, oraindik corpus handiagoak beharko genituzke, doitasuna hobetzeko.

Adibidea	**Alarma faltsuaren kausa**
atxiloketa 1998ko urtarriletik irailaren 16ra ...	Datatzat hartutako data osoa ez den egitura.
Donostian 1960ko Urtarrilaren jaioa	Okerra den esaldia datatzat hartu da.
etorriko da 1997ko irailaren 26ko 1:15etan	Ordua (1:15) hilaren eguntzat ulertu da.
atzotik 1999ko abenduaren 31 arte	Gramatikak ez du *arte* partikula tratatzen, eta horregatik data zuzena okertzat ematen da.
Primakovek 1998ko irailaren 11n hartu zuen ...	*Primakov* hitz ezezaguna lokatibotzat interpretatzen da.

V.13 taula. Sistemak emandako alarma faltsuak.

V.2.3.3 Ondorioak

Puntu honetan analizatzaile sintaktikoa errore-patroien bidezko detekzioan aplikatu dugu. Hauek dira egindako lanaren ezaugarri aipagarrienak:

- Corpusetan oinarritutakoa. Sistema erabilgarria egiteko, esaldi zuzen zein okerren benetako adibideen gainean probatu behar da. Hau begi-bistakoa dela eman arren, ez da orain arte erroreen detekzioan egin den lan gehienen ezaugarria izan. Hau egiteko corpus handiak behar dira, neurri handi batean eskuzko markaketa-lana eginez.

- Errorearen detekziorako metodo erabilienak konbinatu dira (errore-patroiak, *chart*-aren berrerabilera eta erlaxazioaren aldaera bat) emaitza onak lortuz. Horrela, lengoaiari buruzko ezagumendu positiboa (gramatikaren erregeletan adierazitakoa: zer dagoen ondo) eta negatiboa konbinatzen dira (errore-patroien debekuak: zer dagoen txarto (Heaton eta Turton 1987)), lengoaien irakaskuntzan egiten den antzera. Beste alde batetik, erlaxazioaren eskuzko idazketak lan linguistiko handia suposatzen du.

- Baterakuntzan oinarritutako eta egoera finituko analizatzaile sintaktikoen erabilera sekuentziala egokia da erroreen patroien tratamendurako. Baterakuntzan oinarritutako analizatzaileak osagai sintaktiko nagusiak aurkitzen ditu, eta horien gainean patroi konplexuak defini daitezke erroreak harrapatzeko.

Hauek aurreikusten ditugu izango direla sistema hau zabaltzeko hurrengo pausoak:

?? Errore mota gehiagoren azterketa. V.7 taulan, datez aparte, beste errore sintaktiko lokal batzuk agertzen dira, metodo honen bidez tratagarriak izan daitezkeenak. Gure sistemak, erroreen tratamenduan ari diren sistema guztiak bezala, eskalatze-arazoa dauka, errore-kopuruen gehikuntzak errore-patroiak kodetzeko eskuzko lana eskatzen baitu. Errore mota gehiago aztertzen direnean, erregelen arteko elkarrekintza kontu handiz aztertu beharko den arren, espero dugu errore ezberdinetako erregelak independenteak izango direla neurri handi batean. Beste puntu interesgarri bat baliabide linguistikoen berrerabilpena da: daten erroreak tratatzeko orduan, zenbait patroik arau linguistiko orokorrak definitzen dituzte, eta beste batzuk datentzako bereziak dira.

Horrez gain, patroien bidezko tratamendua lagungarria izan daiteke komunztadurak eta antzeko erroreak tratatzeko ere. § V.2.2n erlaxazioan oinarritutako metodoa aztertu dugu, bere arazo handienetako bat testu errealetara aplikatzeko zailtasuna zela. Bi metodoen konbinaziorako hipotesia aztertzen ari gara: patroien bidez azpiesaldi edo sintagmen testuinguruak aukeratuko dira, beren barruan erlaxazioa aplikatuz (Oliva 1997). Horrela, testu errealen tratamendua egingarria izan daiteke.

?? Erroreak detektatzeko erregelen ikasketa automatikoa. Behin erroreak markatuta dauzkan corpusa edukita, gure kasuan eskuz landu ditugu detekzio edo zuzenketarako erregelak. Horregatik, ikasketarako metodoen azterketa oso interesgarri ikusten dugu (Golding eta Roth 1996, Mangu eta Brill 1997), sistema garatzeko lan bakarra erroreen adibideak markatzea izango litzatekeelako.

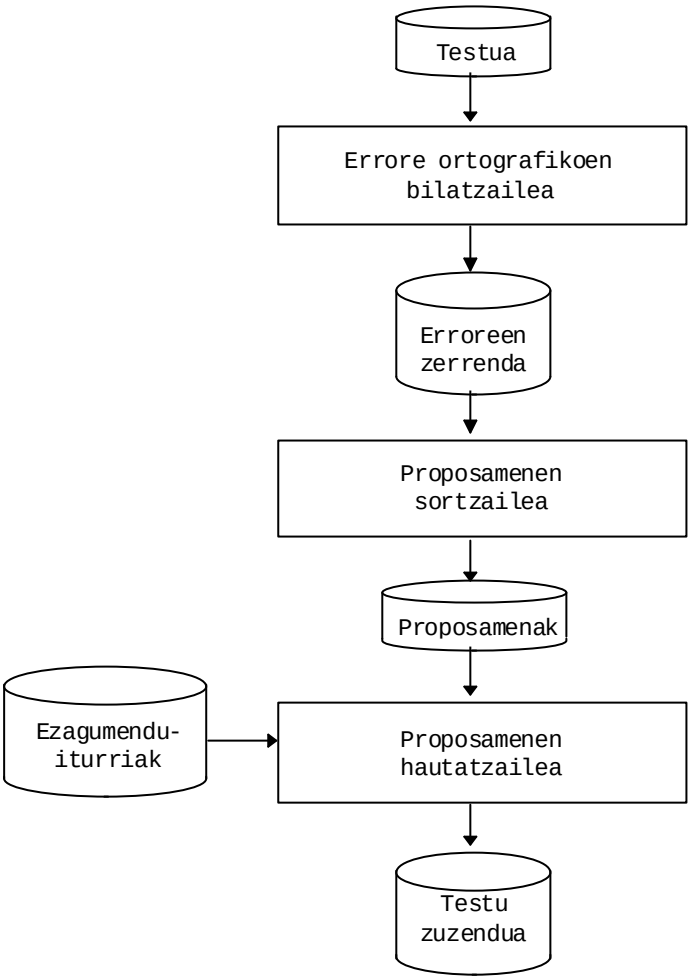
?? Corpus handien bilketa. Erroreen benetako agerpenak izateko, corpus handiak beharko dira etorkizunean. Internet-en gorakada dela eta, lan hau erraz daiteke, baina gogoratu behar dugu corpus horiek lortu ondoren oraindik erroreen eskuzko markaketa egongo dela.

V.2.4 Errore ortografikoen zuzenketa

V.2.4.1 Sarrera

§ V.2.2n eta § V.2.3n errore sintaktikoen detekziorako bi hurbilpen ezberdin aztertu ditugu. Lehenago, errore ortografikoak aipatu ditugunean esan dugu detektatu ondoren zuzenketa geratzen dela, proposamen bat baino gehiago egoten delako askotan, eta horien artean zuzena dena aukeratzeko ezagumendu sintaktiko eta semantikoa behar dela. Errore ortografikoekin gertatzen den bezala, metodoak definitu behar dira zuzentzaile sintaktiko baten zuzenketa-proposamenak baztertzeko, eta horregatik puntu honetan lan horri ekingo diogu. Eremu zabal hau murriztearren, gure kasuan zuzentzaile ortografikoen proposamenak diskriminatzera mugatuko dugu lan hau, eta horren arrazoi nagusia esperimentaziorako datuak lortzeko erraztasuna da, erroreen agerpenak ateratzeko nahikoa izango delako eskuragarri dagoen zuzentzaile bat testu bati pasatzea. Gainera, proposamenen diskriminazioa egiteko ezagumendu-iturri ezberdinen ahalmena neurtu nahi dugu (ikus V.6 irudia). Ingelesa izan da linguistika konputazionalean gehien landu den hizkuntza, eta horregatik tresna ugari daude eskuragarri bere tratamendurako.

Arrazoi honengatik, ingelesa aukeratuko dugu esperimentaziorako hizkuntzatzat, euskararentzat baino tresna gehiago garatu direlako eta probarako eskuragarri daudelako.



V.6 irudia. Erroreen zuzenketarako proposatutako sistemaren eskema.

Erroreen zuzenketa automatikoaren problema oraindik ikerlerroa da, erabilgarri dauden teknikak mugatuak daudelako estaldura eta doitasunean. Zuzentzaile ortografiko gehienek, erroreak detektatzeaz gain zuzenketarako kandidatuak diren hitzen multzoa aurkezten diote erabiltzaileari. Dena dela, kasu askotan beharrezkoa da zuzenketa guztiz automatikoa egitea (OCR programetan, edo denbora errealean lan egin behar duten ahotsaren tratamendurako edo itzulpen automatikorako sistemetan).

Testuinguruaren araberako errore ortografikoen zuzenketa, azken finean, proposamen bakoitza ordezkatuz lortzen diren sententzien arteko aukeraketa egitea da (Mays *et al*. 1991). Arazoa horretarako behar diren ezagumendu motak (lexikoa, fonologikoa, sintaktikoa, semantikoa edo estatistikoa) adierazi, erabili eta konbinatzea da.

Esperimentuak egiteko, ingeleserako Ispell zuzentzaile ortografikoa aukeratu dugu. Bere ezaugarri nagusien artean estaldura zabala, fidagarritasuna eta malgutasuna aipatuko ditugu. Proposamenak aukeratzeko, murriztapen-gramatikaren formalismoa aukeratu dugu ezagutza sintaktikoaren ordezkari izateko, bere desanbiguaziorako ahalmenagatik (ikus § III.3.1). Dentsitate kontzeptuala (Agirre eta Rigau 1996) izenen bereizketa semantikoa egiteko erabiliko da, izenen arteko distantzia neurtuko baitu Wordnet (Miller 1990) erabiliz. Testuinguruarekiko lotura semantikoa corpusetatik ere atera da, kokakidetza eta agerkidetzen ezaugarri estatistikoen bidez (Yarowsky 1994). Azkenik, hitzen maiztasunak ere tratatuko dira, dokumentukoak zein corpus orokorretakoak, sistema konbinatuko ezagumendu-iturriak osatzeko. Lan honen beste helburu bat lengoaiaren eredu desberdinen azterketa egitea izan da, erroreenaz gain desanbiguaziorako beste aplikazioetan tratatu ahal izateko. Horregatik, nahita ez genuen errore-eredurik egin (tekleatze-erroreak, errore fonologikoak, OCR erroreak, ...), jakinda horiek gehituz gero zuzenketa-tasa hobetu ahal izango dela. Sistemaren ebaluaziorako bi testu-sorta erabili dira: Brown corpusetik (Francis eta Kucera 1967) programa baten bidez sortutako errore artifizialak eta Bank

of English[36] corpusetik ateratako benetako erroreak. Hemen aurkeztuko ditugun esperimentuak (Agirre *et al*. 1998ab) lanetan argitaratuta daude.

V.2.4.2 Errore ortografikoen zuzenketa automatikoa

Ingeleseko errore ortografikoen zuzenketa automatikoan egin dugun lana aurkezteko, lehenengo erabilitako tekniken deskribapen laburra emango dugu, ondoren egindako esperimentuak eta emaitzak azaltzeko.

V.2.4.2.1 Erabilitako teknikak

Erabilitako tekniken azalpen laburra emateko, hau esango dugu:

?? Murriztapen-gramatika (MG). Formalismoa bereziki prestatuta dago anbiguotasuna ebazteko, eta horregatik egokia da proposamenen diskriminaziorako. Gure aplikazio honetan, ezagutzen ez den hitz bakoitzeko Ispell-ek sortutako proposamenen analisi morfologikoak (ENGTWOL; Karlsson 1995) jarri genituen. V.4 adibidean *bos* hitz okerraren *bop* zuzenketa kentzen du, testuinguru horretan aditzaren interpretazioa ez delako zuzena.

Ispell-en eta MGren estaldura zabaleko lexikoiak independenteak direnez, batean eta bestean oker edo ezezaguntzat emandako hitzak desberdinak ziren. Arazo hori saihesteko, hitz bat lexikoi bietako batean zegoenean zuzentzat eman genuen.

```
...
"<our>"
    "our" PRON PL ...
"<bos>" ; ERRORE ORTOGRAFIKOA
    "boss" N S
    "boys" N P
    "bop"  V S
    "Bose" <Proper>
"<are>"              ...
```

```
...
"<our>"
    "our" PRON PL ...
"<bos>" ; ERRORE ORTOGRAFIKOA
    "boss" N S
    "boys" N P
    ~~"bop"  V S~~
    "Bose" <Proper>
"<are>"              ...
```

V.4 adibidea. Murriztapen-gramatikaren aplikazioak aditzaren interpretazioa kentzen du.

?? Dentsitate kontzeptuala (DK). MGk askotan ezin du bereizketarik egin kategoria bereko proposamen bat baino gehiago dagoenean, ezagutza sintaktikoa erabiltzen duelako. Horregatik Wordnet-en oinarritutako hitzen adiera-desanbiguatzailea erabili genuen (Agirre eta Rigau 1996). Gure kasuan, desanbiguatzaileak testuingurutik semantikoki hurbilen dagoen proposamena aukeratu beharko luke. Wordnet-en orduko egoera zela eta, neurri hau proposamen guztiak izenak direnean bakarrik aplikatu ahal izan da.

[36] http://titania.cobuild.collins.co.uk/boe_info.html

?? Maiztasunei buruzko estatistikak (DM eta BM). Hitz-formen maiztasunak errorea agertu den dokumentutik hartu ziren (Dokumentuko maiztasuna, DM) edo Brown corpuseko dokumentuetatik (Brown-eko maiztasuna, BM).

?? Testuinguruaren estatistikak (TS). Yarowsky-k (1994) emandako ideiak jarraituz, proposamenek inguruko hitzekin duten hurbiltasuna neurtzeko metodoak erabili genituen. Testuinguruari buruzko ezaugarriak Brown corpusetik lortu ziren (esperimentu honen probarako testuak kenduta). Hartutako ezaugarriak hauek ziren: hitz-formen bigramak, hitz-formen trigramak eta ?20 hitzeko leihoko hitzak.

?? Beste heuristikoak (H1 eta H2). Maiuskulaz hasitako proposamenak baztertu egin ziren hitz okerra minuskulaz hasi eta minuskulaz hasitako beste proposamenak zeudenean. V.4 adibidean *bos* hitzaren laugarren irakurketa (*Bose*) ken liteke, hasierako hitzarekin duen distantzia bikoa delako (H1 heuristikoa). Heuristiko hau oso fidagarria izan da[37], eta horregatik ondoren azalduko diren esperimentu guztietan aplikatuko da.

Lehen emaitzak atera eta gero, 4 karaktere baino gutxiagoko erroreek (*si*, *teh*) proposamen gehiegi (askotan 30 proposamen baino gehiago) ematen zutela ikusi genuen, desanbiguatzeko oso zailak direnak. Hitz hauek gure metodoaren
errore-iturri nagusietako bat zirenez, emaitzak hitz hauek sartuz eta sartu gabe lortu genituen (H2 heuristikoa).

