ERROREAK AUTOMATIKOKI DETEKTATZEKO TEKNIKEN AZTERLANA ETA EUSKARARENTZAKO APLIKAZIOAK

Maite Oronoz, Arantza Díaz de Ilarraza, Koldo Gojenola
(Universidad del País Vasco/Euskal Herriko Unibertsitatea)

Abstract

In this article, we study the techniques used for detecting errors in Natural Language Processing (NLP). We classify the techniques according to their approach (symbolic or empirical), and then we describe them in depth. Following that, we describe the systems we have developed for detecting syntactic errors in Basque, by using that technique as a criterion for the classification of those systems, and enhancing it with examples.

1. Sarrera

Ikasten ari garenean, bai orokorrean, baita hizkuntzen ikaste-prozesuan ere, bidetik aldentzea, hots erroreak egitea, ezinbestekoa da. Jendeak ezin du hizkuntzarik ikasi lehenengo sistematikoki errorerik egiten ez badu (Dulay et al. 1982). Erroreak garai batean gaizki ikusiak baziren ere, ikaste-prozesuaren parte kontsideratzen dira gaur egun. Hizkuntza zuzen ikasteko, ikasketa-prozesuaren uneren batean egindako erroreak detektatzea, eta beharrezkoa kontsideratzen denean zuzentzea, garrantzitsua da.

Erroreak testuetan automatikoki detektatzeko, IXA taldean ordenagailuak eta zehazkiago hizkuntzaren tratamendu automatikoko edo lengoaia naturalaren prozesamenduko (LNP) teknikak erabiltzen ditugu. Artikulu honen helburua, erroreak automatikoki detektatzeko teknikak zeintzuk diren aztertzea eta horiek euskarazko zenbait errore detektatzeko nola erabili diren azaltzea da.

Erroreak edo gaizki erabilitako egiturak detektatzea ataza oso garrantzitsua da hizkuntzaren tratamendu automatikoko ondorengo bi alorretan:

— Ortografia- eta gramatika-zuzentzaileetan.
— Ordenagailuz Lagundutako Hizkuntzen Irakaskuntzan[1] (OLHI).

Erroreak automatikoki nola detekta daitezkeen azaltzen hasi aurretik, *errorea* kontzeptua bera definituko dugu lehenengo, eta erroreen detekzio automatikoa gero.

[1] Irakaskuntza terminoa modu honetan idatzita, «ikaskuntza» eta «irakaskuntza» terminoak bildu nahi ditugu.

1.1. Errorea

Hizkuntzaren erabilera estandarraren paradigmetatik hizkuntza-erabiltzaile indibidualek egiten dituzten desbideratzeei *errore* deitzen diete batzuek (adibidez, www.wikipedia.com hiztegi entziklopedikoan).[2] Vandeventer-en arabera (2003) erroreak *zuzena* zer den adierazten duen araua existitzen delako soilik detekta ditzakegu. Laburtuz, *errorea* honela definituko dugu: egitura zuzenetik, arautik, desbideratzen den egitura da.

Egin dugun definizioa orokorra da eta hizkuntzarekin lotutako edozein erroretarako, errore ortografiko, sintaktiko, semantiko edo pragmatikoetarako izan daiteke baliagarria. Lan honetan *edizio mekanikoaren* ondorioz sortutako erroreak alde batera utzi eta errore *kognitiboei* erreparatuko diegu. Testu zati batean errorea dagoela esatea ez da lan erraza izaten, eta askotan eztabaidagarria izan daiteke kontu hau.

1.2. Erroreak automatikoki detektatzea

Errorea terminoa definitu ondoren, artikulu honen izenburuan haren jarraian ipini dugun terminoaren inguruan arituko gara orain: *detekzioa.* Guk, errorearen *detekzioa, diagnosia* eta *zuzenketa* bereiziko ditugu. Kasu honetan, dudak, lehen bi terminoen artean suerta daitezke, batzuetan detekzioa eta diagnosia terminoak fenomeno bera izendatzeko erabiltzen baitira.

Gure ustez, testu bat aztertu ondoren, erroreak daudela esatea eta erroreak *kokatzean* datza erroreak *detektatzea,* zergatirik edo argibiderik eman gabe. Erroreak agerian jartzen dira inolako deskribapenik egin gabe. *Diagnosia* egiten denean, ordea, errorea deskribatu egiten da; gaizki dagoena aztertu egiten da zehatz-mehatz arazoa zein den adierazteko; kokatzea baino gehiago egiten da. *Zuzenketa* egitean egitura erroredunaren ordezko zuzenak ematea da helburua, gehienetan aldez aurretik errorearen diagnosia eginez, eta gutxi batzuetan, diagnosirik egin gabe. Ikus ditzagun definizioak adibideak erabiliz:

— Esaldia:

- Zentral nuklearrak zakar erradiaktiboa eratzen dute.

— Detekzioa:

- [Zentral nuklearrak] zakar erradiaktiboa eratzen [dute].

— Diagnosia:
Esaldiko subjektuak, *Zentral nuklearrak,* eta aditz laguntzaileak, *dute,* ez dute numeroan edo/eta kasuan komunztatzen.

— Zuzenketak:

1. Zentral nuklearrak zakar erradiaktiboa eratzen du.
2. Zentral nuklearrek zakar erradiaktiboa eratzen dute.

---

[2] Ingelesezko bertsioan, *error in oral and written language.*

Sarreratxo hau egin ondoren, tekniken deskribapen zabala egingo dugu *teknikak* izeneko atalean (2 atala). Aspaldidanik gabiltza IXA taldean euskarazko erroreak automatikoki detektatzen, horretarako teknika ezberdinak erabiliz. Horren adibide batzuk emango ditugu, 3 atalean. Amaitzeko, ateratako ondorio batzuen berri emango dugu 4 atalean.

2. Teknikak

Erroreak detektatzeko eta zuzentzeko erabiltzen diren teknikak ugariak dira, hizkuntzaren tratamendu automatikoan erabiltzen diren ia teknika guztiak eginkizun zehatz honetarako erabil daitezkeela esan baitezakegu. Teknika guztiek zenbait eredu formaletan dute oinarria. Jurafsky eta Martin-en arabera (2000), eredu horiek *egoera-makinak, erregela-sistema formalak, analisi logikoa, probabilitatearen teoria* eta *ikasketa automatikoan* oinarritutako erremintak dira.

Egoera-makinak eta *erregela-sistema formalak* fonologiako, morfologiako eta sintaxiko ezagutza lantzeko tresna nagusiak dira.

Semantikarekin, pragmatikarekin eta diskurtsoarekin lan egitean erabiltzen den eredu formala *analisi logikoa* da normalean, baina alor horiek ez dira artikulu honetan aztergai, eta beraz, ez dugu analisi logikoari buruz hitzik egingo.

Probabilitatea da aitatu dugun azken eredu formala. Probabilitatearen erabilera nagusia anbiguotasun-arazoak ebaztea da eta LNPko arazo gehienak "sarrera (*input ingelesez*) anbiguo batentzako N aukera emanda, probabilitate handienekoa aukeratu" moduan definitu daitezkeenez, oso erabilia da. *Ikasketa automatikoarekin* oso lotuta dago probabilitatea.

Tekniken eredu formalak aztertu ondoren, oinarrian erabiltzen duten informazio mota irizpide hartuta, honako sailkapena egin dugu:

1. Hizkuntza-ezagutzan oinarritutako teknikak.
2. Corpusetan oinarritutako teknikak.

Hizkuntza-ezagutzan oinarritutako teknikek, teknika sinboliko izena ere jasotzen dute. Metodo sinbolikoak honela definitzen dira (Dale 2000) lanean: "symbolic methods —where knowledge about language is explicitly encoded in rules or other forms of representation—". Hau da, teknika sinbolikoetan hizkuntzari buruzko ezagutza erregeletan edo beste adierazpideetan kodetzen da modu esplizituan. Teknika sinbolikoak oso erabiliak izan dira azken 40 urteetan, baina mundu errealeko arazoei aplikatzen zaizkienean *hedagarritasunaren arazoa* dute.

Arazo honen aurrean, azken 10 edo 15 urteotan *corpusetan oinarritutako teknikek* edo teknika *enpirikoek* hazkunde izugarria izan dute. Horietan, datu kopuru oso handiak erabiltzen dira (LNPren kasuan corpus oso handiak), eta horiekin batera, estatistikaren inguruko prozedurak. Ezagutza testuetatik automatikoki erauzi egiten da, erregeletan eskuz kodetu beharrean. Hau horrela izanik, hurbilpen hau hizkuntza-ezagutzarekiko independentea da.

Gaur egun, giza-hizkuntzaren prozesamenduaren gakoa teknika sinboliko eta enpirikoen konbinazio egokian dagoela onartzen dute ikertzaile gehienek.

Metodo sinbolikoak erabiltzen dituzten teknikak hurrengo atalean aztertuko ditugu (2.1). Enpirikoak, berriz, 2.2 atalerako utzi ditugu.

2.1. Hizkuntza-ezagutzan oinarritutako teknikak (sinbolikoak)

Hizkuntza-ezagutzan oinarritutako tekniketan erregelak eskuz garatzen dira eta testuek informazio linguistiko handia erabilita analizatuta egon behar izaten dute. Hiru azpimultzotan banatu ditugu behar duten informazio linguistikoaren arabera:

1. Ezagutza morfologikoa edo hitz-mailako informazioa behar dutenak.
2. Ezagutza sintaktikoa edo esaldi-mailako informazioa behar dutenak.
3. Esaldi-mailatik haratago dagoen ezagutza —semantika, pragmatika eta diskurtsoa— behar dituztenak.

Multzo hauek 2.1.1, 2.1.2 eta 2.1.3 ataletan aztertuko ditugu, hurrenez hurren. Garrantzirik handiena sintaxia oinarrian duen atalari eman diogu, hauxe baita bereziki landu duguna.

2.1.1. Ezagutza morfologikoa (hitz-maila)

Hitza, sintagma eta goragoko mailetako egituren osagaia den unitatea da. Hitzen barne-egitura *morfologiak* lantzen du eta hitza lemekin eta zenbaitetan morfemekin osatuta egoten da. Morfemetan esaldi-mailako funtzioak topa ditzakegunean, honek propietate morfologikoak eta sintaktikoak edukitzen ditu, eta kasu horietan *morfosintaxia* terminoa erabiltzen da. Euskara hizkuntzaren izaera eranskaria dela eta, morfosintaxia terminoa oso erabilia da.

Hitzekin maila morfologikoan lan egiteko *adierazpen erregularrak* eta *automatak* erabiltzen dira normalean. Teknikaren aukeraketa, hizkuntzaren ezaugarriek erabakitzen dute maiz. Adibidez, ingelesaren antzeko hizkuntzetan errore ortografikoak lantzeko, hitz baten forma flexionatu guztiak zerrenda batean gorde daitezke, eta zerrenda horrekin hizki-zuhaitz moduko egoera finituko automata bat eraiki. Horrela, hitz zuzenak ezagutuko lirateke eta hitz erroredunak baztertu. Suomiera, hungariera, turkiera eta euskara moduko flexio-maila handiko hizkuntzetan, ordea, arazoak egoten dira forma guztiak hiztegian gorde ahal izateko. Arazoa konpontzeko premiazkoa da hitzak morfologikoki behar bezala lantzen dituzten teknikak erabiltzea.

Batzuetan, egiaztatzaileetan hizkuntza bakoitzerako izaten duten hiztegian lemak eta morfemak gordetzeaz gain, hitzen sorkuntzarako erregelak gordetzen dira. Adibidez, ingeleserako hiztegian *twenty/H* bikotea gordeko genuke. Zuzentzaileak *twentieth* topatuz gero, H erregelaren arabera, *ieth* ezabatuko luke eta *y* gehitu. Hiztegian *twenty* balego, *twentieth* zuzena litzateke.

