Euskal Herriko Unibertsitatea / Universidad del País Vasco

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Lengoaia eta Sistema Informatikoak Saila

Entitate-Izenak Euskaraz: Identifikazioa, Sailkapena, Itzulpena eta Desanbiguazioa

Izaskun Fernandez Gonzalezek

Informatikan Doktore titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

Donostia, 2012ko otsaila.

Euskal Herriko Unibertsitatea / Universidad del País Vasco

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Lengoaia eta Sistema Informatikoak Saila

Entitate-Izenak Euskaraz: Identifikazioa, Sailkapena, Itzulpena eta Desanbiguazioa

Izaskun Fernandez Gonzalez Iñaki Alegriaren eta Nerea Ezeizaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2012ko otsaila.

Lan honetarako, ikertzaileak prestatzeko Eusko Jaurlaritzaren doktoretza aurreko beka bat izan nuen 2003 eta 2004 urteetan (BFI04.432).

Un melenudo está haciendo autostop en la carretera. Se detiene Lazkao Txiki y el joven le pregunta:
- Oiga, ¿me falta mucho para León?.
- El rabo, hijo, el rabo.

Lazkao Txiki

Con constancia y tenacidad se obtiene lo que se desea; la palabra imposible no tiene significado.

Napoleón Bonaparte

Etxekoei

Eskerrik asko!

Azken hamarkada honetan egunero buruan egon den tesi hau egin ahal izateko jende asko izan dut alboan, eta horiei guztiei eskerrak emateko tartea hartu nahi dut.

Lehenik eta behin Iñakiri eta Nereari nire esker beroenak eman nahi dizkiet. Ez zarete soilik nire alboan tesi-zuzendari gisa egon, momentu kritikoetan beti *gabinete de crisis*ak montatzeko prest egon zarie, eta behar izan dudan bakoitzean bultzadatxoa ematen eta laguntzen hortxe egon zarete. Hori guztiagatik eta ikerketa munduan murgiltzen erakusteagatik, mila esker. Zuen laguntza gabe ez nintzateke lerro hauek idazten egongo.

Aspaldi izan bazen ere, ezin dot ahaztu Klarak, Rubenek eta Irenek egindako eskuzko lanak (corpusak prestatu, erregelak idatzi), horiek gabe tesi honen hasiera izan zen Eihera eta geroagoko beste batzuk egitea ezinezkoa izango baitzan.

Xavier Carrerasi, ikerketa honen hasieran eskainitako laguntzagatik eta *Abionet* uzteko eta azaltzeko erakutsitako eskuzabaltasunagatik. Baita Sydney Unibertsitateko hizkuntzen teknologien ikerketa taldeari beraien hizkuntzen independentea den entitateak aztertzeko tresna erabiltzeko aukera emateagatik. Eta Jordi Atseriasi *Yahoo!*k semantikoki etiketatutako ingelesezko Wikipedia bertsioa luzatzeagatik.

Itsa, mila esker laguntzeko beti erakutsi didazun prestutasunagatik, nire txapak entzuteko zure dibana beti prest izateagatik, eta beti gauzak beste ikuspuntu batetik begiratzeko zure jokabidea nirekin konpartitzeagatik. Lan honetan eta hemendik kanpo oso lagungarri izan zara niretako.

Ezin ahaztu, *tropikoan* hainbeste ordu konpartitu eta bide luze hau xamurragoa izaten lagundu zenuten bulegokideei (Idoia, Itsa, Aran, Ainara). Eta bulegoan ez zeundeten Ixa kide eta kide ohi guztiei ere, mila esker.

Aitor, Eneko, Oier eskerrik asko desanbiguazioko kontuetan nirekin izan

duzuen pazientziagatik. Nola ez, Bertol eta Maite zuei ere mila esker edozein izan dela arazoa beti laguntzeko prest egoteagatik. Eta orokorrean latexeko kontuekin laguntzeko prest egon zareten guztiei (Aitor, Maite, Bertol).

Txosten honen zuzenketan lagundu didazuen guztiei ere eskerrik asko (Iñaki, Nere eta Olatz).