?? Oinarrizko tekniken konbinazioa botoak erabiliz. Teknika guztien konbinazioak izan genituen gogoan, adibidez: MG+BM, BM+DM, MG+DM+TS, ... Botoetan oinarritutako metodoak boto gehien jasotzen duen proposamena hobetsiko du, botoaren balioa zenbaki osokoa izanik. Metodo honen abantaila handia sinpletasuna da, eta beste teknika sofistikatuagoek (pisuen zenbaki zatiki onenak lortzeko optimizazio-algoritmoak erabiliz) bezain emaitza onak lortzen ditu.

Teknika bakoitzaren botoaren pisua alda daiteke, adibidez, MGri bi boto eta BMri boto bat emanez (konbinazio hau MG2+BM1 izendatuko dugu). Honek esan nahi du BMren kandidatua(k) bakarrik hautatuko dela MGk beste aukerarik ez badu hartzen. Pisuen konbinazio ezberdinak probatu ziren.

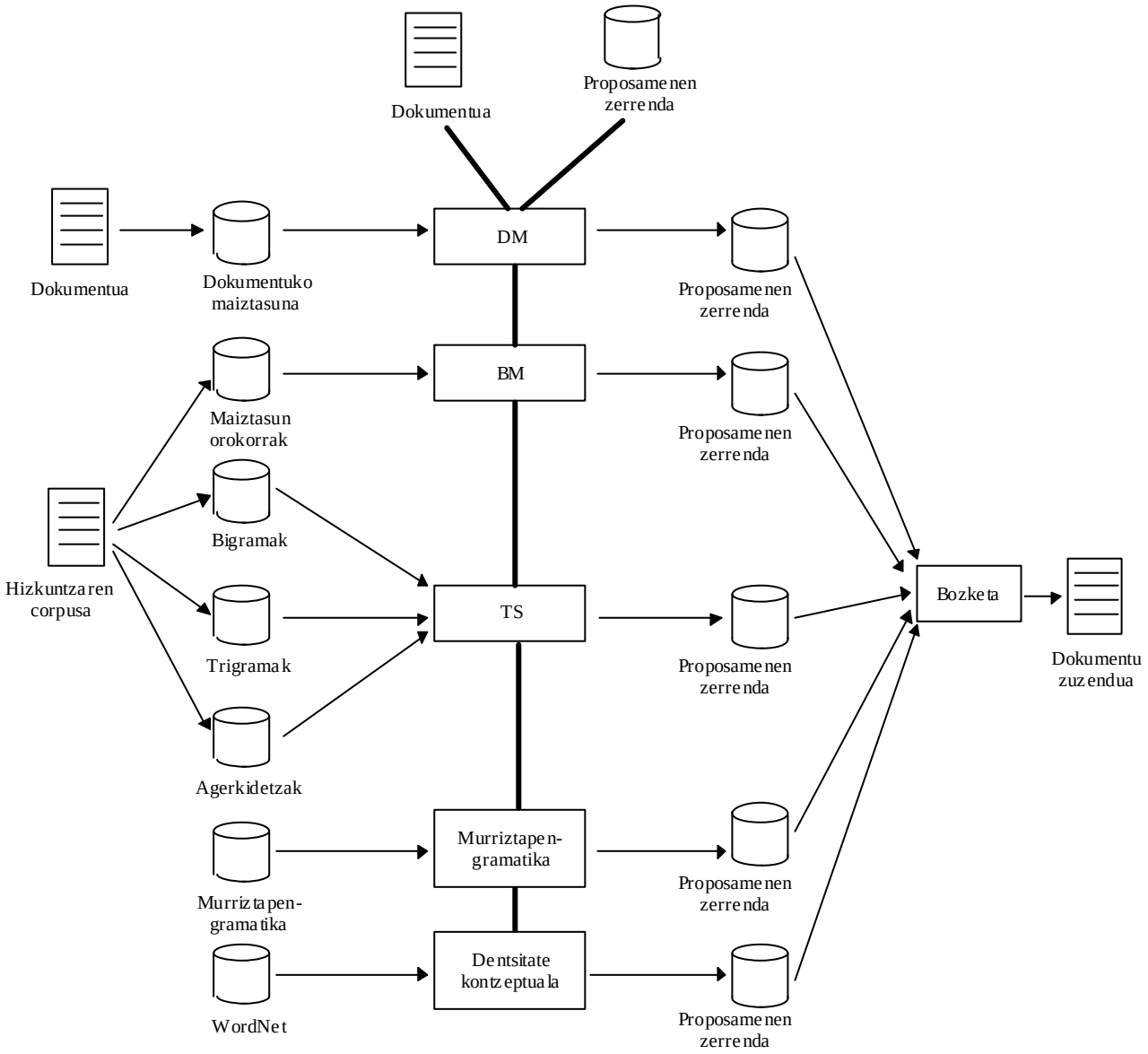
Testu-sorta batentzako teknika eta pisuen konbinazio onenak alda daitezkeenez, erroreen corpusa bitan zatitu genuen, lehenengo erdian konbinazio guztiekin saiatuz, eta konbinazio onenen proba bigarren erdian eginez (ikus geroago proba-kasuen antolaketa).

[37] Ispell-en estaldura handiko hiztegian izen propioen zerrenda luzea dagoenez, askotan minuskulaz hasitako erroreentzat izen propioen interpretazioak ematen dira, gehienetan zuzenak ez direnak.

V.2.4.2.2 Esperimentuak

Ezagumendu mota bakoitza erabiliz, proposamen bat aukeratzeko zuzentzaile bana osatu genuen, eta era ezberdinetan konbinatu genituen (ikus V.6 irudia, Agirre-ren tesitik (1999) hartuta). Lehenbiziko fase batean aukera guztiak ebaluatu eta onenak hartu ziren ausaz sortutako erroreen corpusaren zati batean. Bukatzeko, konbinazio onenak benetako erroreetako testuetan probatu ziren.

Lehenago esan bezala, bi corpus erabili ziren esperimenturako. Lehenengoa erroreak ausaz modu sistematikoan sortuz atera zen Brown corpusaren zati batetik, eta bigarrena benetako erroreak zeukan corpus batetik. Lehenengo zatia egokia zen esperimentaziorako, erroreak automatikoki sortuak zirenez egiaztapena zuzenean egin zitekeelako. Bigarrenak, aldiz, erroreen zuzenketaren eskuzko markaketa eskatzen zuen, baina beste alde batetik egoera erreala islatzen zuen (Brown corpusean, bera sortzeko modua dela eta, ez dago ia errore ortografikorik, zuzendua izan baita).



V.6 irudia. Proposamenaren hautapenerako ezagutza-iturriak eta konbinazioa egiteko sistema.

Modu artifizialean sortutako erroreen corpusak (corpus artifiziala deituko diogu laburtzearren) honako ezaugarriak dauzka: SemCor-etik (Brown corpusaren azpimultzoa) lagin bat atera zen, ausaz 150 paragrafo hartuz. Honek 5050 esaldiko eta 12.659 *token*-eko hasierako corpusa ematen du. Errore ortografikoen simulazioa egiteko *antispell* izeneko programa bat exekutatu zen, Damerau-ren erregelak aplikatuz 20 hitzeko errore bat sortzeko (hitz ezezagunak ukitu gabe). Antispell 8 aldiz exekutatu zen hasierako corpus horren gainean, testu bera baina errore

ezberdinak dituzten 8 corpus sortuz. Esaldi batean errore bat baino gehiago egon liteke, eta paragrafo batzuetan ez zegoen errorerik.

Benetako erroreen corpusa (hemendik aurrera corpus erreala deituko diogu) *Bank of English Corpus*-eko aldizkarien testuez osatua dago, lehenago zuzenketa ortografikorik egin gabea. Honelako corpusak lortzeak zailtasuna du, egun zuzentzaile ortografikoen erabilera zabaldua baita. Beste arazo bat erroreen markaketan eta zuzenketaren aukeran egin behar den lan handia da, ebaluazio automatikoa egin ahal izateko gero. Lehen aipatu dugu errore artifizialen corpusa bi azpimultzotan banatu zela. Lehenbizikoa, 1, 2, 3 eta 4 multzoez osatua, metodoa fintzeko (botoak eta konbinazio onenak ateratzeko) erabili zen. Bigarren azpimultzoa eta testu errealak azken ebaluazioa egiteko erabili ziren.

Corpus biak Ispell-etik pasatu ziren, hitz ezezagun bakoitzeko bere zuzenketarako proposamenak txertatuz. V.13 taulan ikusten da errore ortografikoak ausaz sortuz gero %23,5a benetako hitzak direla, eta lan honetatik kanpo geratuko dira. Testu errealetan antzeko kontaketa egin ez genuen arren, gaineko azterketa eginda antzeko portzentajea gertatzea espero dezakegu.

	1. erdia	**2. erdia**	**erreala**
hitzak	47584	47584	39733
erroreak	1772	1811	-[38]
ez-hitzak diren erroreak	1354	1403	369
Ispell-en proposamenak	7242	8083	1257
proposamen bat baino gehiagoko erroreak	810	852	158
hitz luzeen erroreak (H2)	968	980	331
hitz luzeentzako proposamenak (H2)	2245	2313	807
hitz luzeak (H2) proposamen askorekin	430	425	124

V.13 taula. Esperimenturako datuak corpus artifizialean (2 zati) eta errealean.

Benetako erroreen corpusarekin jarraitutako prozedura hau izan zen: Ispell aplikatu ondoren ez zen zuzenketarik eman 150 hitzentzat (gehienak izen propioak eta erdarakadak), eta 300 inguru zeuden jarraian doazen bi hitz lotuz edo atzizkien lotura bereziak aplikatuz lortuak (Ispell-ek ondo ezagutu zituen gehienetan). Horiek kenduta, 369 hitz-forma oker geratu ziren. Proposamenak aztertu ondoren ikusi genuen proposamen zuzena Ispell-ek emandakoen artean zegoela ia kasu guztietan.

Hitz okerren proposamenen hautapena aztertuz, ia erdiek proposamen bakarra dute. Proposamen bat baino gehiagokoek emango dute egin behar den lanaren neurria. Adibidez, corpus errealean 158 hitz-forma daude 1.046 proposamenekin, hitzeko 6,62 aukerako batez bestekoa emanez. Hiru letra edo gutxiagoko hitzak kontatzen ez badira, orduan 807 proposamen daude, hau da, 4,84 aukera hitzeko.

Hiru balio kontsideratu ditugu emaitzak ebaluatzerakoan:

- Estaldura: erroreetariko zenbatetan teknikak (edo tekniken konbinazioak) erantzuna ematen duen.
- Doitasuna: erroreetariko zenbatetan uzten duen proposamen zuzena hautatutako proposamenen artean.
- Errore bakoitzeko utzitako batez besteko proposamen-kopurua. Hau da, askotan doitasuna eta estaldura elkarren kontrakoak direnez, batzuetan teknika batek proposamen bat baino gehiago aukeratzen du, horrela doitasun handiagoa lortuz.

38 Hitz ezezagunak aztertu genituenez, ez ziren kontatu benetako hitzak ematen dituzten erroreak.

Konbinazio hoberenak aurkitzeko, garapenerako corpusean (errore artifizialen lehen erdia) lortutako emaitzak aztertu ziren. Konparazioa errazago egiteko, proposamenaren ausazko aukeraketaren emaitza ere atera zen. H1 heuristikoa kasu guztietan aplikatu zen, eta esperimentuak H2 erabiliz eta erabili gabe egin ziren. Bigarren fase batean, konbinazio hoberenak corpus artifizialaren bigarren erdiaren gainean ebaluatu ziren, emaitzak lehenengo zatia lortutakoen parekoak zirela, portzentajeak pixka bat jaitsi arren. Proba honekin konbinazio hoberenak errore ezberdinetan mantentzen direla baieztatzen zen.

	estaldura %	**doitasuna %**	**# proposamenak**
Oinarrizko teknikak			
ausazko aukeraketa	100.00	69.92	1.00
ausaz+H2	89.70	75.47	1.00
MG	99.19	84.15	1.61
MG+H2	89.43	90.30	1.57
DM	70.19	93.05	1.02
DM+H2	61.52	97.80	1.00
BM	98.37	80.99	1.00
BM+H2	88.08	85.54	1.00
TS	97.02	89.10	1.02
TS+H2	85.64	91.50	1.01
Konbinazioak			
MG1+DM2	100.00	87.26	1.42
MG1+DM2+H2	89.70	90.94	1.43
MG1+DM1+BM1	100.00	80.76	1.02
MG1+DM1+BM1+H2	89.70	84.89	1.02
MG1+DM1+TS1	100.00	90.80	1.24
MG1+DM1+TS1+H2	89.70	93.10	1.20
MG1+DM1+TS2	100.00	89.70	1.04
MG1+DM1+TS2+H2	89.70	91.80	1.03

V.14 taula. Konbinazio hoberenak (testu errealetako erroreak).

Azken pausoan konbinazio hoberenak benetako errore ortografikoen corpusean ebaluatu genituen. Arreta proposamen bat baino gehiagoko kasuetan jarriz gero, hobeto ikus daiteke teknika bakoitzaren ekarpena. V.14 taulak emaitza orokorrak azaltzen ditu, eta V.15n proposamen anitzeko hitzenak (emaitza hoberenak grisez agertzen dira). Azken honetan, hitzeko 6,62 proposamen daude kasu orokorrean (corpus artifizialean baino 2 gutxiago), eta 4,84 H2 heuristikoa aplikatuz (corpus artifizialean baino 1 gehiago). Hurrengo lerroetan emaitzen balorazioa emango dugu, testu errealekin hasi eta ondoren corpus artifizialarekin konparatuz.

	estaldura %	doitasuna %	# proposamenak
Oinarrizko teknikak			
ausazko aukeraketa	100.00	29.75	1.00
ausaz+H2	76.54	34.52	1.00
MG	98.10	62.58	2.45
MG+H2	75.93	73.98	2.52
DM	30.38	62.50	1.13
DM+H2	12.35	75.00	1.05
BM	96.20	54.61	1.00
BM+H2	72.84	60.17	1.00
TS	93.21	74.16	1.05
TS+H2	67.28	75.36	1.03
Konbinazioak			
MG1+DM2	100.00	70.25	1.99
MG1+DM2+H2	76.24	75.81	2.15
MG1+DM1+BM1	100.00	55.06	1.04
MG1+DM1+BM1+H2	76.54	59.68	1.05
MG1+DM1+TS1	100.00	78.51	1.56
MG1+DM1+TS1+H2	76.54	81.58	1.53
MG1+DM1+TS2	100.00	75.94	1.09
MG1+DM1+TS2+H2	76.54	78.11	1.08

V.15 taula. Proposamen bat baino gehiagoko erroreen emaitzak (testu errealetako erroreak).

Teknika bakoitza modu independentean aplikatuz[39]:

?? Teknika guztiak dira ausazko aukeraketa baino hobeak.

?? TSk dauka doitasun altuena (%74), %93ko estaldurarekin.

?? BMk doitasun txikiagoa (%54) baina estaldura zabala du (%96).

?? DMk doitasun txikiagoa (%62) eta estaldura txikiagoa (%30) ditu.

?? MGk %62ko doitasuna eta ia %100eko estaldura du, baina proposamen asko mantentzen ditu (2,45).