Euskarazko hitzen analisirako Koskenniemi-k (1983) garatutako *bi mailatako (lexikoa, azalekoa) morfologia*n oinarritutako analizatzailea erabiltzen da, zeinetan analisia eta sorkuntza egiten diren. Hitzen analisia egiteko, sarreran *azaleko* forma (adibidez, *etxean*) jasotzen da eta emaitzan dagokion adierazpen *lexiko* zuzena itzultzen da (adibidez, *etxe[IZE][ARR] + an[INE]*). Sorkuntzan, forma lexikala emanda, azaleko forma lortzen da. Bi mailatako erregelek, transformazio morfofonologikoen eraginez sakoneko eta azaleko mailen artean sortzen diren diferentziak adierazten dituzte. Eta honetaz gain, eransketa eta ezabaketaren ondorioz sortutako erroreak deskribatzen dituzte. *Xuxen* zuzentzaile ortografikoan (Agirre et al. 1992) erabiltzen da.

Flexio handiko turkiera hizkuntzarako, erroreak onartzen dituzten egoera finituko eredu ezagutzaileak erabiltzen ditu Oflazer-ek (1996). Egoera finituko ezagutzaileak, hitz zuzenetatik *urruntzen* diren karaktere-kateen ezagutza ahalbidetzen du. Karaktere-kate bat beste batean bihurtzeko zenbat edizio-eragiketa (eransketa, ezabaketa, aldaketa, transposizioa) behar diren neurtzen duen *edizio-distantzia* izeneko kontzeptua erabiltzen du erroreak detektatzeko.

2.1.2. Ezagutza sintaktikoa (esaldi-maila)

Hitza aztertu ondoren, unitate handiago bat aztertzera joko dugu: esaldia. Sintaxiak hitzak sintagma izeneko egituretan antolatzen ditu eta egitura sintagmatiko horiek, era berean, perpausetan/esaldietan[3] antolatzen ditu.

Esaldi bat sintaktikoki analizatzen denean, egituraren bat esleitzen zaio. Egiturak esaldiko osagai linguistikoak errepresentatzen ditu, eta beren arteko harreman gramatikalak azalarazten ditu.

Lan hori guztia modu mekanikoan gauzatzen duten algoritmo eta programek *parser* edo analizatzaile sintaktiko izena hartzen dute. Analizatzaileetako askotan hizkuntza-ezagutza *gramatiketan* kodetzen da *hizkuntza-egiturak* deskribatzen dituzten erregelen bidez. Erroreen detekzioaren lan eremuan, hizkuntza-egiturak gramatikalak edo ez-gramatikalak izan daitezke.

Gramatika bat idazteko garaian, aukera ezberdinak egin daitezke (Gojenola 2000). Erroreen detekzioaren ikuspuntua kontuan hartuta erabil daitezkeen irizpideen zerrendapena egingo dugu:

— Hizkuntzaren teoria vs hizkuntzaren tratamendu automatikoa. *Hizkuntzaren teoria* hizkuntza baten (edo hizkuntzen arteko) oro deskribatzeaz arduratzen den bitartean, hizkuntza modu automatikoan ulertzea eta sortzea du helburu *hizkuntzaren tratamendu automatikoak.* Azken honetarako ordenagailuak erabili beharrak, eragin handia izaten du sintaxiaren eta semantikaren teorizazioan, teoriak konputagarriak izan behar duten eragiketetara murriztu behar baitira. Hori egitean, LNP sistemek gehienbat formaltasuna, erazagutzaile izateko gaitasuna eta egokitasun linguistikoa mantendu behar dituzte (Shieber 1986). Hizkuntzaren teoriako formalismoetan normalean datu gutxi eta laborategikoak erabiltzen dira, erabilitako esaldiak gramatikalak dira eta desanbiguazioa, puntuazioa eta antzeko fenomenoak ez dira tratatzen. Aplikazio errealetarako erabiltzen diren gramatiketan, berriz, datu askorekin eta esaldi errealekin lan egin behar da: esaldi gramatikalekin eta esaldi ez-gramatikalekin.
— Esaldiaren analisi osoa vs analisi partziala. Testu errealekin lan egin beharrak, *analisi partzialaren* erabilera bultzatzen du. Edozein egitura, gramatikala zein ez-gramatikala, analizatzeko gai den sistema behar dugu, eta sistema horietan ezin izaten dira aurreikusi hizkuntza-egitura posible guztiak; beraz, ezin izaten da beti analisi osoa egin. Analisi partzialak, fidagarritasuna eta sendotasuna

[3] Artikulu honen helburua teknikak aztertzea da, eta ez egitura linguistikoak, beraz, hemendik aurrera «esaldi» kontzeptu orokorra erabiliko dugu, perpausen eta esaldien arteko bereizketarik egin gabe.

ditu helburu, sakontasuna eta osotasuna neurri batean galduaz. Beste aldetik, badira sistemak, teoria linguistiko baten azterketarako pentsatuak, *analisi osoa* bakarrik lortzea helburu dutenak, fenomeno linguistiko interesgarriak aztertzeko pentsatuta daudelako, eta ez testu errealetako esaldiak.

— Osagai-egitura vs mendekotasun-egitura. Osagai-egituretan oinarritutako analisiaren ideia nagusia hau da: hitz multzoak *osagai* deituriko unitate bakar moduan kontsidera daitezke (Jurafsky eta Martin 2000). Adibidez, izen-sintagma izeneko hitz multzoak unitate bakar moduan jokatzen du maiz. Mendekotasun-egituretan egitura sintaktikoak dependentzia izeneko erlazio bitar asimetrikoekin elkarren artean lotutako elementu lexikalekin osatuta daude (Nivre 2005).

Sintaxia lantzeko teknikak era askotarikoak dira, baita sintaxi-mailako erroreak detektatzeko teknikak ere. Guk Chomskyren hierarkiako eredu ezagutzaileak —egoera finituko mekanismoak— eta eredu sortzaileak —gramatikak—aztertu ditugu sailkapenerako eredu baten bila. Egoera finituko mekanismoak eta, beraz, lengoaia erregularrak, morfologia deskribatzeko nahikoak izan ohi dira. Sintaxia deskribatzeko, berriz, baliabide ahaltsuagoak behar izaten dira. Orokorrean, zeregin horretarako, *testuingururik gabeko gramatikak (TGG)* erabili izan dira (Aho et al. 1985), esaldien egitura hierarkiko eta errekurtsiboak definitzeko egokiak baitira. Bestalde, egoera finituko mekanismoak (automatak eta transduktoreak) oso erabiliak izan dira hizkuntzaren tratamendu automatikoan, eredu linguistikoetan oinarritutako patroiak definitzeko duten erraztasuna dela eta. Egin dugun azterlanean, bai *TGGak baita egoera finituko mekanismoak* ere, testu erreal erroredunetan gertatzen diren egiturak deskribatzeko erabili direla ikusi dugunez, sistemak irizpide honen arabera antolatu ditugu:

— Egoera finituko mekanismoak erabiltzen dituzten teknikak.
— Gramatikak oinarrian dituzten teknikak.
— Bestelakoak.

Egoera finituko mekanismoak

Lan honetan zehar, egoera finituko mekanismoei buruz ari garenean, patroi-erregelak aipatuko ditugu, behin eta berriro. Egia esateko, guk baliokidetzat hartuko ditugu, patroi-erregeletan oinarritutako formalismo gehienek patroiak automata bihurtzen baitituzte. Patroi-erregelak egitura edo ereduren bat definitzen duten erregelak dira. Gramatika-zuzentzaileetan, erregelak egitura ez-gramatikalak deskribatzeko erabiltzen dira. OLHI sistemetan, berriz, batzuetan funtzio hori dute, baina maiz, sistemak espero duen erantzuna eta ikasleak eman diona konparatzeko erabiltzen dira.

Erregelen bidez deskribatzen ditugun patroiak dagoeneko morfosintaxi-mailan analizatuta dagoen testu baten kontra parekatzen ditugunean, patroi-parekatze teknika erabiltzen ari gara. Erregelek egitura oker bat deskribatzen dutenean, metodoa testuetan erroreak topatzeko erabil dezakegu. Patroi-parekatzearen teknikak abantaila anitz ditu; besteak beste: i) ez da esaldi osoa analizatu behar egitura mugaturen batean errorea dagoela salatzeko, ii) erregelak modu indibidualean erabil daitezke eta iii) erroreari buruz xehetasun handiko mezuak eskaini daitezke komentario la-

gungarriekin. Gainera, erregelak ulerterrazak izan ohi direnez, erabiltzaileek erregela multzoak erraz heda ditzakete errore-kasu sinpleak modu inkrementalean garatuz. *Patroi-erregelak* erregela-lengoaia ezberdinak erabiliz idatz daitezke, batzuk ezagunak diren formalismoetan (*Murriztapen Gramatika* eta *Xerox Finite State Tool* tresnak erabiliz), besteak norberak garatutakoetan.

Murriztapen Gramatika formalismoa (MG) (Karlsson et al. 1995, Tapanainen 1996) azken urteotan azaleko sintaxia eta desanbiguazioa lortzeko sistemarik arrakastatsuenetarikoa izan da. Desanbiguazioa egiten duela diogunean, forma baten interpretazioen artean, analisi zuzenak/egokiak aukeratzen dituela esan nahi dugu. Nahiz eta tresnaren helburu nagusia desanbiguazioa den, formalismoa testuari etiketak esleitzeko gai da islapen-erregelen bidez. Islapen-erregelak erabilita errore baten hasiera eta errorearen bukaera deskribatzeko gai bagara, MGk errorea mugatzeko eta, beraz, detektatzeko aukera ematen digu.

Erroreen detekziorako maiz erabili da MG. Katalanerako (Badia et al. 2004), suedierarako (Birn 2000, Arppe 2000) eta norvegierarako (Johannessen et al. 2002) sistemetan erabili dute formalismo hau modu arrakastatsuan. MG erabiliz definitutako errore-patroiak analizatzaile morfologiko baten bidez analizatutako testuari aplikatzen zaizkio. Analizatzaileak informazio zuzena analizatzeko prestatuak egoten direnez, egokitzapenak, murriztapenen "erlaxazioak" egin behar izaten dira testu ez-gramatikalak analizatu ahal izateko.

Xerox-eko ikerketa-taldeak sarreratzat adierazpen erregularrak jasotzen dituen eta horiek automata transduktore bihurtzen dituen *Xerox Finite State Tool (XFST)* tresna garatu du (Karttunen et al. 1997, Aït-Mokhtar eta Chanod 1997). Tresnak eragiketa multzo aberatsa du eta ezaugarri honek adierazpenerako indar handia ematen dio.

Hashemi et al. (2003) lanean erroreen detekziorako XFST erabiltzen duen sistema bat deskribatzen da. Lan horretan erabilitako teknika transduktoreen arteko kenketan oinarrituta dago. Egitura linguistiko zuzen eta okerrak deskribatzen dituen *gramatika zabal* bat konpilatuz lortutako transduktorearen, eta egitura zuzenak soilik deskribatzen dituen *gramatika estuaren* bidez lortutako transduktorearen arteko kenketa egiten dute. Kenketa egin ondoren lortutako automatek, egitura okerrak detektatzeko gaitasuna dute.

Aipatutako formalismo ezagunez gain, patroietan oinarritutako errore-erregelak idazteko modu gehiago definitu dira. Naber-ek (2003) ***Ex**tensible **M**arkup **L**anguage*-n (XML) eta *Python* programazio-lengoaian idatzitako erregelak erabili zituen erroreak deskribatzeko.

Patroi-erregelak definitzeko erregela-lengoaia egituratu propioa erabiltzen dute Granska gramatika-zuzentzaile hibridoan (Carlberger et al. 2002, Domeij et al. 1999). *Granskan,* erregelak aplikatzeko estatistika erabiltzen dute. Aldez aurretik, kategoria bakoitzeko aplikatzeko probabilitate handieneko erregela kalkulatzen da. Testuan hitz edo etiketa konkretu bat agertzen denean, estatistikei kasu eginez, dagozkion erregelak soilik aplikatzen dira, ez guztiak.