Teknikerreko lankideei ere eskerrak eman nahi dizkiet, bereziki Xabier Garcia de Kortazarri eta Aitor Arnaizi tesiko gaietan beti izan duten eskuzabaltasunagatik.

Eta orain etxekoentzako tartea ailegatu da. Aita, ama, zuei esker iritsi naiz ona. Eta nola ez Aneri esker ere, ahizpa txiki on bezala beti laguntzeko prest eta *apoyo morala* ematen egon zarelako.

Lagunei, beti hortxe egoteagatik.

Eta Mikeli eman nahi nizkioke azken eskerrak. Denbora guzti honetan izan duzun pazientziagatik, honek hainbatetan gure prioridadeak baldintzatu baditu ere beti aurpegi ona jartzeagatik, eta nire gora beheretan, ez direla gutxi izan, beti hortxe egoteagatik. Guztiagatik, bihotz bihotzez mila esker.

Asko zaretenez hemen agertu beharrekoak seguru baten bat ahaztu zaidala. Hala bada barkatu eta eskerrak zuri ere.

Eta orain bai, eskerrak hau bukau dan!

Laburtzapenak

Euskaraz:

bESA: ESA orekatua, *balanced ESA*
CoNLL: *Conference on Natural Language Learning*
EDBL: Euskararen Datu-Base Lexikala
EusWN: Euskal WordNet
ESA: Analisi Semantiko Esplizitua, *Explicit Semantic Analysis*
HAD: Hitzen Adiera-Desanbiguazioa
HMM: Markoven Eredu Ezkutua, *Hidden Markov Model*
HP: Hizkuntzaren Prozesamendua
IA: Ikasketa Automatikoa
IE: Informazioaren Erauzketa, *Information Extraction*
IR: Informazioaren Eskurapena, *Information Retrieval*
LOC: Tokia, *Location*
MCR: *Multilingual Central Repository*
ME: Entropia maximoa, *Maximum Entropy*
MISC: *Miscellaneous*
MFS: Maiztasun Handieneko Adiera, *Most Frequent Sense*
MT: Itzulpen Automatikoa, *Machine Translation*
MUC: *Message Undertanding Conference*
NB: *Naïve Bayes*
NE: Izendun Entitateak, *Named Entity.*
Tesi-lan honetan laburdura hau Izendun Entitateen espresio konplexuenak diren Entitate-Izenak aipatzeko erabiltzen da ere.
NEC: Izendun Entitateen Sailkapena, *Named Entity Classification.*
Tesi-lan honetan Entitate-Izenen Sailkapena adierazteko ere erabiltzen da.
NED: Entitate-Izenen Desanbiguazioa, *Named Entity Disambiguation*
NER: Izendun Entitateen Identifikazioa, *Named Entity Recognition.*
Tesi-lan honetan Entitate-Izenen Identifikazioa adierazteko ere erabiltzen da.
NERC: Izendun Entitateen Identifikazioa eta Sailkapena, *Named Entity Recognition and Classification.*
Tesi-lan honetan Entitate-Izenen Identifikazioa eta Sailkapena adierazteko ere erabiltzen da.
NET: Entitate-Izenen Itzulpena, *Named Entity Translation*
ORG: Erakundea, *Organization*

OTH: Bestelakoak, *Others*
PER: Pertsona, *Person*
PoS: Kategoria gramatikala, *Part of Speech*
QA: Galdera-Erantzun, *Question Answering*
SVM: Sostengu-Bektore Bidezko Makinak, *Support Vector Machine*
TA: Transliterazio-Automata
TC: Testu Sailkapena, *Text Classification*
tf-idf: *Term frequency - inverse document frequency*
VSM: Bektore-Espazioen Eredua, *Vector Space Model*
WePS: *Web People Search*
WIL: Hizkuntzen arteko Wikipediako Estekak, *Wikipedia Interlingual Links*
WIR: Webean oinarritutako IR, *Web based Information Retrieval*
WN: WordNet
XFST: *Xerox Finite State Transducer*