Teknikak konbinatzerakoan, emaitzak hobetu egiten dira:

?? MG+DM+TS konbinazioak emaitza onenak ematen ditu estalduran (%100etik hurbil) eta doitasunean (%78a V.15 taulan).

?? TSk pisu bikoitza hartzen badu, MG1+DM1+TS2 konbinazioan, doitasuna gutxi jaisten da (%76), baina proposamen-kopurua onargarriagoa da (1,56tik 1,09ra).

?? %70eko doitasuna lortzen da MG1+DM2 probatuz. Ikusten denez, MGk DM teknikaren estaldura igotzen du, proposamen-kopuruaren igoerarekin batera (1,9). DMren estaldura handiagoa izango balitz, proposamen-kopurua neurri berean jaistea espero genezake, corpus artifizialen parera iritsiz (1,28).

39 Besterik ez bada esaten, aipamenak testu errealei buruz dira (VI taula).

?? MG1+DM1+BM1 konbinazioak antzeko estaldura ematen du hitz bakoitzeko ia interpretazio bakarra utziz, baina doitasuna %55era jaisten da.

?? Estaldura osoa beharrezkoa ez bada, H2 heuristikoaren erabilerak doitasuna %3a igotzen du konbinazio guztietarako.

Argigarriagoa da corpus artifizialeko eta V.15 taularen arteko konparazioa. Azken taula honetan teknika guztien performantzia jaisten da. MG, TS eta BM 15, 10 eta 20 puntu jaisten dira, hurrenez hurren. DMk doitasunean 20 puntu eta estalduran 50 puntuko jaitsiera du. Azken honen performantziaren okertzea espero zitekeen, benetako erroreen corpuseko dokumentuen luzera 50 hitzekoa delako batez bestekoan. Hau corpusaren ezaugarri berezitzat jo behar dugu, ez baikenuen testu luzeagoak lortzerik izan, eta neurri handiagoko dokumentuak izatera, testu artifizialen antzeko estaldura espero genezake.

Konbinazio onenak mantendu egiten dira testu errealetarako. Doitasun altuena MG+DM+TS konbinazioarentzat da (H2rekin eta H2 gabe). Proposamen-kopurua handiagoa da testu errealetan (1,56 eta 1,12). Honen kausa bat DM eta TS errore guztiak ez estaltzea izan liteke, eta ondorioz MGren proposamen asko ez dira ukitzen.

Testu errealen performantziaren jaitsiera azaltzeko faktore ezberdinak aipatuko ditugu. Lehenengo, dokumentuen luzera txikiagoak eragin handia izan du DMn. Bigarren, erroreen sorrera ere ezberdina da: errore ortografikoak sortzeko erabili dugun algoritmoak (antispell) gehiago hartzen ditu maiz agertzen diren hitzak, askotan hitz laburrak. Eredu hau benetako erroreekin zein puntutaraino bat datorren aztertu beharko dugu.

V.14 taularen emaitzak hartuz, zuzenketa automatikorako sistema baten irteera zein den aurreikusi dezakegu. MG1+DM1+TS2 konbinazioarekin testu bateko errore guztientzako 25etik 24 aldiz proposamen bakarra lor daiteke %90eko doitasunarekin (1,04 proposamen batez besteko) edo bestela doitasun altuagoa lor daiteke (%93), %90eko estaldura eta 1,20 proposamenekin (MG+DM+TS+H2).

V.2.4.3 Ondorioak

Egindako lana laburtzeko, esan behar dugu errore ortografikoen zuzenketarako teknika ezberdinen azterketa egin dugula, helburua errore bakoitzeko proposamen bakarra ematea izan dela. Problemaren aspektu zailenetakoa emaitzen proba egitea izan da, datuak urriak izaten direlako. Horregatik, modu artifizialean sortutako erroreen corpusa erabili dugu, azken probarako benetako erroreen beste corpusarekin batera.

Aurretik erroreen zuzenketa automatikoan egindako lanekin konparatuz gero (Yarowsky 1994, Golding eta Schabes 1996), esan behar dugu sistema horiek lehenago aurreikusitako hitzen nahasketan oinarritzen zirela (azentuak edo antzeko hitzak). Gure lanean, berriz, sistemak aukera bat hartu behar zuen edozein errorerentzat, eta ezinezkoa zen aldez aurretik proposamenen diskriminaziorako aukerak biltzea. Gure kasuan, gainera, errore guztiak zuzendu nahi dira, nahiz eta kasu batentzako datuak urriak izan, eta lan horietan agerpen ugari zeukaten hitzak hartzen ziren.

Emaitzak aztertu eta gero, esan dezakegu hobetu egiten direla testuinguru gehiago hartzen den heinean. Hitz-formen maiztasuna irizpide gordina baina lagungarria da zuzenketa aukeratzeko. Doitasuna igo egiten da testuinguru hurbilagoak hartu ahala, dokumentuaren maiztasunak, murriztapen-gramatika edo testuinguruko ezaugarrien modukoak.

Benetako erroreen corpusetik ondorio batzuk atera ditzakegu. Lehenengo, zuzenketa Ispell-en proposamenen artean dagoela ia %100ean, honek errore guztiak zuzendu daitezkeela emanez. Bigarren, uneko sistematik espero daitekeen irteera errore ortografikoen 25etik 24tan proposamen bakarra lortzea dela, estaldura osoarekin eta %90eko doitasunarekin, edo bestela %93ko doitasuna eta %90eko estaldura, batez besteko 1,20 proposamen utziz.

Bi teknikek, Brown corpuseko maiztasuna eta Dentsitate Kontzeptualak, ez dute emaitza erabilgarririk eman. DK bakarrik erroreen proportzio txiki bati aplikatzen zaio, honek Wordnet-en estaldurarekin erlazio zuzena du, eta horregatik ezin dugu beste ondoriorik atera.

Etorkizunerako emaitzak hobetzea espero dugu. Lehenengo, errore errealen corpusak dokumentu oso laburrak zeuzkan, eta honek DMren performantzia jaistea ekarri du. Testu luzeagoak lortzera, emaitza hobeak aterako genituzke. Bigarren, esan behar da Ameriketako ingelesaren testuinguruko ezaugarriak Erresuma Batuko

ingeleseko testuak zuzentzeko erabili zirela. Lengoaien arteko korrespondentzia-arazoak kenduz gero, emaitzen hobekuntza lor daiteke. Bukatzeko, erabilitako teknikak hobetzea ere badago. Errore ortografiko arrunten eredurik ez genuen erabili (hauek errorearen sorrerarekin lotuta egongo dira: OCR, tekleatze-erroreak, ...). Testuinguruko ezaugarriei buruz esan behar dugu hitz-formen ezaugarriak bakarrik erabili ditugula, baina kategoria edo lemak erabiltzeak hobekuntzak ekar ditzake. Bestalde, tekniken konbinaziorako beste metodoak proba daitezke, nahiz eta hemen ez dugun uste aldaketa handia gertatuko denik.

V.3 Erroreen detekzio eta zuzenketari buruzko lanen ondorioak eta hurrengo pausoak

Kapitulu honetan sintaxiaren erabileraren azterketa egin dugu errore ortografiko eta sintaktikoen detekzio eta zuzenketan. Aplikazio hori lehenago garatutako tresna morfologiko eta sintaktikoetan oinarritu da, eta tresna berriak osatzen eta azkartzen diren heinean erroreen tratamendurako ahalmenak ere gorantz egingo du.

Egindako lanaren ekarpen nagusiak hauek dira:

?? Euskaraz egindako errore sintaktikoen sailkapena egin da. Euskararen ezaugarriak direla eta, berezitasunak dauzka gehien landu diren ingelesa edo espainiera bezalako hizkuntzekin konparatuz. Sailkapenarekin batera, erroreen lehen azterketa egin da, tratamendu automatikorako bideragarritasunaren arabera.

?? Murriztapen sintaktikoen erlaxazioa esperimentatu da errore sintaktikoen detekziorako. Teknika hau egokia da euskaraz maiz gertatzen diren erroreak tratatzeko, komunztadurak kasu. Teknikaren ekarpena isolatzeko, probak gramatika sintaktikoaren estalduran zeuden esaldien gainean egin dira. Esperimentuek frogatu dute teknika honek errore gehienak detektatzen dituela, ia guztietan proposamen zuzena emanez (beste batzuen artean). Tratatutako errore motak zabaltzen diren heinean, murriztapenen askatze graduala beharrezkoa dela frogatu dugu.

?? Gramatikaren estalduratik kanpo dauden eta fenomeno sintaktiko lokaletan egindako erroreak tratatzeko, errore-patroietan oinarritutako detekzioa probatu dugu. Esperimentua corpusetan aurkitutako erroreen gainean egin da, horrela erroreen detekzioa eta alarma faltsuen arazoa elkarrekin tratatuz. Horretarako baterakuntzan oinarritutako analizatzaile sintaktikoa eta egoera finituko patroiak erabili dira. Daten adierazpenak aukeratu dira sistema probatzeko, erroreen iturri aberatsa direlako eta beste erroreen adierazgarritzat har daitezkeelako. Emaitza onak lortu dira, doitasunean zein estalduran.

?? Azkenik, zuzenketa automatikoa ere landu da, errore ortografikoen proposamen bakarra lortzeko asmoz. Horretarako, testu errealetako erroreen gainean egin dira probak, ezagumendu mota desberdinen (sintaktikoa eta semantikoa) ekarpena konbinatuz. Lortutako sistemak estaldura eta doitasun onak lortu ditu, kasu gehienetan proposamen bakarra utziz.

Bukatzeko, esan behar dugu etorkizunerako ikerlerroa izaten jarraituko duela erroreen tratamenduak, testu errealei aplikatuz modu eraginkorrean zuzenketa-tasa altuak eta alarma faltsuen kopuru baxua lortzen duten sistemak egiteko. Arlo honetan egin beharreko lanen artean hauek aurreikusten ditugu (Kukich 1992, Oliva 1997):

?? Errore-patroien bidezko tratamenduaren zabalpena. Datekin egin den moduan, testuinguru sintaktiko lokaleko azterketarekin detekta daitezkeen errore gehiago aztertu eta tratatzeko asmoa dugu. Horien artean postposizioak edo azpikategorizazioarekin lotutako erroreak aipa daitezke (ikus V.8 taula).

?? Analisi sintaktiko sendoa lortzen den neurrian, murriztapen sintaktikoen erlaxazioan oinarritzen diren sistemen inplementazioari ekin ahal izango zaio. Euskararen kasuan, erlaxazioa zuzenean edozein esaldiri aplikatzea kostu handikoa denez (exekuzio-denboraren aldetik, eta gainera etekin gutxikoa esaldi bakoitzean errore mota posible guztien agerpena frogatu nahi bada), gure ustez patroietan oinarritutako sintaxia eta erlaxazioa konbinatu beharko dira: patroiek detektatu beharko dituzte errorea eman dezaketen esaldien zatiak, eta ondoren baterakuntzan oinarritutako analizatzailea aplika dakioke zati horri. Horrela, bakarrik testuinguru horretan posible diren errore batzuk bilatuko dira.

?? Linguistika konputazionalaren garapen-prozesuan gertatu den bezala, corpusen erabilera erroreen tratamenduan ere zabaldu egin beharko da. Ez da soluzio erraza daukan problema, corpusen lorpenak dauzkan problemez gain, erroreen tratamendurako corpusak lortzeak corpus berezituak (hau da, erroreak dauzkaten corpusak) behar dituelako. Gainera, corpus horiek lortu eta gero, ebaluaziorako erabilgarri izateko eskuzko lan handia egin beharko da. Aurrekoarekin lotuta, esan behar dugu erroreen detekziorako sistemen mugarik handiena eskala zabaltzeko (*scale up*) zailtasuna dela, erroreak banan-banan edo errore-multzoka tratatu behar direlako, eta errore-kopurua infinitutzat kontsidera daitekeelako. Horregatik, testu errealen tratamendurako erroreen adibideak lortzea eskuz egin beharko da ia beti, gainditzeko zaila den muga jarriz (salbuespentzat hitz jakinen nahasketa (*weather/whether*), errore fonologikoak edo daten moduko erroreak har daitezke, hauetan lanaren zati bat automatikoki egin daitekeelako baina, hala ere, eskuzko lan handia eskatu dute).

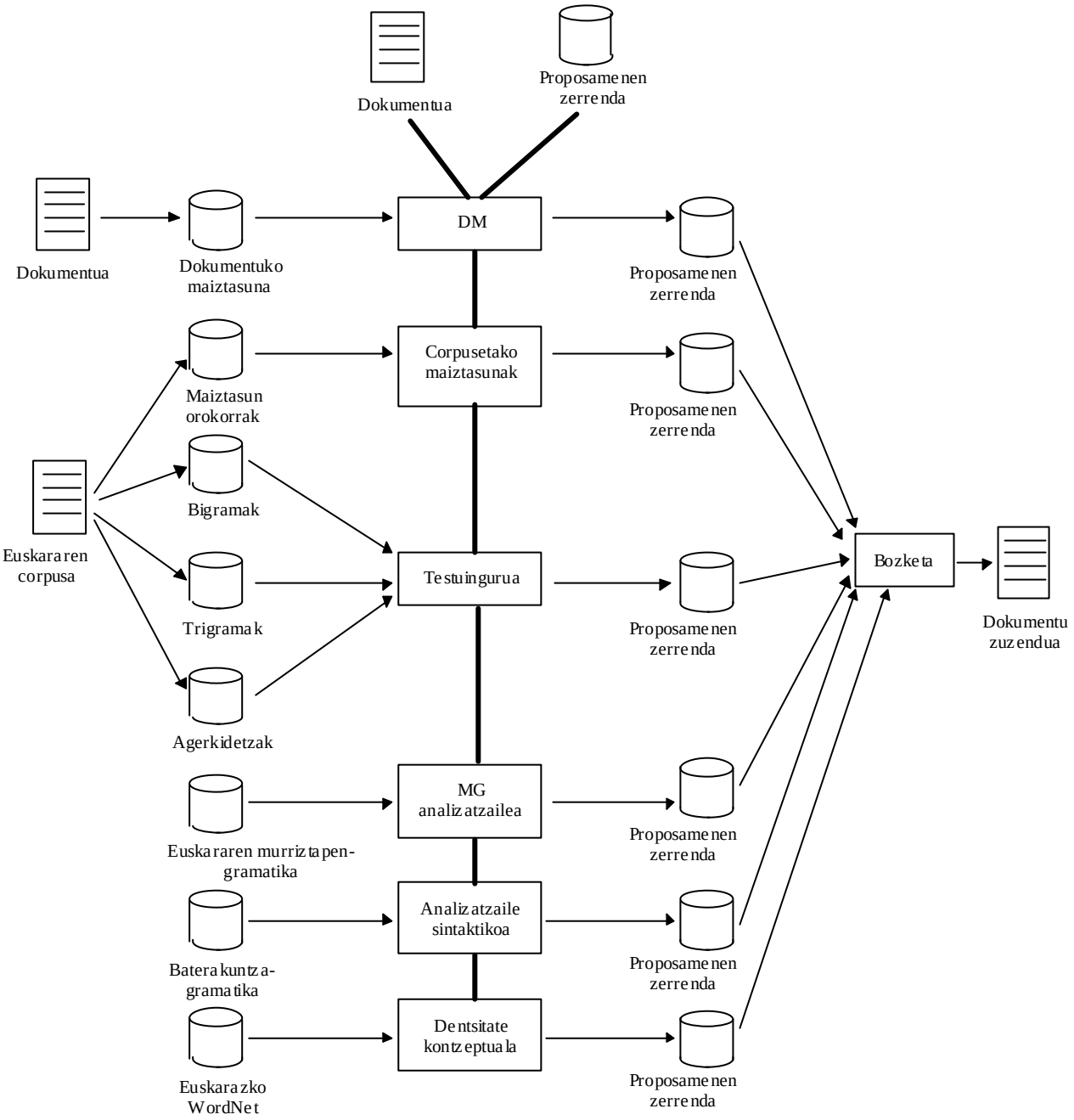
?? Erroreak detektatzeko erregelen ikasketa automatikoa. Behin erroreak markatuta dauzkan corpusa edukita, gure kasuan eskuz landu ditugu detekzio edo zuzenketarako erregelak (gramatikaren gaineko erlaxazioak markatuz edo erroreak harrapatzeko patroiak definituz). Bestalde, errore ortografikoen zuzenketarako, automatikoki lortutako informazioak erabili dira (maiztasunak, testuinguruari buruzko erregelak). Horregatik, ikasketarako metodoen azterketa oso interesgarri ikusten dugu (Golding eta Roth 1996, Mangu eta Brill 1997), horrela sistema garatzeko lan bakarra erroreen adibideak markatzea izango litzatekeelako.