Testuingururik Gabeko Gramatikak (TGG)

Aipatua dugu testuingururik gabeko gramatiken egokitasuna esaldien egitura hierarkikoa definitzeko. TGG sinpleak eta *baterakuntzan* oinarritutakoak bereiz ditza-

kegu. TGG sinpleetan erregelekin osagai atomikoak deskribatzen diren bitartean, gainontzeko TGGetan osagaiei informazioa gehi diezaiekegu ezaugarri-egituren bitartez (gramatikaren sinpletasuna eta trinkotasuna bultzatzen da). Ezaugarri-murriztapenen arteko bateragarritasuna egiaztatzeari eta bateragarriak direnean informazio hori trinkotzeari, baterakuntza deitzen zaio. Baterakuntza oinarritzat duten formalismorik ezagunenak, *Lexical Functional Grammar (LFG), Generalized Phrase Structure Grammar (GPSG)* eta *Head-Driven Phrase Structure Grammar (HPSG)* ditugu.

Oinarrian TGGak dituzten sintaxi-analizatzaileak asko erabili izan dira erroreen detekziorako, emaitza onargarriak lortuaz. Erroreen detekziora bideratuta ez dauden TGGak ez dira prestatuta egoten sintaxi-egitura ez-zuzenak analizatzeko, beraz, gaizki eratutako esaldi bat jasotzen dutenean, ez dira gai izaten esaldiaren analisi-zuhaitza eraikitzeko. Hori egin ahal izateko, moldaketaren bat behar izaten dute. Analisi-zuhaitza ez sortzeko arrazoia murriztapenak ez betetzea denean, murriztapenak *erlaxatzearekin* nahikoa da. Beste batzuetan aldatzen dena gramatika-egitura bera denean, *errore-gramatikak* erabiltzen dira.

Baterakuntza + Erlaxazioa

Baterakuntza-formalismoak egokiak dira *murriztapenen erlaxazio* izeneko teknikaren aplikaziorako. Baterakuntzan, ezagutza sintaktikoa ezaugarri-balio egituren bidez adierazten dela esan dugu. Egitura horietan bete beharreko murriztapenak ekuazioen bidez deskribatzen dira. Esaldi baten analisi-zuhaitza lortzen ez denean, ekuazio horietako batzuk kendu egiten dira (erroreen sorburua izan daitezkeenak, adibidez, numeroa komunztaduran) analisia lortu ahal izateko. Kendutako ekuazioek adieraziko dute akatsaren iturria. Patroi-erregeletan gertatzen ez zen bezala, erroremezu zehatzak ematea zaila da analizatzaileak esaldi zuzenen analisia egiten baitu.

Erroreen detekziorako murriztapenen erlaxazioa erabiltzen duten sistema ugarik teknika hau beste batzuekin konbinatzen dute. Testu erroredunetan topa daitezkeen erroreak era askotakoak direnez, estrategia ezberdinak jarraitzen dira errore mota bakoitzaren detekziorako. Sistema batzuetan, *erroreei aurre hartzen* dieten erregela esplizitu lokalekin eta heuristikoekin konbinatzen da erlaxazioa (Bustamante eta León 1996, Teixeira Martins et al. 1998). Beste zenbaitetan, murriztapenen erlaxazioa errore-gramatika batekin konbinatzen dute (Vosse 1992).

Erlaxazioaren teknika erabiltzean, analizatzaileak analisi posible ugari eman ohi ditu, esaldi bat analizatzeko erregela erlaxatu gabeak eta erlaxatuak erabiltzen baitira. Erlaxazio-konbinazioak murrizteko (Gojenola 2000), lanean ezaugarrien erlaxazio mailakatua proposatzen dute.

Baterakuntza + Errore-gramatikak

Esaldi erroredunetan hizkuntza-egiturak aldatzen direnean, esaldiaren analisi-zuhaitzak lortzea ez da posible izaten. Analisia ahalbidetzeko, hizkuntza-egitura okerrak *errore-gramatiketan* deskribatzen dira. Errore-gramatika bat errore tipikoak adierazten dituzten esaldi egiturak deskribatzen dituen gramatika bat da. Hizkuntza-egitura zuzenak deskribatzen dituen gramatikarekin sarrerako testu erroredunaren analisia egitea ezinezkoa denean, errore-gramatikari pasatzen zaio testua. Errore-gramatikako

egituraren batekin analisia posible bada, errore-mezu bat erakusten da. Kasu honetan ere, egoera finituko mekanismoetan gertatzen zen moduan, erroreei aurrea hartu behar zaie, hau da, aldez aurretik ezagutu behar ditugu errore posibleak.

Foster eta Vogel-en lanean (2004) errore-gramatiken alde egiten da eta erlaxazioaren eta analisi partzialen bateraketa erabiltzen duten teknikei zenbait kritika egiten zaie. Hona hemen kritika horietako batzuk:

1. Erlaxazioaren teknika erabiltzean esaldi zuzen batean eta esaldi oker batean hitz kopuru bera egoten dela suposatzen da. Eta, beraz, hitzak esaldian ekiditen edo gehitzen direneko kasuak ez omen dira lantzen. Errore-gramatika fenomeno hori lantzeko gai da.
2. Esaldi okerra analizatu ahal izateko erlaxazioa erabiltzean, murriztapen ugari erlaxatzeko beharra egon daiteke, eta horien ondorioz, zentzurik gabeko analisiak lor ditzakegu. Hori ez da errore-gramatikekin gertatzen, errore-erregelak modu kontrolatuan gehitzen baitira.
3. Errore-gramatikek gramatikala ez denaren eredu linguistiko bat ematen dute eta azpitik ezagutza dutenez, diagnosirako informazio zehatza emateko gai dira.

Orain arte, Chomskyren hierarkiarekin zerikusia duen sailkapena erabili dugula esan dugu: teknikak egoera finituko mekanismoetan eta gramatiketan sailkatu ditugu. Irizpide horren ordez irizpidetzat esaldiko egitura sintaktikoak errepresentatzeko modua hartzen badugu, erroreak lantzeko teknikak *osagai-egituretan* eta *mendekotasun-egituretan* sailka ditzakegu. Gure kasuan, interes berezia dugu mendekotasun-egituretan oinarritutako tekniketan, beraz, horien inguruan arituko gara hurrengo lerroetan.

Mendekotasun-egiturak oinarri dituzten teknikak

Mendekotasun-gramatikak sintaxia adierazteko modu bat izan dira tradizionalki. Tradizio horrek gramatika-teoria eta -formalismo multzo handi bat du azpian, eta horiek ideia bera (Nivre 2005):

> The fundamental notion of *dependency* is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of a sentence.

Mendekotasun-gramatiken inguruan egiten diren lanek maiz Tesnière-ren lanetik edaten dute (Tesnière 1959). Mendekotasun-erlazioak, hitz-formen artean definitzen dira gehienetan, baina adabegitzat sintaxi-osagaiak, lemak eta morfemak (Mel'čuk 1988) dituzten mendekotasun-zuhaitzak existitzen dira. Erlazioei dagokionez, nahiz eta gehienetan sintaxi-mailakoak diren (subjektu, objektu...), zenbait kasutan semantika-mailakoak izaten dira (egile, jasale...). Gu, sintaxi-mailako erlazioak dituzten mendekotasun-zuhaitzen inguruan arituko gara. Ohiko mendekotasun-egiturez gain, bereziki esteka-gramatikak (*Link grammar* ingelesez) aipatuko ditugu (Sleator eta Temperley 1993), erroreak detektatzeko saiakerak egin baitira sistema gramatikal formal honekin.

Esteka-gramatiketan hitzen arteko estekak ipintzeko bi parametro nagusirekin egiten da lan: noranzkoarekin eta distantziarekin. Hiztegi batean hitz bakoitzak eskuinera ("+") edo ezkerrera ("–") zein gramatika-kategoriako hitzekin egin dezakeen

lotura (D-, O- e.a.)[4] definitzen da eta hitzen artean beharrezko baldintzak betetzen badira, hitzen arteko lotura sortzen da. Arku bat ezarri ahal izateko baldintza hauek bete behar dira: i) hitz bakoitzaren baldintza lokalak bete behar dira, ii) estekak ezin dira gurutzatu esaldian eta iii) hitzek grafoa osatu behar dute. Esaldi erroredunak analizatu ahal izateko, detektatu nahi den errore motarentzako garrantzitsuak diren hitzen definizioa aldatu egin behar da (Brehony eta Ryan 1994).

Mendekotasun-egiturak hitzen ordena librea duten eta beraz, sintaxi-egitura ezjarraituak dituzten hizkuntzetarako egokiak dira (Mel'čuk 1988). Erroreak detektatzeko mendekotasun-egiturak erabiltzen dituzten txekierarako (Holan et al. 1997), errusierarako (Mitjushin 1996) eta frantseserako (Courtin et al. 1991) sistemak topatu ditugu.

Txekierarako egindako lanean zuhaitzetan gerta daitezkeen trinkotasun ezak metaerregeletan deskribatzen dira (Holan et al. 1997). Esaldia analizatzerakoan metaerregelaren bat erabiltzen bada, eraikitako zuhaitzean ikur negatibo bat txertatzen da. Erregeletan murriztapenen erlaxazioa egiteko aukera dago. Analisia faseka egiten da eta *ondoz ondoko harreman sintaktikoak deskribatzen dituen ezaugarria* (*projectivity*)[5] eta zuhaitzeko ikur negatiboak kontuan hartzen dira, zuhaitzak errorerik ote duen edo ez erabakitzen duen modulura bidaltzeko.

Sintaktikoki ez-konektatuta egotearen ezaugarria erabiltzen dute errusieraz erroreak detektatzeko (Mitjushin 1996). Konektatutako segmentuen kopurua 1 denean, egitura sintaktiko osoa lor daiteke analisian. Bestela, hasierako esaldian aldaketa batzuk egin behar dira esaldi erroredunaren aldaeraren baten analisia lortu arte. Horrela, esaldi erroredunaren zuzenketa lortzen dutela diote. Ez dute diagnosia egiten, batzuetan ez baitakite zein errore gertatu den.

Frantseserako egindako lanean (Courtin et al. 1991), mendekotasun-zuhaitzei baldintza-ekintza moduko erregelak aplikatzen dizkiete komunztadura-erroreak detektatu ahal izateko.

Bestelakoak

Atal honetan inolako multzotan sailkatu ezin ditugun teknikak azalduko ditugu motz motzean.

1. DeSmedt-ek (1995) *predikatuak zuzendurikoa* izeneko teknika proposatzen du. Teknika honetan lehenik aditza identifikatzen da, eta gero, esaldiaren gainontzeko elementuak aditzaren osagarri moduan identifikatzen saiatzen da. Adibidez, aditz trantsitibo bat edukita, objektua izan daitekeen sintagma baten bila hasiko da analizatzailea. Esaldiko elementu guztiei, aditzarekiko betetzen duten funtzioa esleitzen zaie. Elementuren bat funtziorik gabe geratzen bada, esaldia ez dela guztiz zuzena suposatzen dute.

---

[4] Hitz baten definizioan D- topatuz gero, ezkerrera D konektore bat (ustez determinatzailea) behar duela adierazten da. Adibidez, *cat* hitzak ezkerrera *the* edo *a*. . . behar du hauek sintagma batean lot daitezen.

[5] «A dependency graph satisfies the constraint of projectivity with respect to a particular linear order of the nodes if, for every arc $h \rightarrow d$ and node w, w occurs between h and d in the linear order only if w is dominated by h» (Nivre 2005).

2. Vandeventerrek (2003) *zatien berrinterpretazioa* erabiltzen du. Ideia nagusia honakoa da: "sintaxi-analizatzaile bat analisi osoa lortzeko gai ez denean, analisi partzialak edo zatiak lortzen ditu. Zati hauek bilduz gero, esaldi osoa lortuko dugu. Beraz, zatiak aztertuko ditugu, esaldi osoaren analisia ez lortzearen zergatia ezagutzeko." Analizatzailearen estaldura nahikoa bada, analisi osoa ez lortzearen arrazoia errore bat izan daiteke.

2.1.3. Ezagutza semantikoa

Erroreen sailkapenak aztertuz gero, *errore semantikoaren* kontzeptua ez dela gehiegi aipatzen ikusiko dugu. Ortografia- eta sintaxi-mailan zuzenak diren baina kokatuta dauden testuinguruan zentzurik ez duten hitz edo esaldiek *errore semantiko* bat sortzen dutela esan ohi da. Hori, ordea, *hitz zuzenen erabilera nahasketa* izeneko kontzeptuari egiten zaion definizioa da.

Jo dezagun hitz zuzenen erabilera nahasketa labur-labur azaltzera, adibide baten bidez. Har dezagun **Atetik begiratu eta esan zegoela ikusi zuen* esaldia. Bertan, *esan* eta *esna* hitz-formak nahastu egin dira, baina bi hitzak zuzenak direnez, latza gertatzen da nahasketaren ondorioz sortutako errorea automatikoki detektatzea. Testuingurua behar da hitza egoki erabili den ala ez jakiteko. Kukich-en arabera (1992), "real-word errors" deitutakoak, hots, baliozko hitzetan gertatzen diren erroreak, erroreen % 25-40 izaten dira; beraz, honakoa garrantzi handiko gaia da. *Context-sensitive spelling correction* edo *testuingurua aztertuta ortografiaren zuzenketa* du izena arazo honetaz arduratzen den alorrak.

Errore semantiko eta *hitz zuzenen erabilera nahasketa* kontzeptuen artean diferentzia bat egitekotan, errorea egitearen arrazoi posiblera jo dezakegu. Hitza testuinguru okerrean kokatu bada esanahi zehatza ezagutzen ez delako, *errore semantikoa* dugula aurrean esan dezakegu (adib. *bilatu* vs *topatu*). Tipografia edo tekleatze-arazoak direla eta antzeko hitza sortu bada, edo hitzak homofonoak izateagatik nahastu badira, nahastutako hitzen esanahia ezaguna eta zeharo ezberdina izanik, *hitzen nahasketa* sinple baten aurrean gaudela esango dugu, inolako kutsu semantikorik ez duena (adib. *hura* vs *ura*).

Errore semantikoa egiten denean, esanahi zehatza ezagutzen ez dela esan dugu. James-ek (1998) zentzu-erlazioen nahasketak, i) esanahi orokorragoa duten hitzak zehatzagoen ordez erabiltzea (hiperonimoa, hiponimoaren ordez), ii) termino zehatzegia erabiltzea (hiponimoa, hiperonimoaren ordez), iii) bi hiponimo-anaien artean desegokiena erabiltzea edo iv) hurbileko sinonimoen artean okerra erabiltzea, errore semantiko mota desberdinak kontsideratzen ditu. Horietaz gain, kolokazio-erroreak ere aipatzen ditu. Hau da, esanahiagatik elkarren ondoan ager ez daitezkeen hitzak. Adibidez, *makila okertu* dela esatea egokia da, baina *eguraldia* ezin da *okertu.* Antzekoak dira lekuz alda ezin diren hitz bikoteak. Adibidez, *karakola eta barea* esango dugu eta ez **barea eta karakola.*

Idazleak gaizki erabilitako hitzaren esanahia ezagutzen ote duen ala ez jakitea oso zaila denez, hitz zuzenen erabilera nahasketak eta errore semantikoak detektatzeko eta zuzentzeko teknika berdinak erabiltzen dira. Teknika horiek ikasketa automatikoari dagokion atalean (2.2.2 atalean) aipatuko ditugu. Teknika sinbolikoen azalpen orokorrak amaituta, teknika enpirikoen inguruan arituko gara hurrengo lerroetan.

2.2. Corpusetan oinarritutako teknikak (enpirikoak)

Azken 10 edo 15 urteotan aipatzekoa izan da hizkuntzaren tratamendu automatikoan teknika enpirikoen garapena. Teknika horiek datu kopuru handien erabileran dute oinarria, datuak arakatzeko prozedura estatistikoak erabiltzen direlarik.

Hurbilpen enpirikoen ezaugarri nagusia corpus etiketatuen erabilera da (gure kasuan, testuzko corpusak). Teknika enpirikoetan ataza ezberdinak lantzeko beharrezko hizkuntza-ezagutza corpusean agertzen diren elementuetatik (eta beren maiztasunetatik) erauzten da. Gertaera linguistikoen deskribapen zabala izateko, corpusak tamaina handia behar izaten du.

Erroreen detekzioari dagokionez, corpusean oinarria duten teknikak ez dira gehiegi erabili. Oso zaila da teknika enpirikoak erabili ahal izateko egokia den testu erroredun kopuru handia biltzea. Gainera, testuan topa daitezkeen errore posible guztien etiketatzea ataza konplexua eta garestia da. Hala ere, zenbait abantaila badituzte teknika enpirikoek. Ez dira aldez aurretik ezagutu behar errore posibleak. Testuetan dute isla. Desabantaila moduan esan dezakegu, erroreak detektatu ondoren sistema hauetan normalean ez dela errore-mezurik ematen, horretarako erregela bereziak erabiltzen ez badira, behintzat.

Zenbaitetan Interneteko corpus izugarria erabili da terminoren baten zuzentasuna neurtzeko (Moré et al. 2004). Ideia sinplea da: Interneten oso maiz azaltzen diren terminoak zuzenak dira. Gutxi azaltzen direnekin, ordea, duda izan dezakegu. Hurbilpen honek Interneten argitaratzen diren testuen zuzentasuna suposatzen du. Hipotesia, ordea, ez da erraz neurgarria.

Teknika enpirikoak orokorrean *teknika estatistikoetan* eta *ikasketa automatikoan* sailkatzen dira. Batzuetan, ordea, teknika horien arteko banaketa ez da argia. Teknikak teilakatu egiten dira, bi diziplinetan datuen analisia egiten baita, eta maiz ikasketa automatikoa erabiltzen duten sistemak estatistikarekin konbinatuta erabiltzen baitira.

2.2.1. Metodo estatistikoak

Esan dugun bezala, testu kopuru handia eskura duten hizkuntzetan estatistika erabili ahal da informazio linguistiko aberatsa testuetatik erauzteko. Zenbait gramatika-zuzentzailek metodo estatistikoak baliatzen dituzte erroreen detekziorako atazak garatzeko. Normalean, maila morfologikoan etiketatutako corpus bat erabiltzen da gramatika-kategorien etiketen zerrenda bat eraikitzeko. Etiketa-sekuentzia batzuk (n-grama) oso arruntak dira (adibidez *izena determinatzailea*), eta beste batzuk, oso gutxitan agertzen dira (adibidez, *izena determinatzailea determinatzailea*). Corpusetan maiz gertatzen diren sekuentziak *zuzentzat* joko dira, eta arruntak ez direnak, okertzat. Metodo estatistikoekin honakoak hartu behar dira kontuan: i) Etiketa zehatzen agerpen ezaren arrazoia *datu urritasuna* izan daiteke, eta beraz, egitura ez-ohikoak errore kontsidera ditzakegu, *agerpen urrikoak* besterik ez direnean. ii) n-grametan hitzaren kategoria soilik kontuan hartuz gero, gramatika-kategoria berdineko baina oker erabilitako hitzetan gertatutako erroreak ez lirateke detektatuko. Metodo hauekin alarma faltsu kopuru handia sortzen da. Estatistika hutsezko zuzentzaile bat sortzea zaila da, eta normalean, hauek erregeletan oinarritutako teknikekin hedatzen dira.

Informazio zuzena oinarrian hartu beharrean, *errore-probabilitatea* neurtzen saiatu izan dira zenbait egile (Atwell 1987): erroreetan zein gramatika-etiketa konbina-

zio agertzen den aztertu nahi izan dute. Hurbilpen hau baztertua izan da erroreekin etiketatutako corpus ezagatik.

Estatistika hutsezko sistemei hizkuntza-ezagutza gehituz gero, erroreen detekzioaren doitasuna izugarri hobetzen dela, eta alarma faltsuen kopurua asko jaisten dela, egiaztatu dute zenbait lanek (Bigert eta Knutsson 2002, Sjöbergh 2005).

Azken aldian, hizkuntzaren tratamendu automatikotik aldentzen diren teknikak erabiltzen hasiak dira erroreak detektatzeko. Adibidez, Kumar eta Nair-en lanean (2007) inmunologia-sistema artifizialeko (*Artificial Immune System (AIS)* ingelesez) oinarriak erabiltzen dituzte gramatika-erroreak detektatzeko. Erroreak testuetako patogenoak bailiran tratatzen dituzte, eta antigorputz detektatzaileak eraikitzen dituzte horiek detektatzeko. Hizkuntza-informazioa kategoria gramatikalen bigrametan, trigrametan eta tetragrametan biltzen denez, estatistika oinarria duen lanen artean kokatu dugu honakoa, nahiz eta algoritmoak AIS eremukoak diren. Teknika honekin detekta daitezkeen zortzi errore mota topatu dituzte soilik, eta gramatika-liburu batetik ateratako esaldiak erabili dituzte probarako, hots, laborategiko esaldi motzak.

2.2.2. Ikasketa automatikoa

Ikasketa automatikoaren helburu nagusia, ordenagailuek *ikas* dezaten teknikak garatzea da. Zehazkiago, adibide moduan emandako informazio ez-egituratutik abiatuta, portaera zehatz batzuk garatu nahi dira. Hizkuntzaren tratamendu automatikoari dagokionez, informazio ez-egituratu hori corpusetan egoten da. Corpusetatik datuak jaso eta testuan egon daitezkeen egiturak inferitzen dira datu-meatzaritza izenekoa eginez. Teknika hau alor desberdinetan erabil daiteke, eta horien artean erroreen detekzioan eta, batez ere, hitz zuzenen erabilera nahasketan.

Corpus erroredunak biltzeko, batzuek eskuzko etiketatzea erabili duten bitartean (Izumi et al. 2003), beste batzuek lan hau ekiditeko erroreak automatikoki sortzen dituzte, era berean erroreak etiketatuz (Sjöbergh eta Knutsson 2005). Hitz zuzenen erabilera nahasketaren atazarako ez da testu erroredunik behar, ongi samar idatzitako edozein testu erabil daiteke; beraz, ez da arazo handirik egoten corpusak lortzeko. Hori izan da azken ataza honetan ikasketa automatikoak lortu duen arrakastaren arrazoi nagusia.

Erroreak detektatzeko lanetan erabilitako algoritmoei dagokionez, *transformazioan oinarritutako ikasketa*[6] (Sjöbergh eta Knutsson 2005), *memorian oinarritutako ikasketa*[7] (Örvar Kárason 2005) eta *entropia handieneko eredua*[8] (Izumi et al. 2003) erabiltzen dituzten lanak topatu ditugu. Ikasketa prozesurako, errore mota bakoitzeko ikasketa automatikoko modulu bat entrena daiteke, edo errore mota desberdinetarako modulu bakarra. Lehen hurbilpenak, zuzenketak proposatu ahal izatearen abantaila du.