?? Sintaxiaren prozesuan garatutako tresnak erabili ditugunez, ez ditugu asko landu erroreen tratamendu estatistikoak (hitz-formen bigrama/trigramak ezik). Gramatika sintaktikoa

hedatzeko, euskararen neurri estatistikoen aplikazioaren erabilera (Ezeiza 2000) aztertzea interesgarri ikusten dugu, eta honek erroreen tratamenduaren hobekuntza ekarriko duela uste dugu. Antzera gertatuko da azpikategorizazioari buruzko informazioa lortzen den neurrian, errore asko azpikategorizazioaren erabilera okerrarekin lotuta daudelako.

?? Detekzioaz gain, zuzenketak ere gainditu beharreko arazoak planteatzen ditu. Errore ortografikoen zuzenketak hitz baten aldaketa eskatzen du, eta ez dauka lan asko egin beharrik, baina errore sintaktiko baten zuzenketa egiteko, askotan errorea eta bere inguruko beste osagaiak aldatu beharko dira (komunztadura-errore bat dagoenean, adibidez, aditza edo izen-sintagmaren ezaugarriak alda litezke), beti ere esaldi osoaren zuzentasuna mantenduz. Honek sorkuntza morfologikoa eta kasu batzuetan sintaktikoa egitea eskatuko du.

?? Ingeleserako errore ortografikoen zuzenketari ekin diogu emaitza onekin, eta horregatik oso interesgarri ikusten dugun urratzeko bidea euskarari aplikatzea da. Hizkuntza eranskarien tratamenduak arazo berriak dakartza, zuzenketarako proposamenak askotan deklinatuta joango direlako, eta proposamenen aukeraketarako joko aberatsagoak ager daitezkeela sumatzen dugulako. Adibidez *gizonk* hitzak bi proposamen izango ditu (*gizonak* eta *gizonek*), bakoitzak bi interpretaziorekin. Horien artean erabakitzeak sintaxiaren erabilera beharko du. Zuzenketarako ezagumendu-iturrien artean euskararen murriztapen-gramatika, baterakuntzan oinarritutako gramatika, maiztasunak eta agerkidetzak aipa ditzakegu. V.7 irudian modu horretako sistema baten arkitektura aurkezten da, ingeleserako egindakoan oinarrituta, baina euskararako bereziki landu diren baliabideak aipatuz. Baterakuntza-gramatika erabil liteke proposamenen aukeraketarako, hipotesi nagusitzat sintagma luzeenak ematen dituzten proposamenak hartuko direla. Egoera finituko patroiak ere erabil litezke errore jakin batzuetan proposamenak diskriminatzen.



V.7 irudia. Proposamenaren hautapenerako ezagutza-iturriak eta konbinazioa egiteko sistema.

VI Beste aplikazioak

Aplikazioen atala bukatzeko, morfosintaxia eta sintaxian egindako lanaren ondorioz garatu diren bi tresna azalduko ditugu. Batetik, analizatzaile morfosintaktikoa euskararen lematizatzaile/etiketatzaile (EUSLEM izenekoa) baten osagaia izango da (§ VI.1). Bestetik, azaleko sintaxiaren tratamenduaren bidez ikasleen testuetako egitura sintaktikoen patroien maiztasunak neurtzeko tresna bat inplementatu da (§ VI.2)

VI.1 EUSLEM

Corpus bateko testu-hitz bakoitzari bere lema eta kategoria esleitzeko erabiliko den tresna informatikoa da EUSLEM izeneko lematizatzaile/etiketatzailea (Aduriz *et al*. 1996ab). Funtsezko tresna da corpus-analisirako, indexazio automatikorako edo analisi sintaktiko eta semantikorako. Testu errealei aplikatzeko tresna denez, egindako zenbait sistema konbinatuko ditu. Hona hemen bere osagai funtsezkoenak:

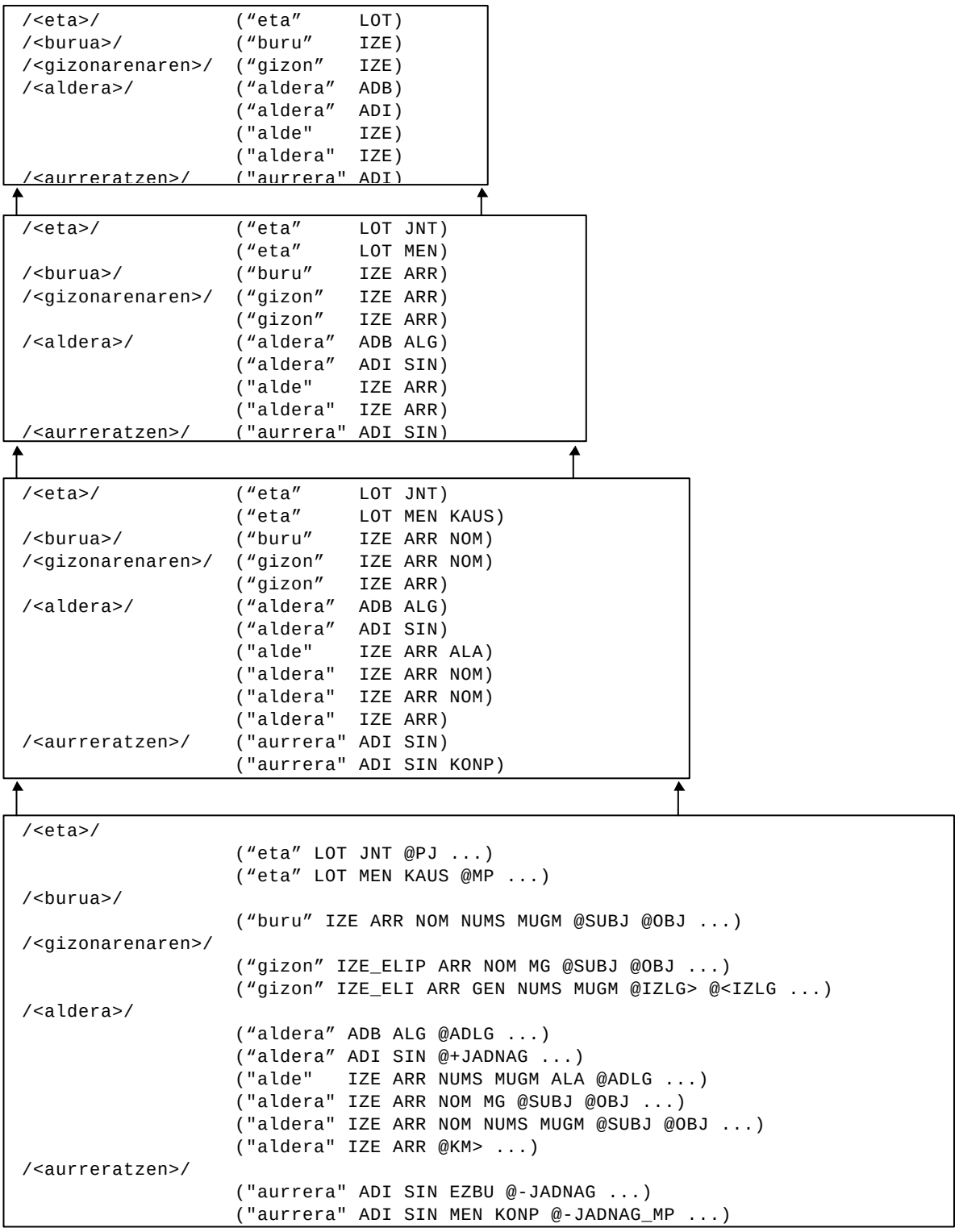
a) Euskararen datu-base lexikala (EDBL). Lehen esan bezala, EUSLEMerako ez ezik beste aplikazio askotarako oinarri lexikala dugu EDBLn.

b) Etiketatzeari begira, aurretik definituriko etiketa-sistema erabili da. Sistema mailakatua da, kategoria soila ematetik analisi morfologiko osoa (deskonposaketa, morfema bakoitzaren informazioa,...) emateraino heda daitekeena.

c) Analizatzaile Morfologikoa (MORFEUS). Euskaraz, hizkuntza eranskaria eta morfologikoki konplexua izanik, ezinbestekoa da analisi morfologikoaren berri emango duen analizatzaile morfologiko bat.

d) Analizatzaile morfologikoaren emaitzak hitz bakoitzaren analisi posible guztiak ematen ditu. Horren gainean aplikatzen da murriztapen-gramatika. Honek, informazio linguistikoan oinarrituz, desanbiguazio morfologikoa burutzen du. Horrez gainera, estatistikan oinarrituriko metodo bat ere aplikatuko da, aurrekoak desanbiguatu gabe uzten dituen kasu bakanetan (Ezeiza *et al*. 1998).

e) Bestalde, EUSLEMek Hitz Anitzeko Unitate Lexikalak (HAUL) ere tratatzen ditu (Aduriz *et al*. 1996c). Unitate lexikal hauen tratamenduak zailtasun gehigarriak ditu. Horregatik, HAULak zehatz deskribatzeko adierazpide formal bat garatu da.



VI.1 irudia. EUSLEMen irteerarako informazio-mailak[40].

40 Irudiaren beheko partean dagoen laugarren mailan, analizatzaile morfologikoaren emaitza (II.2 kapituluan azaldutakoa) aterako da eta, beraz, irudia sinplifikazioa da. Horregatik jarri dira hiru puntu analisi bakoitzaren bukaeran, informazio gehiago dagoela adierazteko.

Beraz, tesi honetako lanak ekarpenak izango ditu EUSLEMen hiru atal garrantzitsutan:

a) Alde batetik, EUSLEMen irteera analisi morfosintaktikoaren emaitza izango da.
§ II.2n esan da emaitza hori osoa izango dela, hitzaren barruko interesgarri den informazio guztia kodetuz. Informazio-aberastasun honek malgutasuna emango dio EUSLEMi, eta hortik informazioa aplikazio-eremu askotarako iragazi ahal izango da. EUSLEMen diseinuan erabaki da lau informazio-maila egongo direla (ikus VI.1 irudia), kategoria soila ematetik informazio morfosintaktiko osorainoko tartea hartuz. Hau da definitutako mailen azalpen laburra:

?? Lehen mailan, hemeretzi kategoria nagusiak (izena, aditza, ...) sartu dira. Hau lematizazio arrunta egiteko oinarrizko etiketa-sistema da.

?? Bigarren mailan kategoria bakoitza azpikategorien etiketekin birfintzen da. Adibidez, aditzak sinple eta konposatuen artean desberdinduko dira.

?? Beste informazio morfosintaktikoak gehituko zaizkio hirugarren mailari, kasua edo numeroaren modukoak (adibidez, kategoria, azpikategoria eta kasua hartuta, 318 etiketa konposatu desberdin aurkitu dira probarako testu batean). Maila honetan erabiltzaileak zehatz dezake behar duen informazioa.

?? Azken mailan analizatzaile morfosintaktikoaren informazio guztia emango da, funtzio sintaktikoak barne. Maila honetako etiketa konposatuen kopurua asko igotzen da (2943 etiketa lehen aipatutako testuan).

```
"<Herriarenak>"
     "herri" IZE ARR DEK GEN NUMS MUGM DEK ABS NUMP MUGM @OBJ @PRED
     "herri" IZE ARR DEK GEN NUMS MUGM ELI DEK ABS NUMP MUGM @OBJ @PRED
"<ziren>"
     "izan" ADT B1 NOR NR_HK ERL MEN ERLT @+JADNAG_IZLG>
"<mendiak>"
     "mendi" IZE ARR DEK ABS NUMP MUGM @OBJ @PRED
"<ataldu>"
     "ataldu" ADI SIN AMM PART ASP BURU DU NOTDEK @-JADNAG
"<eta>"
     "eta" LOT JNT AORG @PJ
"<saldu>"
     "saldu" ADI SIN AMM PART ASP BURU NOTDEK @-JADNAG
     "saldu" ADI SIN AMM PART NOTDEK @-JADNAG
"<egin>"
     "egin" ADI SIN AMM PART ASP BURU NOTDEK @-JADNAG
"<zituzten>"
     "*edun" ADL B1 NOR_NORK NR_HK NK_HK @+JADLAG
"<$.>"
     PUNT_PUNT
```

VI.1 adibidea. Murriztapen-gramatikaren oraingo sarrera.

b) Beste alde batetik, analizatzaile morfosintaktikoa EUSLEMen barne-funtzionamendurako ere lagungarria izango da, murriztapen-gramatikak hartzen

duen sarrera morfosintaxiaren irteera izango delako. Une honetan murriztapen-gramatikak VI.1 adibideko formatuan hartzen ditu testuak.

Ikusten denez, oraingo deskribapen morfologikoa segmentatzailearen irteera da, eta horregatik kasu batzuetan bi balio daude numero, kasu eta mugatasunerako, hau da, lehen aipatutako problemak (ikus § II.2) agertzen dira. Arazo hori gainditzeko, murriztapen-gramatikan kasu horiek tratatzeko erregela bereziak landu dira, erregelen definizioa konplikatuz eta orokortasuna galduz. Analisi morfosintaktikoak egoera honen konponketa ekarriko du, gramatikarien lana erraztuz, eta azken gramatikaren ulergarritasuna eta mantentzea hobetuko du.

c) Desanbiguazioa euskararen murriztapen-gramatika erabiliz. Honen garapena § II.3n azaldu da. Taldean metodo estokastikoen desanbiguazioa ere aztertu da, baina honen emaitzak ez dira izan ezagutza linguistikoan oinarritutakoaren metodoarenak bezain onak. Horregatik desanbiguaziorako moduluan lehenengo murriztapen-gramatika aplikatuko da, eta ondoren berak utzitako anbiguotasunak ebazteko tratamendu estatistikoa, hitz bakoitzeko interpretazio bakarra geratzeko.

Puntu honi bukaera emateko, esan dezakegu egindako hitzaren gramatika zein murriztapen-gramatika funtsezko tresnak izango direla euskararen tratamendurako. Hitzaren gramatika hitz-mailan oinarritzen diren moduluetan erabiliko da, bai barne-funtzionamendurako bai kanporako emaitzak aurkezteko orduan. Murriztapen-gramatika euskara bezalako hizkuntzetan desanbiguaziorako ezinbesteko tresna izango da. Horregatik, tresna horiek funtsezkoak dira lematizatzaile/etiketatzaile batean.