Hitz zuzenen erabilera nahasketaren atazarako oso teknika egokia da, arazoa desanbiguazio-ataza moduan ikustea erraza baita. Era formalean, honela definitzen da arazoa (Carlson et al. 2001) lanean: sarrerako esaldi bat eta *helburu hitz* bat emanda,

---

[6] Transformation-Based Learning.
[7] Memory-based Learning.
[8] Maximum Entropy (ME) model.

jakin nahi da hitz hori noiz den zuzena edo noiz aldatu behar dugun. Hitzen arteko anbiguotasuna *nahasketa multzoekin* modelatzen da (adibidez, ingelesezko (*among, between*), (*cite, sight, site*)...). *C* = *w*1, ..., *wn* nahasketa multzo batean *wi* hitz bakoitza anbiguoa da multzoko beste hitzekiko. Testuingurua kontuan izanik zein hitz den egokiena aukeratu behar da. Arazo honi buruzkoak dira (Golding eta Roth 1999, Carlson et al. 2001) lanak. Horietan, nahiz eta emaitza onak lortzen diren (% 92-95eko zuzentasuna), lan eremua nahasketa multzo zehatz batzuetara mugatuta dago.

Aipatutako teknikak datu kopuru eskergarekin probatuta, emaitzek bat egiten dutela frogatu da. Hori dela eta, LNPko ikerlarion etorkizuneko lana etiketatutako corpusak sortzera bidera dadin proposatzen dute Banko eta Brill-ek (2001), ikasketa teknikak probatu eta konparatzera baino.

2.3. Tekniken era bilera gramatika-zuzentzaileetan

Gramatika-zuzentzaile bat garatzen hasi aurretik, bi gauza hartu behar izaten dira kontuan:

1. *Landu beharreko hizkuntzaren ezaugarri sintaktikoak zeintzuk diren.* Ezaugarri sintaktikoek zuzentzaileak detektatuko duen fenomeno multzoa zehaztuko dute. Adibidez, euskara moduko hizkuntza eranskari batean, non aditz laguntzaileen osaketa nahiko konplexua den, aditza eta osagai gramatikalen (subjektu, objektu, zehar-objektu) arteko komunztadura oso garrantzitsua izango da; eta suedierarako, adibidez, ez (Arppe 2000).
2. *Zein formalismo linguistiko daukagun eskuragarri hizkuntzaren analisia eta geroago erroreen detekzioa eta zuzenketa egin ahal izateko.* Zerotik hasi nahi ez badugu behintzat, aurretik garatutako tresnak berrerabiltzeak garrantzi handia izango du.

Gramatika-zuzentzaileei gainbegiratu bat eman diegu 1 taulan, horietako bakoitza bi parametroren arabera sailkatuz: i) erroreak detektatzeko erabilitako teknika eta ii) helburu duen hizkuntza.

3. Euskararentzako aplikazioak. Adibide bat

Aipatutako teknikak euskaraz gertatzen diren erroreak detektatzeko erabiltzen direnean topa daitezkeen ezaugarri eta zailtasunak erakusteko, atal hau adibide batekin hasiko dugu.

Lan bat egin dugu Murriztapen Gramatika erabiliz euskarazko postposizio-lokuzioetan gertatzen diren erroreak detektatzeko eta zuzentzeko (Díaz de Ilarraza et al. 2008, Oronoz 2009). Zehazkiago, murriztapen-gramatikako islapen-erregelak erabili ditugu, erroreak automatikoki etiketatzeko. Islapen-erregela hauetan errore-patroiak definitu ditugu, eta lehenago ere aipatu dugun moduan, horiek morfosintaxi-mailan analizatutako egiturei aplikatu dizkiegu. Lortu dugun sistemaren baliagarritasuna neurtzeko, erroreak postposizio-lokuzio zuzenetan, postposizio-lokuzio erroredunetan eta postposizio antzeko egituretan probatu ditugu.

Landu dugun egitura linguistikoa —postposizio-lokuzioa— aipuak erabiliz zehaztuko dugu:

Hizkuntza	Gramatika-zuzentzaileak eta erabilitako teknikak							
	Sinb.						Enp.	
	Egoera finituak			Gramatikak				
				TGG				
	MG	XFST	Best.	+ Erlax.	+ Errore gram.	Mend.	Estat.	IA
Katalana (Badia et al. 2004)	✓	—	—	—	—	—	—	—
Suediera *Grammatifix* (Arppe 2000) (Birn 2000) (Hashemi et al. 2003) *Granska* (Carlberger et al. 2002) (Domeij et al. 1999) *Scarrie* (Paggio 2000) (Bigert eta Knutsson 2002) (Sjöbergh 2005) (Sjöbergh eta Knutsson 2005)	✓ — — — — — —	— ✓ — — — — —	— — ✓ — — — —	— — — ✓ — — —	— — — — — — —	— — — — — — —	— — — — ✓ ✓ —	— — — — — — ✓
Islandiera (Örvar Kárason 2005)	—	—	—	—	—	—	—	✓
Norvegiera (Johannessen et al. 2002)	✓	—	—	—	—	—	—	—
Espainiera *Gram Check* (Bustamante eta León 1996)	—	—	—	✓	—	—	—	—
Italiera *JDII* (Bolioli et al. 1992)	—	—	—	✓	—	—	—	—
Brasilgo portugesa *ReGra* (Teixeira Martins et al. 1998)	—	—	—	✓	—	—	—	—
Arabiera *Arabic Gram Check* (Shaalan 2005)	—	—	—	✓	—	—	—	—
Euskara (Gojenola 2000)	—	—	—	✓	—	—	—	—
Nederlandera (Vosse 1992)	—	—	—	✓	—	—	—	—
Txekiera (Holan et al. 1997)	—	—	—	—	—	✓	—	—
Errusiera (Mitjushin 1996)	—	—	—	—	—	✓	—	—
Frantsesa (Courtin et al. 1991)	—	—	—	—	—	✓	—	—
Ingelesa (Naber 2003) (Foster eta Vogel 2004) *Abi Word* (Brehony eta Ryan 1994) (Atwell 1987) (Chodorow eta Leacock 2000) (Cucerzan eta Brill 2004) (Kumar eta Nair 2007) (Izumi et al. 2003)	— — — — — — — — —	— — — — — — — — —	✓ — — — — — — — —	— — — — — — — — —	— ✓ — — — — — — —	— — ✓ ✓ — — — — —	— — — — ✓ ✓ ✓ ✓ —	— — — — — — — — ✓
Ingelesezko nahasketa multzoak (Mangu eta Brill 1997) (Golding eta Roth 1999) (Carlson et al. 2001)	— — —	— — —	— — —	— — —	— — —	— — —	— — —	✓ ✓ ✓

1 taula

Gramatika-zuzentzaileek erabiltzen dituzten teknikak

> Egungo ikuspegitik, bada, postposizioen artean bi mota nagusi bereiz daitezke: alde batetik lemari erantsirik ageri diren atzizkiak (1), eta bestetik, osagai nagusia elementu beregaina dutenak (2).
>
> (1) etxe*tik* (2) etxe*aren gainetik*

Zabala eta Odriozola-ri (2004) jarraiki lehenengo multzoko postposizioei *atzizki-postposizioak* deituko diegu *(posposiciones sufijales)*. Bigarren multzokoak, ostera, *post-posizio-lokuzioak* direla esango dugu.

Euskaltzaindiaren arabera:

> Postposizio diren neurrian kasu markaren bat hartzen dute [...] Kasu marka hau aldakaitza izan ohi da.
>
> (3) Elizaren *bitartez* (4) Astearen *buruan*
>
> Ezin da esan **elizaren bitartean, *elizaren bitarteari* edo **astearen buruaz.*
>
> (Euskaltzaindia 1985: 438).

Aipuaren amaieran aipatzen diren moduko erabilera okerrak dira aztertuko ditugunak. Erabilera oker horien deskribapena errazte aldera, nahiz eta unitate osoa postposizio-lokuziotzat jo (ikus 3 adibidea), *postposizio-atzizki* (PA) (*-ren*) eta *osagai beregain* (OB) (*bitartez*) terminoak zehaztuko ditugu tokenei erreferentzia egiteko.

Postposizio-lokuzio egitura nola ulertu dugun azaldu ondoren, ikus dezagun egiturotan azaldu ohi den erroreetako bat detektatzeko modua adibide batekin.

3.1. Adibidea: postposizio-lokuzioetako erroreen detekzioa patroi-erregelak baliatuz

Postposizio-lokuzioetan gertatzen diren erroreak detektatzea ez da berehalako lana. Zailtasun handiak topatu ditugu egituron anbiguotasuna dela kausa, eta baliabide ugari erabili behar izan dugu lana aurrera eraman ahal izateko. Ikus dezagun adibide moduan *buruz* hitz-forma osagaitzat duen errore bat detektatzeko erabilitako islapen-erregela (ikus 1 eta 3 irudiak). *Buruz* postposizio-lokuzioaren erabilera zuzenak "*-ra/-i -Ø/-ko*" eskema jarraitzen du. Hots, postposizio-atzizkian adlatibo (*-ra*) edo datibo (*-i*) marka du, eta *buruz* osagai beregaina markarik gabe, edo lekuzko genitibo (*-ko*) markarekin agertzen da. Adibidez, *oihanera buruz abiatu zen edo liburu honi buruz mintzatu gara* egiturak zuzenak dira (Euskaltzaindia 1985), baina ez, **bidaia buruz hitz egin genuen* egitura. Azken adibideko moduko egitura okerrak detektatu nahi ditugu, eta horretarako, hasiera batean 1 irudiko erregela definitu genuen:

```
LIST POSTPOSIZIOAK-BURUZ = «buruz»;
MAP (¡POSTBURUZ) TARGET IZE-DET-IOR IF (NOT 0 DAT-ALA)
                                        (1 POSTPOSIZIOAK-BURUZ + ADB);
```

1 Irudia

*Buruz*en erabilera oker bat deskribatzen duen islapen-erregela baten lehen hurbilpena

Hau da, idatzi "{POSTBURUZ" etiketa, baldin eta izena, determinatzailea edo izenordaina kategoriako uneko hitzak (TARGET IZE-DET-IOR) ez duen datibo edo adlatibo postposizio-atzizkirik (NOT 0 DAT-ALA), eta gainera, ondorengo hitza (1 posiziokoa), buruz adberbioa den (1 POSTPOSIZIOAK-BURUZ + ADB).

```
0. «<bidaia>»
        "bidaia"IZE ARR DEK ABS MG AORG {POSTBURUZ
        "bidaia"IZE ARR DEK ABS NUMS MUGM AORG {POSTBURUZ
        "bidaia"IZE ARR ZERO AORG {POSTBURUZ
        "bidaiatu"ADI SIN AMM ADOIN NOTDEK
1. "<buruz>"
        "buru"IZE ARR DEK INS MG
        "buruz"ADB ARR ZERO
```

2 Irudia

Bidaia buruz egituraren analisi morfosintaktikoa, errore-etiketekin

1 irudiko islapen-erregelak, 2 irudiko analisian ongi markatu du errorea, baina beste hainbat adibideren azterketak erakutsi digu erregela *orokorregia* dela: 2 irudiko analisian errorea detektatzen duen era berean ipin lezake errore-etiketa *bidaia buruz ari zitzaigun kontatzen* egitura zuzenean. *Buruz* hitza tartean duten eta postposizio ez diren esamolde/egitura zuzen ugari daude, eta horiek ezin ditugu erroretzat jo. Honela landu ditugu ekidin nahi ditugun egiturak:

— Esamoldean aditza lagun duten egiturak (adibidez, *buruz erori/jo, buruz ikasi/hartu, buruz jokatu, buruz ari izan...*) BURUZ-ADI zerrendan bildu ditugu. Aditza, buruz hitzaren aurretik ala atzetik ager daiteke perpaus berean eta ez du zertan ondoan agertu (adibidez, *kanta buruz eta letraz letra ikasi du,*[9] edo *ez ikasi testua buruz*). Aipatutako esamoldeak zuzenak dira, eta errorea detektatzera goazenean, ez daitezela inguruan ageri eskatu dugu 3 irudiko erregelan. Horretarako, BURUZ-ADI zerrendako elementuak ez daitezela agertu buruz OBaren eskuineko edozein posiziotan (NOT * 1 BURUZ-ADI), ezta ezkerreko edozein posiziotan ere (NOT *-1 BURUZ-ADI) zehaztu dugu.
— Aditza osagai ez duten gainontzeko esamoldeetan *buruz* hitzaren kidea haren ondoan azaltzen da eskuineko aldean (*buruz buru* eta ez **buru buruz*). BURUZ-GAIN zerrendan bildutakoek ez dutela buruz hitzaren ondoan agertu behar (NOT 2 BURUZ-GAIN) zehaztu dugu.
— Futbolarekin lotutako zenbait egitura ere ekidin behar izan ditugu alarma faltsuen kopurua jaisteko. *baloi, gol, sare* eta antzeko lemak BURUZ-FUTBOLA zerrendan bildu ditugu eta analizatu beharreko testuetan agertzen direnean, alde batera utzi ditugu.
— Alarma faltsuak ekiditeko definitutako baldintza hauetaz gain, uneko hitza lexikorik gabe (guesserraren bidez) analizatua ez izatea (NOT 0 EZEZAGUNA) eskatu dugu.

[9] **kanta buruz hitz egin* postposizio-egitura, okerra litzateke.

Ikusi dugun moduan, errore-patroiak definitzen dituguneanean, postposizio erroredunaz gain, erroretzat jo daitezkeen egitura zuzenak izan behar ditugu kontuan, baita postposizio ez direnak ere. Alarma faltsuak ekiditeko ezarritako baldintzek, ordea, egitura erroredunen bat ez detektatzea eragin dezakete. Adibidez, **baina hori buruz zerbait esatea diote* egitura erroreduna ez genuke detektatuko, *buruz . . . esan* esamoldea topatu dugulako. Sistemak detektatzen duen errore kopuruaren eta ematen duen alarma faltsu kopuruaren arteko oreka mantentzen saiatu bagara ere, erabaki bat hartu dugu horien arteko erlazioari dagokionez: gure ustez egokiagoa da egitura erroredun bat ez detektatzea, alarma faltsu bat sortzea baino. Erabaki honek badu bere arrazoia: erabiltzaile askori ordenagailuak hutsik gabekoak direla iruditzen zaionez, alarma faltsuek zalantzan ipin dezakete gramatikalki zuzena denari buruzko duten ezagutza.

```
LIST BURUz-ADI = “erori” “jo” “ikasi” “hartu” “jokatu” “ari_izan” “esan” “errematatu”
                 “jakin” “gogoratu”;
LIST BURUz-GAIN = “behera” “belarri” “ buru” “gain” “gora”;
LIST BURUZ-FUTBOLA = “falta” “baloi” “ jokalari” “ gol” “futbol” “sare” “zutoin” ...;
LIST POSTPOSIZIOAK-BURUZ = “buruz”;
MAP ({POSTBURUZ) TARGET IZE-DET-IOR IF (NOT 0 DAT-ALA)
                                        (NOT 0 EZEZAGUNA OR BURUZ-ADI OR BURUZ-FUTBOLA)
                                        (1 POSTPOSIZIOAK-BURUZ + ADB)
                                        (NOT 2 BURUZ-GAIN)
                                        (NOT *1 BURUZ-ADI OR BURUZ-FUTBOLA)
                                        (NOT *-1 BURUZ-ADI OR BURUZ-FUTBOLA);
```

3 Irudia

*Buruz*en erabilera oker bat deskribatzen duen islapen-erregela

Erroreak detektatzeko egitura zuzenak uneoro buruan eduki behar ditugula ikusi dugu, alarma faltsuak ekidin nahi baditugu behintzat. Azkenak lantzeko, gramatiken garapenean corpus erroredunak erabiltzeaz gain, errorerik gabeko corpusak ere erabili behar direla uste dugu.

Ondoren azalduko ditugun esperimentuetan, erroreen detekzioaren eta alarma faltsuen ekiditearen arteko konpromisoa mantendu dugu, detekziorako teknika edozein izanda ere.

3.2. Euskararentzako aplikazioak

Euskarazko erroreak detektatzeko orain arte garatu diren sistema gehienek hizkuntza-ezagutzan oinarritutako teknikak (sinbolikoak) erabili dituzte. Gure ustez, teknikaren aukeraketa guztiz lotuta dago landu nahi den hizkuntza-egiturarekin. Gehiago ere esango dugu: hizkuntza-egitura desberdinak lantzeko, teknika ezberdinak erabili behar dira, errore mota bakoitzarentzako teknika/tresna bat egokiagoa baita.

Gure ustez, elementuen galerarekin, ordezkapenarekin eta gehitzearekin zerikusia duten erroreak lantzeko corpusetan oinarritutako teknikak egokiak dira. Osagai lin-

guistikoen edo horien ezaugarrien arteko adostasuna edo komunztadura neurtu nahi denean, ordea, teknika sinbolikoak egokiagoak dira. Zehazte aldera esango dugu, gramatika-zuzentzaileei gainbegiratua eman diegunean, detektatutako erroreen eta erabilitako tekniken artean harremana topatu dugula: horrela, sintagma-mailako erroreak detektatzeko patroi-erregelak erabili ohi dira (MG, XFST formalismoekin, adibidez), esaldi-mailakoak lantzeko analisi-zuhaitzak (TGGak, mendekotasun-zuhaitzak...) eta hitz zuzenen erabilera nahasketak tratatzeko, teknika sinbolikoak.

Jarraian aztertuko ditugun sistemetan eredu hau bete-betean betetzen da: koma puntuazio-ikurraren galera edo leku-aldaketa tratatzeko ikasketa automatikoa erabili da, teknika enpirikoa, beraz. Datetan, postposizio-lokuzioetan eta determinatzaileetan gertatzen diren erroreak detektatzeko patroi-erregelak erabili dira; eta komunztadura lantzeko, mendekotasun-zuhaitzen gainean aplikatutako patroi-erregelak.

Hona hemen labur-zurrean sistema horien deskribapen bat erabiltzen duten teknika motaren eta aztertzen duten fenomeno linguistikoaren arabera antolatua:

Puntuazio-erroreak (teknika enpirikoak)

— Koma. Lan honek xuxeng gramatika-zuzentzailearen estilo- eta puntuazio-erroreen detekziorako atala osatzea du helburu, horretarako ikasketa automatikoko teknikak erabiltzen dituelarik (Alegria et al. 2006). Koma puntuazio-marka lantzen du bereziki, arrunki markarik anbiguoena eta gutxien landutakoa. Gainera, nahiz eta euskaraz puntuazio-markei buruzko oinarrizko arau batzuk definituak diren, ez da ezer arautu komaren erabilerari buruz. Hortaz, lehen pausoa euskararako komaren erabilerari buruzko teoriak aztertu eta nolabaiteko dokumentua osatzea izan da. Koma-zuzentzailea garatzeko haien artean konbina daitezke bi bide hartu dira: perpausen mugetan oinarritzea edo/eta corpusean oinarritzea. Ideia nagusia honakoa da: esaldia perpausetan banatzea lortuz gero, erraza izan daiteke komak ipintzeko lekua adieraziko duten erregelak idaztea. Hots, lehenengo perpausak ezagutzeko tresna garatu behar da. Hori faltan, corpusetan oinarritzea erabaki da. Trebatzeko corpustzat *Euskaldunon Egunkaria* hartu da, "komaren teoria" hobekien betetzen zuena baitzen. Ikasketarako hiru algoritmo (*Naive Bayes, Support Vector Machine* eta erabaki-zuhaitzak) eta ahalik eta atributu kopuru (informazio linguistiko) handiena erabilita, honako emaitzak lortu dira: trebatze-lanetarako 100.000 hitzetako corpus bat erabiliz, *komarik ez ipintzeko* atazan % 96ko doitasuna eta % 98ko estaldura, eta *koma ipintzeko* atazan % 70eko doitasuna eta % 49ko estaldura.

Sintagma-mailako erroreak (teknika sinbolikoak)

— Datak. Datak ez dira egitura anbiguoak, *'[leku-izena,]*[10] *urtea hilabetea eguna'* katea topatuz gero, data dugu parean. Datetako gramatika-erroreak detek-

[10] [ ] ikurrek leku-izena aukerazkoa dela adierazten dute.

tatzeko eta zuzentzeko, egoera finituko transduktoreak erabili ditugu (Díaz de Ilarraza et al. 2007, Oronoz 2009). Erroreetako ezaugarri eta murriztapenak zehaztu ditugu *Xerox Finite State Tool* (XFST) tresnak gramatikak definitzeko duen formalismoa erabiliz, eta honek automata eta transduktore bihurtu ditu. Automata horiek daten analisi morfosintaktikoari aplikatu zaizkio. Sistema bi egoera finituko transduktore multzok osatzen dute, batak erroreak detektatzen ditu eta besteak data zuzenak sortzen ditu. Sistemak euskaraz daten inguruan maizen egiten diren errore motak lantzen ditu. Hots, leku-izenak inesibo kasua izanik eguna adierazten duen zenbakiak inesibo kasurik ez izatea (**Donostia**[n]**, 2007ko maiatzaren 27a[]*), urtea adierazten duen zenbakia deklinatzeko marratxoa erabiltzea (**Donostian, 1995[-]eko maiatzaren 14an*), hilabeteak genitibo kasua izanik, egunak absolutibo kasua izatea (**Donostian, 1995eko maiatzaren 22[]*)...

Lortu ditugun emaitzak oso onak izan dira, 411 esaldirekin osatutako garapenerako corpus batean % 95,9ko estaldura eta % 97,8ko doitasuna lortu baitugu. Probarako corpusean (247 esaldi), % 92,1eko estaldura eta % 89,7ko doitasuna lortu ditugu.

— DETERMINATZAILEAK. Murriztapen Gramatika erabili da euskarazko determinatzaileekin zerikusia duten erroreak detektatzeko. Landutako erroreak besteen artean (guztira 8 kategoria) errore-kategoria hauetakoak dira: determinatzaileen galera (adib., **kotxe[] erosi nuen*), determinatzaileen errepikapena (adib., *Euskal Herria nazioa bat da), determinatzaileen ordena okerra (adib., **asko lagunak joan ziren, -a* organikoaren ezabaketa... Ikasleen testuekin osatutako 113.290 hitzetako corpus batean, 788 determinatzaile-errore eskuz etiketatu dira. Bost errore-kategoria lantzea erabaki da, eta horretarako, 85 patroi-erregela idatzi dira MG erabiliz. Ebaluazioan % 45,45eko doitasuna eta % 44,78ko estaldura lortu dira corpus erroredunean. Alarma faltsu kopuru handia sortu da hitz ezezagun, analisi oker eta egitura arraroak direla eta. Horiek baztertuta, doitasuna % 80ekoa litzateke.

— POSTPOSIZIO-LOKUZIOAK. Prozesu luze baten ondoren, euskarazko bost postposizio-lokuzio erabilienak aukeratu ditugu (*arte, aurre, bitarte, buruz* eta *zehar*), haietan gertatzen diren erroreak MG erabiliz detektatzeko eta horien ordain zuzenak emateko (Díaz de Ilarraza et al. 2008, Oronoz 2009). Egituren anbiguotasun handia dela eta, sintaxi-erroreak ezagutu ahal izateko besteen artean semantika erabili behar izan dugu. Adibidez, *arte* osagai beregaina duen postposizioaren erabilera aldatu egiten da lekuari edo denborari buruz ari garenean: *gero arte* esan dezakegu, baina ez, **etxea arte* (egokiagoa da *etxeraino*).[11] Erroreen deskribapenerako leku/denbora ezaugarri semantikoak lortzeko, Lersundik bere tesi-lanean (Lersundi 2005) Euskal Hiztegiko definizioetatik erauzitako *genusak*[12] baliatu ditugu. Semantikaren beharra izan dugu era be-

[11] EGLU-1 (395): «Has gaitezen lehenetik. Batak, *-raino*-k, lekua adierazten duela eta hainbestez leku izenekin gehienbat erabili behar dela esan izan da gramatika guztietan; bigarrenak, *arte*-k, aldiz, denbora adierazten duela eta denborazko adberbio eta sintagmekin erabili behar dela batik bat. [...] Bi atzizki horien gurutzaketak askotxo dira.»