VI.2 Ikasleen testuen egitura sintaktiko orokorrak aztertzeko tresna

Egindako analizatzaile sintaktikoa esaldi bateko elementu sintaktiko nagusiak analizatzeko gai denez, baliagarria izan daiteke testuen zenbait ezaugarriren azterketarako. Puntu honetan aipatuko duguna bigarren hizkuntza baten ikaskuntza-prozesua aztertzeko sistema batean erabili da. Sistema horretan euskara ikasten ari diren ikasleen testuak aztertu nahi dira, Maritxalar-en tesian (1999). Tesi horretan morfologiaren inguruan egin da lan gehiena, baina hitz barruko erroreen eta akatsen deskribapenaz aparte, maila desberdinetako ikasleen testuen egitura sintaktikoen maiztasunak atera nahi izan dira, hipotesi nagusia maila ezberdinetako ikasleek egitura sintaktiko ezberdinak erabiltzen dituztela zela. Hori egiaztatzeko, tresna linguistiko eta informatikoa garatu zen, lehen aipatutako oinarrizko analizatzaileak abiapuntutzat hartuz.

Adibidez, hasierako esaldia hau bada:

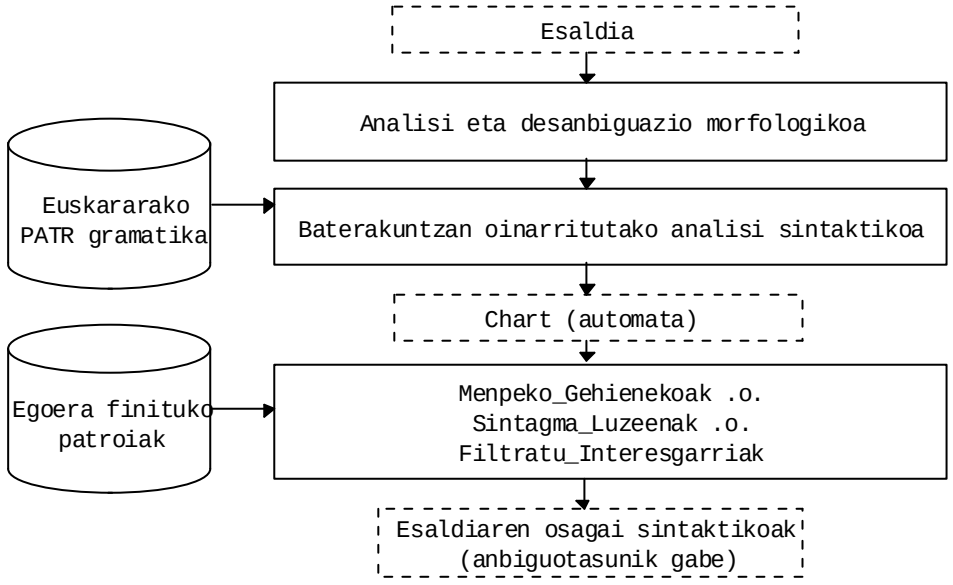
> Horregatik, susmatzekoa da honekin batera, garraioak ere igo egingo direla, bidaiatzea gero eta garestigoa izan dadila zeren eta dibisak erosteko pezeta gehiago ordaindu beharko da erebai.

Emaitzan esaldi horretan zenbat menpeko dauden (menpeko bakoitzaren mota markatuz), zenbat esaldi arrunt, eta beraiek lotzeko nolako juntagailuak dauden jakin nahi da. VI.2 irudian deskribatzen da jarraitutako prozedura. Lehen pauso batean esaldiaren osagai sintaktikoak lortzen dira baterakuntzan oinarritutako gramatikaren bidez. Ondoren egoera finituko gramatika aplikatuko zaio testuari emaitza ateratzeko.

?? Lehenengo menpeko gehien dauzkaten interpretazioak hartuko dira. Pauso honek baztertu egingo ditu interpretazio anbiguo batzuk, adibidez '*jatea eta edatea*' esaldian, *-tea* bukaerako bi elementuek *-tea*-menpekoa/izena-abs (eratorpenaren bidez) anbiguotasuna dute, eta horregatik lau interpretazio posible ditugu:

jatea	*eta*	*edatea*
menpekoa		menpekoa
izena-abs		izena-abs

Horietatik, biak menpekotzat hartzen dituen interpretazioa hobesten da.



VI.2 irudia. Esaldi baten osagai sintaktikoak ateratzeko tresna.

?? Bigarren pausoan sintagma luzeenak hartzen saiatuko da, hau da, *politak* eta '*etxe politak*' bi izen-sintagmen artean bigarrena hartuko dugu. Honek aukera asko baztertuko ditu.

?? Hirugarren pauso batean esaldiaren interpretazio guztietatik informazio interesgarria iragazten da, honela interpretazio bakarra utziz kasu guztietan. Adibidez, esaldi batek mota honetako lau interpretazio izan ditzake
esaldi sinplea + *juntagailua* + *esaldi sinplea*. Interesatzen dena patroi orokorra denez, honek aukera asko baztertuko ditu.

Aurreko esalditik lortutako emaitza VI.1 taulan agertzen den modukoa izango da. Informazio hori atera ondoren maila desberdinetako ikasleen esaldien konplexutasuna aztertu da, ondorio nagusiak maila baxuagoko ikasleek testu bera idazteko maila altuagokoek baino esaldi gutxiago erabiltzen dituztela, eta gainera esaldi horiek lotzeko mekanismo sintaktikoak pobreagoak direla, gehienetan koordinazioa erabiliz. Maila altukoek, aldiz, esaldi luzeagoak eta menpekoen erabilera aberatsagoa egiten dute. Adibidez, maila baxuko ikasleen esaldien laurdena (%25,25) *esaldi sinplea + eta + esaldi sinplea* patroikoak dira, eta maila altukoetan esaldi horiek %5,13a besterik ez du suposatzen. Beste alde batetik, menpekorik gabeko esaldien portzentajea handiagoa da behe mailako ikasleen testuetan (%32,35 versus %69,7).

Esaldia	**Emaitza**
Horregatik,	
susmatzekoa	**Menpekoa(tzeko)**
da	**Esaldi sinplea**
honekin batera,	
garraioak ere igo egingo direla,	**Menpekoa(ela)**
bidaiatzea	**Menpekoa(tzea)**
gero eta garestigoa izan dadila	**Menpekoa(ela)**
zeren eta dibisak erosteko	**Menpekoa(tzeko)**
pezeta gehiago ordaindu beharko da erebai	**Esaldi sinplea**

VI.1 taula. Testuen azterketarako tresnak lortutako emaitzaren adibidea.

VI.3 Ondorioak

Kapitulu labur honetan ikusi dugu egindako tresnak, analizatzaile morfosintaktikoa zein analizatzaile sintaktikoak, erabilgarri direla aplikazio-multzo zabal baterako. Analisi morfosintaktikoa lematizatzaile/etiketatzaile batentzako oinarrizko baliabidea da, lema eta morfemen informazioetatik abiatuta hitz osoaren informazioa lortzeko. Analizatzaile sintaktiko partziala tresna egokia da ikasleen testuetako egitura sintaktikoen erabilera aztertzeko, baita lengoaien irakaskuntzarekin lotutako beste hainbat aplikaziotarako ere.

AURRERA BEGIRAKOAK ETA ONDORIOAK

VII Tesiaren ondorio nagusiak eta etorkizunerako ikerlerroak

VII.1 Lortutako emaitzak

Tesi honetan euskararen sintaxi konputazionalaren azterketa egin da, testu errealen tratamendurako baliabide linguistiko orokorrak garatu dira eta zenbait aplikazio inplementatu dira. Sintaxia oraingoz eremu zabalegia dela kontuan hartuta, lan honetan bere tratamenduaren aspektu batzuk baino ez dira landu, zenbait irizpideren arabera aukeratuta. Lana
ikerketa-talde baten barruan kokatzen denez, sintaxiaren deskribapenerako beste
ikerketa-lanen testuinguruan kontsideratu behar da.

Hasteko, hitzaren barruko informazioa osatzeko hitzaren gramatika morfosintaktiko osoa garatu eta inplementatu da (II. kapitulua). Lehenago eginda zegoen hitzaren segmentazioaren bi mailatako deskribapenari gramatika hau gehituz euskararen morfologiaren tratamendu konputazional osoa biribildu egin da, corpusetako edozein hitzen analisia egiteko tresna lortuz. Euskararen morfologia konplexua dela eta, analizatzaile morfologiko oso hau sintaxiaren tratamenduaren hasiera da, eta hitza edo osagai sintaktiko handiagoak unitatetzat hartzen dituen edozein prozesamendu motatan erabili beharreko tresna izango da. Beste alde batetik, analizatzaile morfologiko osoaren garapenak erakutsi digu formalismo ezberdinen konbinazioak abantailak dituela, formalismo bakoitzaren alde onenak har daitezkeelako. Morfosintaxiaren kasuan bi mailatako formalismoa eta baterakuntza-gramatikak integratu dira osagarriak diren moduan, bukaeran sistema osoaren ahalmen deskriptiboa handituz eta testu errealak tratatzeko eraginkortasuna lortuz (Aduriz *et al*. 2000ab).

Morfosintaxiaren tratamenduaren ondoren sintaxiaren mundu zabalagoan sartu gara III. kapituluan. Hasteko, euskararen estaldura ertaineko baterakuntza-gramatika garatu eta dagokion analizatzailea inplementatu da. Gramatika partziala da, baina corpusetan agertzen diren osagai sintaktiko nagusiak deskribatzen ditu (horien artean izen-sintagmak, adizlagunak, esaldi sinpleak eta menpeko esaldiak), eta horrela esaldien analisi sakon eta osorako lehen pausoa eman da. Analisi osoari ez ekiteko bi arrazoi aipatu dira. Batetik, EDBLren erabilerak sendotasuna eta estaldura lexikal zabala eskaintzen duen arren, oraingoz informazio sintaktiko garrantzitsuen gabezia ere kontuan hartu behar izan dugula (aditzen azpikategorizazioari buruzkoa, adibidez). Bestetik, sintaxiaren eremu zabalak tesi honetan tratatzeko osagaiak mugatzera eraman gaituela. Une honetan, analizatzailea gai da corpus handiak modu eraginkorrean tratatzeko eta corpus horietatik osagai sintaktiko nagusiak ateratzeko.

Egoera finituko sintaxiaren bidetik, euskararen murriztapen-gramatikaren garapenean lan egin dugu, testuetako hitzen desanbiguazioa lortzeko. Gramatikaren garapena lankidetzan egin da (Aduriz 2000, Arriola 2000). Gure

ekarpen nagusia tratamendu informatikoan izan da. Beste hizkuntzetan lortutako emaitzekin konparagarria den desanbiguazio-tasa lortu dugu, euskara bezalako hizkuntzetarako formalismoaren egokitasuna frogatuz. Tresna erabilgarria lortu da, une honetan zenbait aplikaziotan erabiltzen ari dena. Tresnaren erabilera nagusia testuetako anbiguotasunaren jaitsieran egin dugu, ondorengo prozesuetarako lana aurreratuz.

Murriztapen-gramatikaren mugak kontuan hartuta, egoera finituko beste formalismo orokorrago baten egokitasuna aztertu dugu. Formalismo honek adierazpen erregularren bidezko patroietan oinarritutako sintaxia garatzea ahalbidetzen du, patroiekin anbiguotasuna kentzeko murriztapenak, informazioaren iragazleak edo osagai sintaktiko berrien sorkuntza adierazteko. Egin ditugun zenbait lanetan formalismo honen baliagarritasuna frogatu dugu.

Sintaxiaren tratamendurako hiru hurbilpen horien azterketak aukera eman digu bakoitzaren alde onak eta arazoak esperimentatzeko. Horregatik hirurak konbinatzen dituen sistema integratua garatu dugu, bakoitzaren alde positiboak biltzeko asmoz. Hain zuzen ere, hiru tresnen aplikazio sekuentziala probatu eta inplementatu dugu. Murriztapen-gramatika desanbiguazio morfosintaktikorako erabiliko da lehen pauso batean, emaitza desanbiguatuaren gainean baterakuntzan oinarritutako gramatikaren bidez osagai sintaktiko posibleak eraikitzeko. Osagai sintaktikoen anbiguotasuna eta aplikazioen behar desberdinak tratatzeko egoera finituko tresnaren bidezko patroiak defini daitezke, bukaeran corpusen azterketarako tresna malgua eta sendoa lortuz.

Baliabide sintaktiko orokor horien garapena egin ondoren, beraien gaineko aplikazioak landu dira. IV. kapituluan ezagumendu lexikal eta sintaktikoaren aberasketan egindako esperimentuak azaldu dira, corpusetatik aditzaren azpikategorizazio-informazioaren erauzketan. Modu horretan, hasierako baliabide sintaktikoak osatu egin ahal izango dira, era iteratiboan: oinarrizko gramatikak erabiliz informazio linguistikoa atera daiteke gramatika horiek aberasteko, ondoren aberasketa-prozesua behin eta berriro errepikatzeko. Esperimentu horietan azaldu dugunez, emaitzak erabilgarriak dira, estaldura eta doitasun altuekin, eta beraien gainean aditzei buruzko informazioa eskuz edo modu automatikoen bidez atera ahal izango dugu. Beste alde batetik, esperimentu hauek oinarrizko baliabide sintaktikoen baliagarritasuna frogatzeko ere balio izan dute.

Garatu diren tresna sintaktikoen erabilera errore ortografiko eta sintaktikoen detekzio eta zuzenketan ere esperimentatu dugu V. kapituluan. Lehenengo, euskarazko testuetako errore sintaktikoen sailkapena egin da, beste hizkuntzetan gertatzen direnekin konparatuz. Bigarren, baterakuntzan oinarritutako gramatikaren gaineko murriztapen sintaktikoen erlaxazio graduala probatu da komunztadurak eta antzeko erroreak detektatzeko, metodoaren bideragarritasuna frogatuz. Ondoren, errore-patroietan oinarritutako detekzioa probatu dugu, corpusetan aurkitutako daten adierazpenen erroreen gainean, baterakuntzan oinarritutako analizatzaile sintaktikoa eta egoera finituko patroiak konbinatuz. Bukatzeko, zuzenketa landu da, errore ortografikoen proposamen bakarra modu automatikoan lortzeko, eta zenbait ezagumendu motaren (sintaktikoa eta semantikoa) ekarpena neurtu da.

Erroreei buruzko kapitulu hau laburtzekotan, esan behar dugu frogatu dugula ezagutza-iturri ezberdinen ekarpenak gehituz gero LNParen edozein aplikazioren emaitzatan islatuko dela. Konbinazioak proposamen ezberdinen indarrak biltzeko ahalmena duela ikusi dugu.

Azkenik, VI. kapituluak beste bi aplikazio erakusten ditu, baliabide sintaktikoen malgutasun eta erabilpenaren adibidea emateko. Analisi morfosintaktikoa lematizatzaile/etiketatzaile batentzako oinarrizko baliabidea da, lema eta morfemen informazioetatik abiatuta hitz osoaren informazioa lortzeko. Bestalde, analizatzaile sintaktiko partziala tresna egokia da ikasleen testuetako egitura sintaktikoen erabilera aztertzeko.