[12] Genus: definizioaren gune sintaktiko eta semantikoa.

rean *aurre* OBa duen postposizio-lokuzioarekin. Kasu honetan, bizidun/bizigabeak bereizi behar izan ditugu *etxe aurrean* modukoak ontzat emateko, eta **zu aurrean* erakoetan errorea detektatzeko.

Batzuetan gramatiketan zuzentzat jotzen diren egiturak ("-Ø arte*an*") dituzten errore-adibideak topatu ditugu: *proposamen artean, beste arraza artean...* Beste batzuetan, zehaztapen falta topatu dugu: "-en artean" zuzena omen da (adib. *zuhaitzen artean*), baina guk **Mendiaren artean bezala biziko gara* egitura okertzat jo dugu; gramatika-liburuetan, egitura hori postposizio-atzizkian pluraleko mugatasun ezaugarria duten hitzekin soilik dela zuzena zehaztu beharko litzateke.[13]

Postposizio-lokuzioetan erroreak detektatzeko 30 erregela idatzi ditugu, horietatik erdia *arte* postposizioa lantzeko eta 2 soilik *zehar* lantzeko. Hau horrela izanik ere, sistemak *arte* postposizio-lokuzioarekin lortu ditu doitasunari dagokionez emaitzarik okerrenak, eta *zehar* postposizioarekin doitasun hoberena. Erregelen eraginkortasuna testuinguruko baldintzen konplexutasunaren eta anbiguotasunaren araberakoa dela ondorioztatu dugu.

Ebaluazioa corpus erroredun orokor batekin, eta corpus zuzen batekin egin dugu. Zuzenean alarma faltsuak gertatzeko aukera askoz ere handiagoa da, oso errore gutxi baitaude, beraz, emaitzak ez dira oso onak izan (% 40ko doitasuna garapenerako corpusean eta % 42koa probako corpusean).[14] Corpus erroredunean % 96 doitasuna eta % 83ko estaldura lortu ditugu garapenerako atalean, eta % 67ko doitasuna eta % 34ko estaldura probarakoan.

— Bestelakoak. xuxeng euskarazko gramatika-zuzentzailearen lehen bertsioa dagoeneko garatua dago eta Microsoft Word testu-prozesadorean integratua (Otegi 2006). Lehen prototipoak informazio linguistikorik gabe detekta zitezkeen estilo- eta puntuazio-erroreak (puntu eta komaren ondorengo espazioak, esaldi luzeegiak...) detektatzen zituen, baina gaur egungo prototipoak hizkuntza-informazioa lortzeko beharrezko tresnak integratuta dauzka. Hau horrela, sintaxi-mailako hainbat errore detektatzeko gai da. Besteak beste, postposizio-lokuzioetakoak, determinatzaileetakoak, *hilabete bat, 5 urte bete, ez du ezer ez egin* eta antzekoak.

Esaldi-mailako erroreak (teknika sinbolikoak)

— Komunztadura patroi-erregelekin. Perpaus-mailako komunztadura-erroreak detektatzeko (subjektuaren, objektuaren eta zehar-objektuaren komunztadura aditz laguntzailearekin numeroan eta kasuan) *Saroi* izeneko tresna erabili dugu (Oronoz 2009). *Saroi*ren kontsulta-lengoaia erabiliz, mendekotasun-zuhaitzetako egitura eta ezaugarriei buruzko kontsultak egin ditzakegu. Kontsulta horietan, komunztadura-erroreek eduki ditzaketen egiturak deskribatzen direnean *Saroi*k erroreak detektatzeko balio du. Helburu nagusia, corpusetako

[13] Segur aski, munduaren ezagutza suposatzen delako ez da zehazten ezaugarria, baina datu hau erroreak detektatzeko ezinbestekoa da.

[14] Corpus zuzena oso handia denez, ez ditugu errore guztiak etiketatu, beraz, ezin izan dugu estaldura zehaztu.

fenomeno linguistikoen analisia du. Adibidez, 4 irudian *Zentral nuklearrak zakar erradiaktiboa eratzen dute* komunztadura-errorea detektatuko lukeen kontsulta-erregela ikus dezakegu. Erregela, azken esaldiaren analisiaren ondorioz sortzen diren mendekotasun-zuhaitz guztiei aplikatuko zaio (ikus horietako bat 4 irudiaren eskuinaldean).

Esaldi bat analizatzen denean, esaldi horri dagozkion hainbat mendekotasun-zuhaitz lortzen da. Erregela bakoitza, zuhaitz guztietan aplikatzen da. Analisi prozesuan zehar analisi-erroreak metatzen direnez, honakoa erabaki dugu: esaldi batean komunztadura-errorea gertatu dela esango dugu, komunztaduraren detekziorako erregela baten baldintzak esaldiko zuhaitz guztietan betetzen badira.

Esan beharra daukagu euskaraz perpauseko subjektua, objektua eta zehar objektua elidituak egon daitezkeenez, aditz laguntzaileko ezaugarriak horietan bilatzea erabaki dugula, eta ez alderantziz. Hona hemen, detektatutako errore mota batzuk:

- Aditzen azpikategorizazioa erabiliz detekta daitezkeenak:

  (5) **Eta orduan nire amak esnatu zidan eta konturatu nintzen lotan nengoela.*

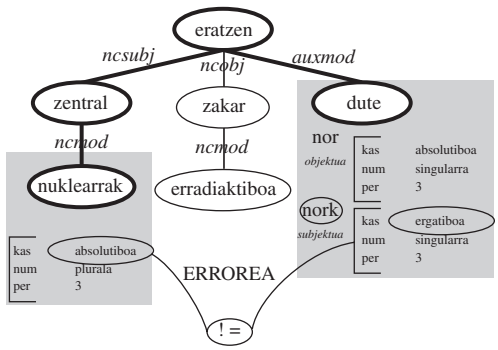
- Behar ez direnean, objektuaren edo zehar-objektuaren agerpena aztertzen dutenak:

  (6) **Erabiltzaileari urruneko beste ordenagailuekin konektatzeko aukera ematen du.*

- Perpauseko funtzio gramatikaletan eta aditz laguntzaileko komunztadura-marketan agertzen diren ezaugarriak konparatuaz ezagut daitezkeenak:

  (7) **Zentral nuklearrak zakar erradiaktiboa eratzen dute.*

```
KOM_SUBJ_KAS_NOR_NORK
(
Bilatu (
@ !ncsubj!ncmod- &
@ !auxmod.paradigma == "NOR-NORK" &
@ !ncsubj!ncmod.kas != @!auxmod.nork.kas)
Ordeztu (
(@!ncsubj !ncmod.kas := @!auxmod.nork.kas)
#  Zentral nuklearrEK zakar erradiaktiboa
   eratzen DUTE.
)
Info (Subjektuak eta aditz laguntzaileak
   ez dute kasuan komunztatzen.)
)
```



4 Irudia

Ezkerraldean, komunztadura-erroreen detekziorako erregela bat.
Eskuinaldean, zein mendekotasun-zuhaitzetan detektatzen duen errorea

4. Ondorioak

Erroreak detektatzea, haien diagnosia egitea eta ordain zuzen posibleak ematea, ezinbestekoa da ortografia- eta gramatika-zuzentzaileetan, eta garrantzitsua, OLHI sistemetan. Hizkuntzaren tratamendu automatikoaren alorrean erroreak detektatzeko eta zuzentzeko erabiltzen diren tekniken eta hauek erabiltzen dituzten sistemen errepasoa egin dugu artikulu honetan (informazio gehiago nahi izanez gero, jo Oronozen 2009ko tesi-txostenera).

Teknikak, bi multzotan banatu ditugu: alde batetik hizkuntza-ezagutzan oinarritutako teknikak edo teknika sinbolikoak daude, zeintzuetan hizkuntzari buruzko ezagutza erregeletan edo beste adierazpideetan kodetzen den modu esplizituan (gogoratu 2.1 atala). Beste alde batetik, corpusetan oinarritutako teknikak edo teknika enpirikoak daude (2.2 atala). Teknika sinbolikoak egoera finituko makinetan eta testuingururik gabeko gramatiketan sailkatu ditugu. Teknika enpirikoek informazio iturria corpusetan duten prozedurak erabiltzen dituzte: estatistika eta ikasketa automatikoa.

Zenbait ondoriotara iristeko parada izan dugu azterketa bibliografikoa egin ondoren:

— Corpusen tamainari dagokionez, corpus gero eta handiagoak lortzea erraza denez, azken urteotako joera corpusetan oinarritutako teknikak erabiltzea da. Gizakiok egindako erregelak ez omen dira inoiz gai izango hizkuntzak duen konplexutasuna jasotzeko. Corpus handiak etiketatuz gero, ordea, testuetatik jasoko dugu behar dugun ezagutza. Corpusen etiketatzean izaten dute teknika enpirikoek zailtasun handiena. Lan hori egina duten hizkuntzetan, ingelesean adibidez, teknika erabilerrazak eta egokiak dira. Euskara modukoetan, berriz, informazio zuzena duen corpusa biltzea zaila bada, interesatzen zaigun atazari dagokion erroredun corpusa biltzea, are neketsuagoa da.
— Ezagutza linguistikoari dagokionez, egindako azterketatik ondorioztatu dugu corpusetan oinarria duten teknikek lema eta batzuetan kategoria gramatikala besterik ez dutela erabiltzen. Teknika sinbolikoak oinarrian dituzten sistemek, berriz, informazio linguistiko kopuru handiagoa erabiltzen dute maiz. Zehatzagoak izateko, sintagmei buruzko informazioa (izen- eta aditz-kateak), eta zenbait kasutan unitate lexikoen arteko erlazioei buruzko informazioa (funtzio sintaktikoak eta beren arteko erlazioak, mendekotasun-erlazioak kasu) erabiltzen dituzte. Azterketatik ateratako beste ondorio bat honakoa da: ezagutza linguistikoa erabiltzeak sistemen doitasuna izugarri hobetzen du.
— Emaitzei dagokionez, aztertutako lan bibliografikoetan maiz ez da ebaluazioari buruzko daturik ematen. Teknika enpirikoak deskribatzen dituzten lanetan zenbait emaitza ematen dira, baina ez, estaldura eta doitasuna; beraz, zaila da batzuetan gainerako sistemekin konparazioa egitea. Teknika sinbolikoak deskribatzen dituzten lanetan, berriz, neurri horiek erabiltzea ohikoagoa da. Gainbegiratutako sistemek testuinguru mugatuko erroreekin (sintagma barruko komunztadura, adibidez) era askotariko emaitzak lortu dituzte. Emaitzarik onenak aldez aurretik eskuz tratatutako laborategiko esaldi motzekin lortu dira (Shaalan 2005). Sintagma-mailako erroreak detektatzeko testu errealak erabili dituzten sistemen artean emaitzarik onenak MG erabiltzen du-

tenek lortu dituzte, % 70 inguruko doitasunarekin (Birn 2000, Johannessen et al. 2002). Testuinguru zabaleko erroreak landu dituzten sistemek (subjektu-aditza komunztadura, adibidez) orokorrean ez dute emaitzarik eman.