VII.2 Zabaldutako ikerlerroak eta perspektibak

Tesi honetako ekarpen nagusiak aztertu eta gero, esan behar dugu sintaxiaren tratamenduaren azterketak bukaera argi eta zehatz batera eraman baino, ikertu gabeko bide berri zabal eta interesgarriak ireki dizkigula. Ondorengo lerroetan, tesiaren lana jarraitzeko aurreikusten ditugun ikerlerro nagusiak emango ditugu.

VII.2.1 Tratamendu morfosintaktikoaren jarraipena

Nahiz eta morfologiaren tratamendua ia bukatutzat eman, oraindik lan egiteko eremuak ikusten ditugu ondoko aspektuetan:

?? Gramatikaren formalizazio linguistikoa. PATR formalismoa erabili dugu deskribapen morfosintaktikorako, horren alde malgutasuna aipatzen genuela beste formalismo konplexuagoak ez erabiltzeko. Baina honek daukan beste ondorio bat erregelen idazketan egin beharreko lana da, printzipio orokorrik definitu gabe ezaugarri guztien tratamendua baterakuntza-ekuazioekin egin behar delako. Printzipioen erabilerak gramatikaren trinkotasuna eta sinpletasuna ekarriko duenez, hasi dugun bide hau jarraitzea komenigarri ikusten dugu. Beste alde batetik, ezaugarri-egitura motadunak erabiltzeak gramatikaren koherentzia eta ziurtasuna gehituko du.

?? Eraginkortasunaren hobetzea. Inplementatutako sistemaren abiadura nahikoa da orain taldearen beharretarako, baina gero eta corpus handiagoen erabilerak etorkizunerako eskakizunak gogortu egingo ditu. Lehenago esan dugunez, egin daitezkeen hobekuntza dezente daude, abiadura magnitude-ordena bat baino gehiago igotzeko. Bestalde, egoera finituko soluzio berrien ekarpena ere kontuan izan beharko da, une honetan ezinezkoak diren tratamenduetarako proposamenak egiten badira.

?? Datu-base lexikaleko informazioak aldatzen edo gehitzen direnean gramatika morfosintaktikoa aldatu edo zabaldu egin beharko da. Adibidez, hitz anitzeko unitateen tratamendua lexikoan sartzeko asmoa dagoenez, horientzako erregelak definituko dira, edo azpikategorizazioa bezalako informazioak lexikoan sartzen diren heinean, berari dagozkion tratamendu morfosintaktikoen definizioak ere gehitu egin beharko dira.

VII.2.2 Tratamendu sintaktikoaren jarraipena

Sintaxiaren tratamendurako esperimentatzeko aukera anitz ikusten dugun arren, hurrengo puntuetan lehentasun handienekoak aipatuko ditugu:

?? Aberasketa lexikal eta sintaktikoa. Lortutako tresnen bidez ikusi dugu corpusen azterketa egin daitekeela, doitasun eta estaldura altuekin. Horregatik, corpusetan kodetuta dagoen ezagumenduaren erauzketarako bidea irekita dago. Hau frogatu egin dugu azpikategorizazio-informazioa ateratzeko prozesuan, baina esan dugunez esperimentu horiek lehen pausoak besterik ez dira, eta egin beharreko lan hauek definitu ditugu:

?? Aditz eta osagarri posibleen adibideak lortu eta gero, azpikategorizazio-patroien definizioa egin daiteke, metodo estatistiko edo erdiautomatikoak erabiliz. Tresna automatikoak direla eta, patroi horiei agerpen-maiztasunak gehi dakizkieke, analizatzaile estatistikoen bidea jorratuz.

?? Ideia beretik abiatuta, egitura sintaktikoen maiztasunak ere neur daitezke, horrela gramatikako erregelen aplikazioa corpusetako maiztasunen arabera baldintzatzeko.

- Sintaxitik harantzago joanez, osagai lexikalen arteko erlazioak ere azter litezke semantikaren mundua ukituz, kokakidetzak edo egitura sintaktiko batzuen bidez definitutako erlazio semantikoak ateratzeko (Berland eta Charniak 1999).

- Baterakuntzan oinarritutako gramatikaren zabalpena. Informazio lexikala aberasten joango denez, egingarria izango da euskararen LFG edo HPSG moduko formalismo ahaltsuagoen bidezko deskribapena. Lan hau moldaketa sinple bat baino askoz aldaketa handiagoa izango da, formalizazio linguistiko zehatza egitea eskatuko duelako, bai une honetan gramatikan tratatzen diren fenomenoena bai gramatikatik kanpo utzi diren guztiena (horien artean esaldi nagusi eta menpekoen arteko erlazioak).

- Mendekotasun-egituren bidezko analisia esaldi osoen tratamendurako. Hau aurreko puntuaren alternatiba izan liteke, baterakuntza-gramatika dagoen bezala utzi eta sortutako osagai sintaktikoen konbinazioak aztertuz, esaldi osoen analisiak zehazteko. Era honetan zenbait lanetan (Järvinen eta Tapanainen 1998, Oflazer *et al*. 1999ab, Basili *et al*. 2000) hasitako pausoak jarraituko genituzke.

- Sintaktikoki etiketatutako corpusen garapena. Pauso hau beharrezkoa da ebaluazioa edo estatistikan oinarritutako analisiak egiteko. Corpus hauek lortzeko egin behar den lan handia kontuan hartuz gero, aztertzeko bide posible bat garatu diren tresna automatikoen laguntza erabiltzea izan daiteke.

- TEI estandarren definizioak tresna desberdinen arteko eta kanpoko komunikaziorako (Artola *et al*. 2000). Tresna linguistikoen emaitzak aplikazio desberdinek erabiltzeko nahitaezkoa da emaitza horiek definizio formal bat izatea, horrela kodeketa-modu partikularren arazoa gainditzeko. Honengatik baliabide linguistikoen sarrera/irteerak TEI gidalerroak jarraituz kodetzeko lanean ari gara une honetan.

VII.2.3 Erroreen tratamendurako lanen jarraipena

Erroreen tratamenduan egin ditugun lanak jarraitzeko honako bide hauek ditugu gogoan:

- Murriztapen sintaktikoen erlaxazioa testu errealen tratamendurako egokitu. Hau bideragarria izateko, patroietan oinarritutako sintaxia eta erlaxazioa konbinatu beharko dira: patroiek detektatu beharko dituzte errorea eman dezaketen esaldien zatiak, eta ondoren baterakuntzan oinarritutako analizatzailea aplika dakioke zati horri. Era horretan esaldi bakoitzeko errore posible guztien hipotesiak aztertzea ekidin ahal izango da, eta testuinguruaren arabera posible diren erroreak bakarrik aztertuko dira.

- Errore-patroien bidezko tratamenduaren zabalpena. Datekin egin den moduan, testuinguru sintaktiko lokaleko azterketarekin detekta daitezkeen errore gehiago aztertu eta tratatzeko

asmoa dugu. Horien artean postposizioak edo azpikategorizazioarekin lotutako erroreak aipa daitezke. Azken hauen tratamendua azpikategorizazio-informazioaren erauzketan lortutako emaitzekin lotuta egongo da.

Esan behar da ere patroien bidezko erroreen tratamendua modu estuan lotuta dagoela erroreak markatuta dituzten corpusen garapenarekin. Une honetan hau da arazo handienetako bat euren tratamendurako.

Corpus horiek lortzen diren neurrian errore-patroi horientzako ikasketa-metodo automatikoen edo erroreen tratamendu estatistikoaren ekarpena probatu ahal izango da.

?? Euskarazko testuetako errore ortografikoen zuzenketa automatikoa. Ingelesarekin egindako esperimentuaren ondoren, euskararekin probak egitea izango da hurrengo urratsa. Ingeleserako baliabide ugari dagoela aipatu dugu zuzenketa automatikoaren proba egiteko hizkuntza aukeratu dugunean, baina euskararen baliabideen aberasketak aurrera jarraitzen badu, euskararen gaineko esperimentu berriak egiteko materiala izango da, horien artean MG eta baterakuntzan oinarritutako gramatikak edo baliabide semantikoak (Agirre 1999) daudela.

?? Tresna sintaktikoen integrazioa hizkuntzaren irakaskuntzarako sistema batean. Morfologiarekin gertatu den antzera (Maritxalar 1999), tresna sintaktikoak erabiliak izan daitezke euskararen irakaskuntzarako sistemetan. Aplikazio horietarako aukera ugari ikusten dugu: ikaslearen gaitasun linguistikoa neurtzeko, erroreen detekzio edo zuzenketarako edo ikasketa-prozesuen laguntza-moduan.

BIBLIOGRAFIA

Konferentzia nagusien erreferentziak egiteko laburdurak erabiliko dira:

- ACL: Association for Computational Linguistics.
- ANLP: Conference on Applied Natural Language Processing.
- COLING: International Conference on Computational Linguistics.
- EACL: European Association for Computational Linguistics.
- IWPT: International Workshop on Parsing Technologies.
- NAACL: North American Chapter of the Association for Computational Linguistics.
- SEPLN: Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural.

Abaitua, J. 1988. *Complex predicates in Basque: from lexical forms to functional structures*. Doktoretza-tesia, University of Manchester.

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X., Artola X., Arriola J.M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R. 1992. *Estudio comparativo de diferentes formalismos sintacticos para su aplicacion al euskara*. Barne-txostena, UPV/EHU/LSI.

Abney S. P. 1991. *Parsing by chunks.* R. C. Berwick, S. P. Abney, and C. Tenny, editoreak, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, Dordrecht.

Abney S. P. 1997. *Part-of-Speech Tagging and Partial Parsing*. S. Young eta G. Bloothooft, editoreak, Corpus-Based Methods in Language and Speech Processing, Kluwer, Dordrecht.

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. 1995. *Different Issues in the Design of a lemmatizer/Tagger for Basque*. From Texts to Tags: Issues in Multilingual Language Analysis. Association for Computational Linguistics SIGDAT Workshop, Dublin.

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. 1996a. *EUSLEM: A lemmatiser/tagger for Basque*. Euralex'96.

Aduriz I., Aldezabal I., Alegria I., Ezeiza N., Urizar R. 1996b. *Del analizador morfológico al etiquetador/lematizador: Unidades léxicas complejas y desambiguación*. SEPLN'96, Sevilla.

Aduriz I., Aldezabal I., Alegria I., Ezeiza N., Urizar R. 1996c. *MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque*. COMPLEX.

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997. *A spelling corrector for Basque based on morphology*. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. 1997. *Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism*. Conference on Recent Advances in Natural Language Processing, Bulgaria.

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

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., ArtolaX., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. 1998b. *A Framework for the Automatic Processing of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada.

Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., ArtolaX., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 1999. *MORFEUS: Euskararako analizatzaile morfosintaktikoa*. Barne-txostena, UPV/EHU/LSI/TR 1-99.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 2000a. *A Word-Grammar Based Morphological Analyzer for Agglutinative Languages*. COLING'2000, Saarbrücken.

Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 2000b. *A Word-Level Morphosyntactic Analyzer for Basque*. Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000), Atenas.

Aduriz I. 2000. *Morfologiatik Sintaxira Murriztapen Gramatika baliatuz*. Tesi-txostena, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M. *1992 XUXEN: A spelling Checker/Corrector for Basque Based on Two-Level Morphology*. ANLP'92, Trento.

Agirre E., Arregi X., Arriola J. M., Artola X., Insausti J. M. 1994. *Euskararen Datu-Base Lexikala (EDBL)*. Barne-txostena UPV/EHU/LSI/TR 8-94.

Agirre, E., Arregi, X., Artola, X., Díaz, A., Sarasola, K. 1994. *Lexical-semantic information and the automatic correction of spelling errors*. Proceedings of the Workshop on Semantics and Pragmatics of Natural Language: Logical and Computational Aspects, Sara, France.

Agirre E., Rigau G. 1996. *Word sense disambiguation using conceptual density*. COLING'96, Copenhagen.

Agirre E., Gojenola K., Sarasola K., Voutilainen A. 1998a. *Towards a Single Proposal in Spelling Correction*. COLING-ACL'98, Montreal.

Agirre E., Gojenola K., Sarasola K., Voutilainen A. 1998b. *Towards a Single Proposal in Spelling Correctio*n. UPV-EHU / LSI / TR 8-98.

Agirre E. 1999. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Aho SA., Sethi R., Ullman J. 1985. *Compilers: Principles, Techniques and Tools*. Addison-Wesley.

Aït-Mokhtar S., Chanod J-P. 1997. *Incremental Finite-State Parsing*. ANLP'97.

Aït-Mokhtar S., Chanod J-P. 1997. *Subject and Object Dependency Extraction Using Finite-State Transducers*. ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications, Madrid.

Aldezabal I., Alegria I., Artola X., Díaz de Illarraza A., Ezeiza N., Gojenola K., Aduriz I., Urkia M. 1994. *EUSLEM: Un lematizador/etiquetador de textos en euskara*. SEPLN'94, Córdoba.

Aldezabal I., Goenaga P., Gojenola K., Sarasola K. 1998. *Subcategorización verbal vasca: propuesta inicial y herramienta de validación*. SEPLN'98, Alicante.

Aldezabal I., Alegria I., Ansa O., Arriola J.M., Ezeiza N. 1999a. *Designing spelling correctors for inflected languages using lexical transducers*. Proceedings of EACL'99, 265-266. Bergen, Norway.

Aldezabal I., Ansa O., ArtolaX., Ezeiza A., Gojenola K., Insausti J.M., Lersundi M. 1999b. *Euskararen Datu-Base Lexikala (EDBL): eskema berriaren proposamena*. Barne-txostena, UPV/EHU/LSI/TR 9-99.

Aldezabal I., Gojenola K., Oronoz M. 1999c. *Combining Chart-Parsing and Finite State Parsing*. Proceedings of the Student Session of the European Summer School in Logic, Language and Computation (ESSLLI'99), Utrecht.

Aldezabal I., Gojenola K., Sarasola K. 2000. *A Bootstrapping Approach to Parser Development*. IWPT'2000, Trento.

Aldezabal I. 2000. *Euskal aditzaren azpikategorizazioa. Azterketa sistematiko-automatikoa*. Tesi-txostena, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

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

Alegria I., Artola X., Sarasola K. 1995 *Improving a robust morphological analyzer using lexical transducers.* Conference on Recent Advances in Natural Language Processing, Bulgaria.

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

Alegria I., Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Urkia M. 1996b. *A Corpus-Based Morphological Disambiguation Tool for Basque*. SEPLN'96, Sevilla.

Allen J., Hunnicut M., Klatt D. 1987. *From text to speech: the MITalk System*. Cambridge University Press.

Allen J. 1995. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company.

Alshawi H. (editorea) 1992. *The Core Language Engine*. Cambridge, MA: MIT Press.

Alshawi H., Moore R. C. 1992. *Introduction to the CLE*. The Core Language Engine, Alshawi, H. editorea, MIT Press.

Antworth E. L. 1994. *Morphological Parsing with a Unification-based Word Grammar*. North Texas Natural Language Processing Workshop, Texas.

Arregi X. 1995. *ANHITZ: Itzulpenean laguntzeko hiztegi-sistema eleanitza*. Tesi-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Arriola J.M., Soroa A. 1996. *Lexical Information Extraction for Basque*. Student Conference in Computational Linguistics, Canada.

Arriola J. M., Artola X., Gojenola K., Soroa A. 1997. *TEI: testu-kodeketarako gidalerroak*. Ekaia, Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria 7.

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

Arriola J.M. 2000. *Hauta-Lanerako Euskal Hiztegi-ko informazio lexikalaren erauzketa erdi-automatikoa eta bere integrazioa sistema konputazional batean*. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Artola X. 1993. *Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena/Conception et construction d'un systeme intelligent d'aide dictionnariale (SIAD)*. Tesi-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A. 2000. *Integration of NLP Tools using SGML-tagged texts.* Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000), Atenas.

Atserias J., Carmona J., Castellón I., Cervell S., Civit M., Màrquez L., Martí M.A., Padró L., Placer R., Rodríguez H., Taulé M., Turmo J. 1998. *Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text*. First International Conference on Language Resources and Evaluation (LREC'98), Granada.

Atwell E. 1987. *Constituent-likelihood grammar*. The Computational Analysis of English, Garside R., Leech G., Sampson G. liburuan.

Atwell E., Elliott S. 1987. *Dealing with Ill-Formed English Text*. In The Computational Analysis of English: a Corpus-Based Approach, Ed. Longman.

Badia T., Egea A., Tuells A. 1996. *SEGMORF: un formalismo para analizadores morfológicos de dos niveles*. SEPLN'96, Sevilla.

Badia T. 1997. *Especificaciones lingüísticas para gramáticas en estructuras de rasgos tipificadas*. Philologia Hispalensis, Vol. XI, fascículo 2, Facultad de Filologia, Universidad de Sevilla.

Basili, R., Pazienza M.T., Zanzotto F.M. 1998. *Efficient Parsing for Information Extraction*. Proceedings of the 13th European Conference on Artificial Intelligence, John Wiley & Sons Ltd.

Basili, R., Pazienza M.T., Zanzotto F.M. 2000. *Customizable Modular Lexicalized Parsing*. IWPT'2000, Trento.

Bates, M. 1978. *The theory and practice of ATN grammars*. L. Bolc (ed.) 'Natural Language Communication with Computers' liburuan, Springer Verlag, Berlin.

Bear J. 1986. *A Morphological Recognyzer with Syntactic and Phonological Rules*. COLING'86.

Bear J. 1988. *Morphology with Two-Level Rules and Negative Rule Features*. COLING'88.

Beesley K. 1998a. *Constraining Separate Morphotactic Dependencies in Finite-State Grammars*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.

Beesley K. 1998b. *Arabic Morphological Analysis on the Internet*. Proceedings of the International Conference on Multi-Lingual Computing (Arabic & English), Cambridge.

Berland M., Charniak E. 1999. *Finding Parts in Very Large Corpora*. ACL'99, Maryland.

Berwick R., Abney S.P., Tenny C. 1991. *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht: Kluwer Academic.

Bod R. 1993. *Using an annotated corpus as a stochastic grammar*. EACL'93, Utrecht.

Bod R., Kaplan R. 1998. *A Probabilistic Corpus-Driven Model for Lexical-Fujnctional Analysis*. COLING-ACL'98, Montreal.

Borsley R., Przepiórkowski A. 1999. *Slavic in Head-Driven Phrase Structure Grammar*. CSLI publications, Stanford.

Brent M. 1993. *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax*. Computational Linguistics, vol. 19(2).

Bresnan J., (editorea) 1982. *The Mental Representaion of Grammatical Relations*. The MIT Press, Cambridge, Massachusetts.

Brill E. 1995. *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging*. Computational Linguistics.

Brill E., Florian R., Henderson J.C., Mangu L. 1998. *Beyond N-Grams: Can Linguistics Sophistication Improve Language Modeling?*. COLING-ACL'98, Montreal.

Brill E., Wu J. 1998. *Classifier Combination for Improved Lexical Disambiguation*. COLING-ACL'98, Montreal.

Brill E., Henderson J., Ngai G. 2000. *Automatic Grammar Induction: Combining, Reducing and Doing Nothing*. IWPT'2000, Trento.

Briscoe T., Carroll J. 1993. *Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars*. Computational Linguistics, vol. 19(1).

Briscoe T. 1994. *Parsing (with) Punctuation etc.* Barne-txostena MLTT-002, Xerox Research Center.

Briscoe T., Carroll J. 1997. *Automatic Extraction of Subcategorization from Corpora*. ANLP'97, Washington.

Brun C. 1998. *Terminology Finite-State Preprocessing for Computational LFG*. COLING-ACL'98, Montreal.

Butt M., Fortmann C., Rohrer. C. 1996. *Syntactic Analyses for Parallel Grammars: Auxiliaries and Genitive NPs*. COLING'96, Copenhagen.

Butt M., Geuder W. 1998. *The projection of arguments - lexical and compositional factors*. CSLI publications, Stanford.

Butt M., King T.H., Nino M.E., Segond F. 1999 *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Lecture Notes, CSLI Publications.

Carbonell J.G., Hayes P.J. 1983. *Recovery strategies for parsing extragrammatical input*. American Journal of Computational Linguistics 9.

Carpenter B., Penn G. 1993. *ALE User's Guide Version 1.0*. Laboratory for Computational Linguistics Technical Report, Carnegie Mellon University, Pittsburgh.

Carroll J. 1993. *Practical unification-based parsing of natural language*. Computer Laboratory, Cambridge University, UK, PhD. thesis, Technical Report 314.

Carroll G., Rooth M. 1998. *Valence Induction with a Head-Lexicalized PCFG*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Granada.

Carroll J., Minen G., Briscoe T. 1998a. *Can Subcategorisation Probabilities Help a Statistical Parser?* Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora, Montreal.

Carroll J, Briscoe T., Sanfilippo A. 1998b. *Parser Evaluation: a Survey and a New Proposal*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.

Carroll J, Minnen G., Briscoe T. 1999. *Corpus Annotation for Parser Evaluation*. Proceedings of Workshop on Linguistically Interpreted Corpora, EACL'99, Bergen.

Carulla M., Oosterhoff A. 1996. *El Tratamiento de la Morfología Flexiva del Castellano mediante reglas de dos niveles en una gramática de unificación*. SEPLN'96, Sevilla.

Chanod J.P., Tapanainen P. 1995. *Tagging French - Comparing a Statistical and a Constraint-Based Method*. EACL'95, Dublin.

Chanod J.P., Tapanainen P. 1996a. *A Non-deterministic Tokeniser for Finite-State Parsing*. ECAI '96 workshop on Extended finite state models of language. Budapest.

Chanod J.P., Tapanainen P. 1996b. *A Robust Finite-State Grammar for French*. ESSLLI'96 Workshop on Robust Parsing, Prague.

Charniak E. 1993. *Statistical Language Learning*. Cambridge, MA: MIT Press.

Chomsky N. 1981. *Lectures on Government and Binding*. Dordrecht.

Church K. 1988. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*. ANLP'88.

Ciravegna F., Lavelli A. 1997. *Controlling Bottom-Up Chart Parsers through Text Chunking*. IWPT'97, Boston.

Collins M. 1997. *Three New Probabilistic Models for Statistical Parsing*. ACL'97, Madrid.

Collins M., Hajic J., Ramshaw L., Tillmann C. 1999. *A Statistical Parser for Czech*. ACL'99, Maryland.

Damerau F.J., Mays E. 1989. *An examination of undetected typing errors*. Information Processing and Management 25, 6.

Díaz de Ilarraza A., Maritxalar M., Oronoz M. 1997. *An implemented interlanguage model for learners of Basque*. Language Teaching and Language Technology, Swets and Zeitlinger Publisher.

Díaz de Ilarraza A. 1990. *Diseño de un Módulo de Interacción Tutor-alumno para un sistema inteligente de enseñanza de la programación*. Tesi-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Doran C., Egedi D., Hockey B. A., Srinivas B., Zaidel M. 1994. *XTAG system - A Wide Coverage Grammar for English*. COLING'94, Kyoto.

Douglas, S. 1991. *A Review and Bibliography of Approaches to Syntactic Error Correction and Robust Parsing*. The Editor's Assistant Project No. IED4/1/1679, Deliverable 4.3.1.

Douglas, S., Dale R. 1991. *Towards a Taxonomy of Errors in Technical Texts.* The Editor's Assistant Project No. IED4/1/1679, Deliverable 3.4.

Douglas, S. 1992. *Customising grammar and style checker rules*. *Intelligent Tutoring Media*, Vol. 3

Douglas, S., Dale R. 1992. *Towards Robust PATR.* COLING'92, Nantes.

Egunkaria 1992. *Estilo liburua*. Egunkaria bilduma.

Euskaltzaindia 1985. Euskal Gramatika Lehen Urratsak-I, Iruñea.

Euskaltzaindia 1994. *Euskal gramatika laburra: perpaus bakuna*. Euskaltzaindia gramatika batzordea.

Euskaltzaindia. *Euskaltzaindiaren arauak. Datak nola adierazi*. 37. unitatea, Bilbo.

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

Ezeiza N., Alegria I., Arriola J.M., Urizar R., Aduriz I. 1998. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98, Montreal.

Ezeiza N. 2000. *Corpusak ustiatzeko tresna linguistikoak/ Herramientas lingüísticas para la explotación de corpus*. Tesi-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Fass D., Wilks Y. 1983. *Preference semantics, ill-formedness, and metaphor*. American Journal of Computational Linguistics 9.

Francis S., Kucera H. 1967. *Computing Analysis of Present-Day American English*. Brown University Press.

Frank A., King T.H., Kuhn J., Maxwell. J. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Proceedings of the LFG98 Conference, Brisbane, Australia. CSLI Online Publications.

Gala N. 1999. *Using the Incremental Finite-State Architecture to create a Spanish Shallow Parser*. SEPLN'99, Lleida.

Gale W.A., Church K.W. 1990. *Estimation procedures for language context: Poor estimates are worse than none.* Proceedings of Compstat-90, Springer-Verlag, New York.

Garside R., Leech G., Sampson G. 1987. *The Computational Analysis of English*. Longman.

Gazdar G., Klein E., Pullum G., Sag I. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Genthial D. 1991. *Contribution à la construction d'un système robuste d'analyse du français*. Thèse de l'université Joseph Fourier, Grenoble I.

Genthial D., Courtin J., Menezo J. 1994. *Towards a More User-Friendly Correction*. COLING'94, Kyoto.

Giguet E., Vergne J. 1997. *From Part of Speech Tagging to Memory-based Deep Syntactic Analysis*. IWPT'97, Boston.

Goenaga P. 1980. *Gramatika bideetan*. Erein

Gojenola K., Sarasola K. 1994. *Aplicaciones de la relajación gradual de restricciones para la detección y corrección de errores sintácticos.* SEPLN'94, Córdoba.

Gojenola K. 1998. *Guneak zuzendutako egitura sintagmatikoen gramatika (HPSG) eta euskararako aplikazioa*. Barne-txostena, UPV/EHU/LSI/TR5-98.

Gojenola K., Oronoz M. 2000. *Corpus-Based Syntactic Error Detection Using Syntactic Patterns*. NAACL/ANLP 2000 Student Research Workshop, Seattle.

Golding A. 1995. *A Bayesian hybrid method for context-sensitive spelling correction*. Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA.

Golding A., Roth. D. 1996. *A Winnow-based Approach to Spelling Correction*. Proceedings of the 13th International Conference on Machine Learning, ICML'96.

Golding A., Schabes. Y. 1996. *Combining trigram-based and feature-based methods for context-sensitive spelling correction*. ACL'96, Santa Cruz, CA.

Gómez Guinovart J. 1996. *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*. Tese de doutoramento, Departamento de Filoloxía Española, Teoría da Literatura e Lingüística Xeral, Universidade de Santiago de Compostela.

Gómez Guinovart J. 1999. *La escritura asistida por ordenador. Problemas de sintaxis y de estilo*. Servicio de publicacións, Universidade de Vigo.

Grefenstette G. 1996. *Light Parsing as Finite-State Filtering*. ECAI'96 workshop on Extended finite state models of language. Budapest.

Grimshaw J. 1990. *Argument Structure*. MIT Press: Cambridge.

Grinberg D., Lafferty J., Sleator D. 1995. *A robust parsing algorithm for link grammars*. IWPT'95, Prague.

Grishman R., Macleod C., Meyers A. 1994. *Comlex Syntax: Building a Computational Lexicon*. COLING'94, Japan.

Gross M. 1997. *The Construction of Local Grammars*. Finite-State Language Processing liburuan, MIT Press.

Grover, C., Carroll J., Briscoe E. 1993. *The Alvey Natural Language Tools grammar (4th release).* Computer Laboratory, Cambridge University, UK, Technical Report 284.

Gunji T. 1987. *Japanese Phrase Structure Grammar*. Dordrecht: Reidel.

HABE 1985. *Akatsen bilduma analitikoa*. Zutabe Aldizkaria

Hajic J., Hladká B. 1998. *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset*. COLING-ACL'98, Montreal.

Heaton, J. B., Turton, N. D. 1987. *Longman Dictionary of Common Errors*. Longman.

Heidorn G. E., Jensen K., Miller L. A., Byrd R. J., Chodorow M. S. 1982. *The EPISTLE text-critiquing system*. IBM Systems Journal, Vol. 21, No. 3.

Hellwig P. 1998. *Natural Language Parsers: a "course in cooking"*. Tutorial Notes, COLING-ACL'98, Montreal.

Hermjakob U., Mooney R.J. 1996. *Learning Parse Decisions From Examples With Rich Context*. ACL'96.

Hindle D. 1989. *Acquiring disambiguation rules from texts.* ACL'89.

Hobbs J.R., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., Tyson M. 1997. *FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text*. Finite-State Language Processing liburuan, MIT Press.

Holan T., Kubon V., Plátek M. 1997. *A Protoype of a Grammar Checker for Czech*. ANLP'97, Washington.

Hudson R. 1990. *English Word Grammar*. Oxford: Basil Blackwell.

Ide N., Veronis J. K. 1995. *Text-Encoding Initiative, Background and Context*. Kluwer Academic Publishers.

Ide N. 1998. *Encoding Linguistic Corpora*. Sixth Workshop on Very Large Corpora, COLING-ACL'98, Montreal.

Ide N., Greenstein D., (Editors-in-Chief) 1999. *Tenth Anniversary of the Text Encoding Initiative.* Computers and the Humanities, Special Double Issue, Volume 33, Nos. 1-2.

In-Sig Y., Kwang-Moo C., Taisook H. 1993 *Syntactic error repair using repair patterns*. Information Processing and Management 47.

Järvinen T., Tapanainen P. 1998. *Towards an Implementable Dependency Grammar*. Proceedings of the Workshop on Processing of Dependency-Based Grammars, COLING-ACL'98, Montreal.

Jensen K. 1987a. *Binary rules and non-binary trees*. Mathematics of Language liburuan, A. Manaster-Ramer editorea, Amsterdam: John Benjamins.

Jensen K. 1987b. *Issues in Parsing*. Computer Science Research Report 13380 12/23/87, IBM Thomas J. Watson Research Center.

Jensen K. 1988. *Why Computational Grammarians Can Be Skeptical About Existing Linguistic Theories*. COLING'98, Budapest.

Jensen K., Heidorn G., Richardson S. 1993. *Natural Language Processing: the PLNLP Approach*. Kluwer Academic Publishers, Boston.

Joshi A. 1985. *Tree-adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions*. D. R. Dowty, L. Karttunen, A. Zwicky editoreak, Natural Language Parsing, New York: Cambridge University Press.