— *Errore motei dagokionez,* erroreak hitz zuzenen nahasketarekin edo gabeziarekin zerikusia dutenean teknika enpirikoak oso egokiak direla ikusi dugu. Sintagma-mailako erroreak patroi-erregelekin (MG, XFST tresnekin, adibidez) lantzea ohikoa dela ondorioztatu dugu; esaldi-mailako erroreak detektatzeko, ordea, osagaien arteko lotura egiteko hizkuntza-egitura konplexuak, analisi-zuhaitzak, esaterako, behar direla uste dugu.

Azken ondorioan aipatutakoak, bat datoz euskararako garatutako sistemetan erabili dugun eskemarekin. Komaren falta edo leku aldatzea teknika enpirikoekin landu da. Sintagma-mailako erroreak, hots, datetan, postposizio-lokuzioetan, determinatzaileetan eta bestelakoetan gertatzen direnak, batez ere Murriztapen Gramatika erabiliz detektatzen dira. Horietako asko, XUXENg gramatika-zuzentzanean integratuta daude dagoeneko. Esaldi-mailako erroreak detektatzeko, komunztadura, esaldien analisi-zuhaitz osoa behar izan dugu.

Erreferentziak

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Urkia, M., Maritxalar, M. eta Sarasola, K., 1992, «XUXEN: A spelling checker/corrector for Basque based on two-level morphology». *Proceedings of ANLP'92,* Italy.

Aho, A.V., Sethi, R. eta Ullman, J. D., 1985, *Compilers: Principles, Techniques, and Tools.*

Aït-Mokhtar, S. eta Chanod, J. P., 1997, «Incremental finite-state parsing». *Proceedings of the fifth conference on Applied Natural Language Processing* 72-79, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alegria, I., Arrieta, B., Díaz de Ilarraza, A., Izagirre, E. eta Maritxalar, M., 2006, «Using machine learning techniques to build a comma checker for Basque». *Proceedings of Coling-ACL* 1-8, Sydney. Australia.

Arppe, A., 2000, «Developing a Grammar Checker for Swedish». *Proceedings from the 12th Nordiske datalingvistikkdager, Department of Linguistics,* Norwegian University of Science and Technology (NTNU). Nordgard.

Atwell, E., 1987, «How to detect grammatical errors in a text without parsing it». *Proceedings of The 3rd Conference of the European Chapter of the Association for Computational Linguistics* 38-45, Copenhagen, Denmark.

Badia, T., Gil, A., Quixal, M. eta Valentín, O., 2004, «NLP-enhanced error checking for Catalan unrestricted text». *Proceedings of the fourth international conference on Language Resources and Evaluation*, LREC, Portugal.

Banko, M. eta Brill, E., 2001, «Scaling to very very large corpora for natural language disambiguation». *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* 26-33, Morristown, NJ, USA. Association for Computational Linguistics.

Bigert, J. eta Knutsson, O., 2002, «Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge». *Romand 2002 (Robust Methods in Analysis of Natural language Data),* Frascati, Italy.

Birn, J., 2000, «Detecting grammar errors with Lingsoft's Swedish grammar-checker», *Proceedings from the 12th Nordiske datalingvistikkdager,* Department of Linguistics, Norwegian University of Science and Technology (NTNU), December 9-10. Nordgard.

Bolioli, A., Dini, L. eta Malnati, G., 1992, «JDII: Parsing Italian with a Robust Constraint Grammar». *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92),* 1003-1007.

Brehony, T. eta Ryan, K., 1994, «Francophone stylistic grammar checking (FSGC) using Link Grammars», *Computer Assisted Language Learning* 7(3).

Bustamante, F. R. eta León, F. S., 1996, «Gramcheck: a grammar and style checker», *Proceedings of the 16th conference on Computational linguistics* 175-181, Morristown, NJ, USA. Association for Computational Linguistics.

Carlberger, J., Domeij, R., Kann, V. eta Knutsson, O., 2002, *A Swedish grammar checker.* Unpublished. Submitted 2002 Association for Computational Linguistics.

Carlson, A. J., Rosen, J. eta Roth, D., 2001, «Scaling up context-sensitive text correction». In Hirsch, H. eta Chien, S. (arg.), *Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference (IAAI-01)* 45-50, Menlo Park CA, august 7-9. American Association for Artificial Intelligence.

Chodorow, M. eta Leacock, C., 2000, «An unsupervised method for detecting grammatical errors», *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics* 140-147, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Courtin, J., Dujardin, D., Kowarski, I., Genthial, D. eta Strube de Lima, V. L., 1991, «Towards a complete detection/correction system», *International Conference on Current Issues in Computational Linguistics* 158-173.

Cucerzan, S. eta Brill, E., 2004, «Spelling correction as an iterative process that exploits the collective knowledge of web users». *Proceedings of Empirical Methods in Natural Language Processing* 293-300.

Dale, R., 2000, «Symbolic approaches to Natural Language Processing». In Dale, R., Moisl, H. eta Sommers, H. (arg.), *Handbook of Natural Language Processing*. Marcel Dekker Inc.

DeSmedt, W., 1995, «Herr Komissar: An ICALL conversation simulator for intermediate German». In Holland, V., Kaplan, J. eta Samsm M. (arg.), *Intelligent Language Tutors: Theory Shaping Technology* 371-381.

Díaz de Ilarraza, A., Gojenola, K. eta Oronoz, M., 2008, «Detecting Erroneus Uses of Complex Postpositions in an Agglutinative Language», *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008),* Manchester, 18-22 August.

—, —, —, Otaegi M., eta Alegria, I., 2007, «Syntactic error detection and correction in date expressions using finite-state transducers», *Proceedings of Finite-State Methods and Natural Language Processing (FSMNLP07),* Postdam, Germany, 14-16 of September.

Domeij, R., Knutsson, O., Carlberger, J. eta Kann, V., 1999, «Granska- an efficient hybrid system for Swedish grammar checking». *Proceedings of the 12th Nordiska Datorlingvistikdagarna (NoDaLiDa),* Trondheim, Norway.

Dulay, H., Burt, M. eta Krashen, S., 1985, *Language Two.* Oxford U. P.

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

Foster, J. eta Vogel, C., 2004, «Parsing ill-formed text using an error grammar», *Artificial Intelligence Review* 21(3-4): 269-291.

Gojenola, K., 2000, *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan.* Doktoretza-tesia, Informatika Fakultatea. Euskal Herriko Unibertsitatea, Donostia.

Golding, A. R. eta Roth, D. A., 1999, «Winnow-Based Approach to Context-Sensitive Spelling Correction», *Machine Learning* 34(1-3): 107-130.

Hashemi, S. S., Cooper, R. eta Andersson R., 2003, «Positive grammar checking: A finite state approach», *Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing 2003,* Mexico City, Mexico, February 16-22, 2588 lib. of Lecture Notes in Computer Science, 635-646. Springer. ISBN 0302-9743.

Holan, T.; Kuboň, V. eta Plátek M., 1997, «A prototype of a grammar checker for Czech», *Proceedings of the fifth conference on Applied Natural Language Processing* 147-154, San Francisco, CA, USA.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T. eta Isahara, H., 2003, «Automatic error detection in the Japanese learners' English spoken data», *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* 145-148, Morristown, NJ, USA. Association for Computational Linguistics. ISBN 0-111-456789.

James, C., 1998, «Errors in Language Learning and Use», *Applied Linguistics and Language Study*, Sydney.

Johannessen, J. B., Hagen, K. eta Lane, P., 2002, «The performance of a grammar checker with deviant language input», *Proceedings of the 19th international conference on Computational linguistics* 1-8, COLING, Taipei, Taiwan. Association for Computational Linguistics.

Jurafsky, D. eta Martin, J. H., 2000, «Speech and Language Processing». *An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, USA. ISBN 0-13-095069-6.

Karlsson, F., Voutilainen, A., Heikkila, J. eta Anttila, A., 1995, *Constraint Grammar: Language-independent System for Parsing Unrestricted Text.* Prentice-Hall, Berlin.

Karttunen, L., Gaál, T. eta Kempe, A., 1997, «Xerox finite state tool». Barne-txostena, Xerox Research Centre Europe.

Koskenniemi, K., 1983, *Two-level Morphology: a general computational model for word-form recognition and production.* University of Helsinki, Helsinki.

Kukich, K., 1992, «Techniques for automatically correcting words in text». *ACM Computing Surveys* 24(4), December.

Kumar, A. eta Nair, S., 2007, «An artificial immune system based approach for English grammar checking». *ICARIS 2007, Lecture Notes in Computer Science, number LNCS 4628* 348-357.

Lersundi, M., 2005, *Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpenak eta postposizioak.* Doktoretza-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea, Leioa.

Mangu, L. eta Brill, E., 1997, «Automatic rule acquisition for spelling correction», *Proceedings of the 14th International Conference on Machine Learning* 187-194. Morgan Kaufmann.

Mel'čuk, I., 1988, *Dependency Syntax: Theory and Practice.* State University of New York Press.

Mitjushin, L., 1996, «An agreement corrector for Russian», *Proceedings of the 16th conference on Computational linguistics* 776-781, Morristown, NJ, USA. Association for Computational Linguistics.

Moré, J., Climent, S. eta Oliver, A., 2004, «A grammar and style checker based on Internet searches», *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004* 1931-1934, Lisbon, Portugal.

Naber, D., 2003, *A Rule-Based Style and Grammar Checker.* Doktoretza-tesia, Universität Bielefeld.

Nivre, J., 2005, «Dependency grammar and dependency parsing». Barne-txostena, Växjö University: School of Mathematics and Systems Engineering.

Oflazer, K., 1996, «Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction», *Comput. Linguist.* 22(1): 73-89. ISSN 0891-2017.

Oronoz, M., 2009, *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura.* Doktoretza-tesia, Euskal Herriko Unibertsitatea.

Örvar Kárason, 2005, *Detecting grammatical errors with memory-based learning.*

Otegi, A., 2006, «Zuzentzaile sintaktikoa Word-en integratzeko liburutegi baten sorkuntza». Barne-txostena, UPV/EHU.

Paggio, P., 2000, «Spelling and grammar correction for Danish in SCARRIE», *Proceedings of the sixth conference on Applied Natural Language Processing* 255-261, San Francisco, CA, USA.

Shaalan, K. F., 2005, «Arabic GramCheck: a grammar checker for Arabic: Research articles», *Software: Practice and Experience* 35(7): 643-665. ISSN 0038-0644.

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

Sjöbergh, J., 2005, «Chunking: an unsupervised method to find errors in text». *Proceedings of NODALIDA 2005*, Joensuu, Finland.

— eta Knutsson, O., 2005, «Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing», *Proceedings of RANLP 2005* 506-512, Borovets, Bulgaria.

Sleator, D. eta Temperley, D., 1993, «Parsing English with a Link Grammar», *Proceedings of the Third International Workshop on Parsing Technologies,* Tilburg, The Netherlands and Durbuy, Belgium, Augus.

Tapanainen, P., 1996, *The Constraint Grammar parser CG-2.* Publications of the University of Helsinki, 27, Helsinki.

Teixeira Martins, R., Hasegawa, R., Volpe Nunes, M. D. G., Montilha, G. eta Osvaldo Ńovais de Oliveira, J., 1998, «Linguistic issues in the development of ReGra: A grammar checker for Brazilian Portuguese», *Natural Language Engineering* 4(4): 287-307. ISSN 1351-3249.

Tesnière, L., 1959, *Éléments de syntaxe structurale.* Paris.

Vandeventer, A., 2003, *Syntactic Error Diagnosis in the context of Computer Assisted Language Learning.* Doktoretzatesia, Universite de Geneve, Geneve.

Vosse, T., 1992, «Detecting and correcting morpho-syntactic errors in real texts», in *Proceedings of the 3rd Conference on Applied Natural Language Processing* 111-118.

Zabala, I. eta Odriozola, J. C., 2004, «Los complejos posposicionales en vasco». In E. Pérez Gaztelu, I. Zabala eta Ll. Grácia (arg.), *Las fronteras de la composición en lenguas románicas y en vasco,* 281-315.