Kaplan R. Kay M. 1981. *Phonological rules an finite-state transducers*. Annual Meeting of the Linguistic Society of America, New York.

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

Karttunen L. 1986. *Radical Lexicalism*. Center for the Study of Language and Information (CSLI), 86-68, Stanford.

Karttunen L. 1995. *The Replace Operator*. ACL'95, Boston.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997. Regular Expressions For *Language Engineering*. Natural Language Engineering.

Karttunen L. 1998. *The Proper Treatment of Optimality in Computational Phonology*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.

Kempe A., Karttunen L. 1996. *Parallel Replacement in Finite-State Calculus*. COLING'96, Copenhagen.

Kempe A. 1997. *Finite State Transducers Approximating Hidden Markov Models*. ACL'97, Madrid.

Kiefer B., Krieger H., Carroll J., Malouf R. 1999. *A Bag of Useful Techniques for Efficient and Robust Parsing*. ACL'99, Maryland.

Kiefer B., Krieger H. 2000. *A Context-free Approximation of Head-Driven Phrase Structure Grammar*. IWPT'2000, Trento.

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

Koskenniemi K., Tapanainen P., Voutilainen A. 1992. *Compiling and using finite-state syntactic rules.* COLING'92, Nantes.

Kuhn J. 1998. *Towards data-intensive testing of a broad-coverage LFG grammar*. In Proceedings of KONVENS 98, Bonn. Peter Lang.

Kuhn J., Eckle-Kohler J., Rohrer. C. 1998. *Lexicon Acquisition with and for Symbolic NLP-Systems -- a Bootstrapping Approach*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.

Kukich K. 1992. *Techniques for automatically correcting words in text*. In ACM Computing Surveys, Vol. 24, N. 4, December, pp. 377-439.

La Serna N., Díaz A., Rodríguez H. 1997. *Parsers Optimization for Wide-Coverage Unification-Based Grammars using the Restriction Technique*. IWPT'97, Boston.

Lapata M. 1999. *Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations*. ACL'99, Maryland.

Levin B. 1993. *English verb classes and alternations*. The University of Chicago Press.

Macleod C., Grishman R., Meyers A., Barrett L., Reeves R. 1998. *NOMLEX: A Lexicon of Nominalizations*. Euralex'98.

Mangu L., Brill E. 1997. *Automatic Rule Acquisition for Spelling Correction*. Proceedings of the 14th International Conference on Machine Learning, ICML'97.

Marcus M. 1980. *A Theory of Syntactic Recognition for Natural Language*. The MIT Press.

Marcus M., Santorini B. 1991. *Building very large natural language corpora: the Penn Treebank*. CIS report, University of Pennsylvania.

Maritxalar M. 1999. *Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Maxwell J.T., Kaplan R. 1996. *An efficient parser for LFG*. Proceedings of LFG'96, Grenoble.

Mays E., Damerau F.J. Mercer R.L. 1991 *Context based spelling correction*. Information Processing and Management 27, 5.

McCord M. 1990. *A System for Simpler Construction of Practical Natural Language Grammars*. Natural Language and Logic, Lecture Notes in Artificial Intelligence, Springer Verlag.

Mellish C. 1989. *Some Chart-Based Techniques for Parsing Ill-Formed Input*. ACL'89.

Menezo J., Genthial D., Courtin J. 1996 *Reconnaisances pluri-lexicales dans CELINE, un système multi-agents de détection et correction des erreurs*. NLP + IA 96, Moncton, N. B., Canada.

Menzel W. 1988. *Error Diagnosing and Selection in a Training System for Second Language Learning*. COLING'88.

Menzel W., Schröder I. 1999. *Error Diagnosis for Language Learning Systems.* ReCALL, special edition, May.

Miller L. A. 1986. *Computers for Composition: a stage model approach to helping*. Visible Language XX 2.

Miller G. 1990. *Five papers on WordNet*. Special Issue of the International Journal of Lexicography, Vol. 3, N. 4.

Min K., Wilson W. 1998. *Integrated Control of Chart Items for Error Repair*. COLING-ACL'98, Montreal.

Mitjushin L. 1996. *An Agreement Corrector for Russian*. COLING'96, Copenhagen.

Mitton R. 1987. *Spelling checkers, spelling correctors, and the misspelling of poor spellers*. Information Processing and Management 23, 5.

Nunberg G. 1990. *The Linguistics of Punctuation*. CSLI Lecture Notes, 4 zenbakia, Stanford.

Oepen S., Callmeier U. 2000. *Measure for Measure: Parser Cross.fertilization - Towards Increased Component Comparability and Exchange*. IWPT'2000, Trento.

Oesterle J., Maier-Meyer P. 1998. *The GNoP (German Noun Phrase) Treebank*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.

Oflazer K., Okan Y. 1996. *A Constraint-based Case-frame Lexicon*. COLING'96 Copenhagen.

Oflazer K. 1999a. *Dependency Parsing with an Extended Finite State Approach*. ACL'99, Maryland.

Oflazer K., Zeynep D., Tür H., Tür G. 1999b. *Design for a Turkish Treebank*. Proceedings of Workshop on Linguistically Interpreted Corpora, at EACL'99, Bergen.

Oliva K. 1997. *Techniques for Accelerating a Grammar-Checker*. ANLP'97, Washington.

Peterson J.L. 1980. *Computer programs for detecting and correcting spelling errors*. Communications of the ACM 23, 12.

Peterson J.L. 1986. *A note on undetected typing erors*. Communications of the ACM 29, 7.

Pollard C., Sag I. 1987. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. CSLI Lecture Notes no. 13, The University of Chicago Press.

Pollard C., Sag I. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press.

Pollock J.J., Zamora A. 1984. *Automatic spelling correction in scientific and scholarly text*. Communications of the ACM 27, 4.

Popowich F., Vogel C. 1991. *A logic-based implementation of HPSG*. Natural language understanding and logic programming III, C. Brown and G. Koch (editoreak), Elsevier Science Publishers.

Prószéky G. 1994. *Industrial Applications of Unification Morphology*. ANLP'94, Stutgart.

Prószéky G. 1996. *Morphological Analyzer as Syntactic Parser*. COLING'96, Copenhagen.

Prószéky G. 1998. *An Intelligent Multi-Dictionary Environment*. COLING-ACL'98, Montreal.

Prószéky G., Kis B. 1999. *A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages*. ACL'99, Maryland.

Przepiórkowski A. 1999. *On Complements and Adjuncts in Polish*. Slavic in Head-Driven Phrase Structure Grammar liburuan, CSLI publications, Stanford.

Rabinovitz R. 1993. *Better writing through electricity*. PC Magazine, Maiatza.

Ramírez F., Sánchez-León F. 1996. *GramCheck: A Grammar and Style Checker*. COLING'96, Copenhagen.

Ramírez F., Sánchez-León F., Declerck T. 1997. *Corrección gramatical y Preprocesamiento.* SEPLN'97, Madrid.

Ritchie G., Pullman S. G., Black A. W., Russel G. J. 1987. *Computational Framework for Lexical Description*. Computational Linguistics, Vol. 13.

Ritchie G., Russel G. J., Black A., W., Pullman S. G. 1992. *Computational Morphology: Practical Mechanisms for the English Lexicon*. The MIT Press.

Roche R., Schabes Y. 1997. *Finite-State Language Processing*. MIT Press.

Rodríguez, C. 1991. *CORRECTOR: un sistema de verificación sintáctica y estilística de textos*. SEPLN'91.

Ruiz J.C., Zubizarreta J.R., Abaitua J. 1990. *Un compilador de LFG y su aplicación al euskara*. SEPLN'90, Donostia.

Sampson G. 1987. *Evidence against the 'Grammatical/Ungrammatical' Distinction.* Corpus Linguistics and Beyond, W. Meijs editorea, Rodopi, Amsterdam.

Samuelsson C., Voutilainen A. 1997. *Comparing a Linguistic and a Stochastic Tagger*. ACL-EACL'97, Madrid.

Sarasola K. 1988. *Caprate: Un Sistema de Interpretación de Problemas en Lenguaje Natural / Caprate: Lengoaia Naturalez idatzitako problemen interpretaziorako sistema*. Tesi-txostena, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea.

Satta G. 2000 *Parsing Techniques for Lexicalized Context-free Grammars*. IWPT'2000, Trento.

Schabes Y., Joshi A.K. 1991. *Parsing with Lexicalized Tree Adjoining Grammar*. Current Issues in Parsing Technology liburuan, Kluwer.

Schabes Y., Waters R.C. 1993. *Stochastic Lexicalized Context-Free Grammar*. IWPT'93, Tilburg-Durbuy.

Schank R., Lebowitz M., Birnbaum L 1980. *An integrated understander*. American Journal of Computational Linguistics 6.

Schiller A. 1996. *Multilingual Finite-State Noun Phrase Extraction*. ECAI'96 Workshop on Extended Finite State Models of Language, Budapest.

Schneider D., McCoy. K.F. 1998. *Recognizing Syntactic Errors in the Writing of Second Language Learners*. COLING-ACL'98, Montreal.

Selkirk E. 1982. *The syntax of words*. MIT Press, Cambridge.

Sells P. 1985. *Lectures on Contemporary Syntactic Theories*. CSLI, Stanford.

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

Shieber S.M. 1988. *Separating Linguistic Analyses from Linguistic Theories*. Natural Language Parsing and Linguistic Theories, U. Reyle eta C. Rohrer editoreak.

Silberztein M. 1997. *The Lexical Analysis of Natural Languages*. Finite-State Language Processing liburuan, MIT Press.

Skut W, Krenn B., Brants T., Uszkoreit H. 1997. *An Annotation Scheme for Free Word Order Languages*. ANLP'97, Washington.

Skut W, Brants T., Krenn B., Uszkoreit H. 1998. *A Linguistically Interpreted Corpus of German Newspaper Text*. Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain.

Sleator D, Temperley D. 1993. *Parsing English with a Link Grammar*. IWPT'93, Tilburg-Durbuy.

Smith J. 1992. *Mark your words with grammar-checking software*. PC/Computing, October.

Sparck Jones K., Galliers J.R. 1996. *Evaluating Natural Language Processing Systems*. Lecture Notes in Artificial Intelligence, Springer.

Sproat R. 1992. *Morphology and Computation*. The MIT Press.

Stede M. 1992. *The Search for Robustness in Natural Language Understanding*. Artificial Intelligence Review 6.

Tapanainen P., Voutilainen A. 1994. *Tagging Accurately-Don´t guess if you know*. ANLP´94.

Tapanainen P. 1996. *The Constraint Grammar Parser CG-2*. University of Helsinki. Publications nº 27.

Tapanainen P. 1997. *Applying a Finite-State Intersection Grammar*. Finite-State Language Processing liburuan, MIT Press.

ten Hacken P., Bopp S. 1998. *Separable Verbs in a Reusable Morphological Dictionary for German*. COLING-ACL'98, Montreal.

Tomabechi H. 1993. *A soft graph unification method for robust parsing*. IWPT'93, Tilburg-Durbuy.

Tomita M. 1986. *Efficient Parsing for Natural Language*.Boston: Kluwer Academic Publishers.

Tomita M. 1988. *'Linguistic' Sentences and 'Real' Sentences*. COLING'88, Budapest.

Trask, L. 1983. *Euskal izen sintagmaren egituraz*. 'Iker-2: Piarres Lafiteri Omenaldia' liburuan, Euskaltzaindia, Bilbo.

Trost H. 1990. *The application of two-level morphology to non-concatenative German morphology*. COLING'90, Helsinki.

Trost H., Matiasek J. 1994. *Morphology with a Null-Interface*. COLING'94, Japonia.

Turcato D., Nicholson D., Heift T., Toole J., Tsiplakou S. 2000. *A Parsing Methodology for Error Detection*. IWPT'2000, Trento.

Tzoukerman E., Liberman M. 1990. *A Finite-State Morphological Analyzer for Spanish*. COLING'90, Helsinki.

Urkia M., Sagarna A. 1991. *Terminología y Lexicografía Asistida por Ordenador. La experiencia de UZEI.* SEPLN'91, Donostia.

Urkia M. 1997. Euskal morfologiaren analisi automatikorantz. Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Uszkoreit H. 1986. *Categorial Unification Grammars*. COLING'86, Bonn.

Uszkoreit H. 1991. *Strategies for adding control information to declarative grammars*. ACL'91, Berkeley.

Uszkoreit H., Backofen R., Busemann S., Diagne A.K., Hinkelman E.A., Kasper W., Kiefer B., Krieger H., Netter K., Neumann G., Oepen S., Spackman S. 1994. *DISCO - An HPSG-based NLP system and its application for appointment scheduling*. COLING'94, Kyoto.

Van Berkel B., DeSmedt K. 1988. *Triphone analysis: A combined method for the correction of ortographical and typographical errors*. ANLP'88, Austin.

Vosse T. 1992. *Detecting and correcting morpho-syntactic errors in real texts.* ANLP'92, Trento.

Vosse T. 1994. *The Word Connection: Grammar-based Spelling Error Correction in Dutch*. PhD Thesis, Unit for Experimental and Theoretical Psichology, University of Leiden, Holland.

Voutilainen, A., Tapanainen P. 1993. *Ambiguity resolution in a reductionistic parser*. EACL'93, Utrecht.

Voutilainen, A. 1994a. *Three studies of grammar-based surface parsing of unrestricted English text*. Ph.D. thesis. University of Helsinki. Publications nº 24.

Voutilainen, A. 1994b. *Designing a parsing grammar*. University of Helsinki. Publications nº 22.

Voutilainen A, Järvinen T. 1995a. *Specifying a shallow grammatical representation for grammatical purposes*. EACL'95, Dublin.

Voutilainen A. 1995b. *A syntax-based part-of-speech analyser*. EACL'95, Dublin.

Voutilainen A. 1997. *Designing a Parsing Grammar*. Finite-State Language Processing liburuan, MIT Press.

Weischedel R.M., Sondheimer N.K. 1983 *Meta-rules as a Basis for Processing Ill-Formed Input*. American Journal of Computational Linguistics, 9.

Wiren, M. 1993. *Fully incremental parsing*. IWPT'93, Tilburg-Durbuy.

XTAG Group. 1995. *A Lexicalized Tree Adjoining Grammar for English*. Technical Report IRCS 95-03, University of Pennsylvania.

Yannakoudakis E.J., Fawthrop D. 1983. *The rules of spelling errors*. Information Processing and Management 19.

Yarowsky D. 1994. *A comparison of corpus-based techniques for restoring accents in Spanish and French text*. Proceedings of the Second Workshop on Very Large Corpora, Kyoto.

Zajac R. 1998. *Feature Structures, Unification and Finite-State Transducers*. Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, Ankara.

Zubimendi J.R., Esnal P. 1993. *Idazkera-liburua*. Eusko Jaurlaritzaren Argitalpen-Zerbitzu Nagusia.

Zubiri I., Zubiri E. 1995. *Euskal gramatika osoa*. Didaktiker.

Zubizarreta J.R. 1992. *Un modelo funcional de diálogo para diálogos orientados por la tarea*. Doktoretza-tesia, Euskal Herriko Unibertsitatea.

Zubizarreta J.R., Jones C. 1994. *Modeling Dialogue by Functional Subcategorization*. COLING'94, Kyoto.

BIGARREN PARTEA: APLIKAZIOAK