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

I. KAPITULUA

Proiektuaren nondik norakoak

I.1 Gaiaren motibazioa eta kokapena

I.1.1 Motibazioa

Gizakiok egunetik egunera handitzen doan informazio kopurua modu erraz batean eskuratu, irakurri eta ulertzeko gaitasuna badugu ere, gaur egun makinak prozesu hori guztia automatikoki burutzeko oso urruti daude, testuak bezalako informazio ez-egituratuari buruz ari garenean behinik behin. Esaterako, web-orri zein blogetan publikatzen diren edukiak, sare-sozialetan agertzen diren elkarrizketak edota notiziak, eta oro har, Interneten edo eduki digital gisa sortzen den edozein informazio interpretatzeko eta ulertzeko gaitasuna dugu gizakiok. Makinek ordea, interpretazio eta ulermen hori gaur egun ezinezkoa dute, nahiz eta hizkuntzaren prozesamenduko (aurrerantzean HP) ikerketek helburu hori geroz eta gertuago izatea lortzen ari diren.

Izan ere, testuen interpretazio eta ulermenaren automatizazio bide horretan, teknologiak eboluzionatu ahala, badira ulermenera iristeko prozesu oso horretan parte hartzen duten eta automatizatzea lortu diren hainbat oinarrizko problema. Esaterako, testuetatik informazioa eskuratu eta erauzten laguntzen duten analisi morfologiko eta analisi sintaktikoa bezalako oinarrizko prozesuak hizkuntza askotarako makinen bidez automatizatzea lortu diren oinarrizko problemak dira. Eta oinarrizko prozesu horien automatizazioari esker, ahalmen handiagoa duten galdera-erantzun sistemak edota itzulpen automatikoko aplikazioen moduko sistema automatikoak garatzeko aukera

sortu da.

HPko oinarrizko prozesu edo eginkizun konkretu horien artean kokatzen da, hain zuzen ere, tesi-lan honen funtsa den entitate-izenen azterketa. Entitate-izenak, pertsona, toki edota erakundeak aipatzen dituzten espresioak dira. Eta tesi-lan honen helburu nagusia euskarazko testuetan agertzen diren espresio horien tratamendua da.

Entitate-izenekin lanean hastean agertzen den lehenbiziko eginkizuna, testuetan zehar agertzen diren espresio horien identifikazioa eta ondoren horien sailkapena da. Alegia, "*Euskal Herriko Unibertsitatean burutzen ari da Izaskun bere doktorego ikasketak*" esaldian, lehenbizi bertan agertzen diren *Euskal Herriko Unibertsitatea* eta *Izaskun* entitate-izenak identifikatu, eta ondoren erakunde- eta pertsona-izen gisa sailkatzea dira bete beharreko eginkizunak, hurrenez hurren.

Entitate-izenen identifikazioak bere zailtasunak baditu ere, sailkapena da bietatik konplexuena. Izan ere, bietatik sailkapena baita entitate-izen batean egon daitekeen anbiguotasunari aurre egin behar dion ataza. Entitate-izenen sailkapen atazan anbiguotasuna bi modutara ager daiteke: espresio berak bi sailkapen posible izan ditzake, alegia, espresio berak mota desberdinetako bi entitate-izenei erreferentzia egin diezaioke; eta sailkapen berari egokitu arren, espresioak bi entitate-izenei aipamena egin diezaieke. Bi anbiguotasun kasu horien adibide dira I.1.1 adibidean aipatzen diren espresioak, hurrenez hurren.

I.1.1 Adibidea Entitate-izenen sailkapen atazaren anbiguotasun arazoen adibideak.

Walt Disney: marrazki bizidunen erakundeari zein bere sortzaile den pertsona kategoriako entitate-izenari egin diezaioke erreferentzia.

Armstrong: *Lance Armstrong* txirrindulari zein *Neil Armstrong* astronauta, bi pertsona-izenak adieraz daitezke abizen horrekin.

Lehen motako anbiguotasunak entitate-izenen sailkapen probleman eragin zuzena badu ere, bigarrenak ez du horrelako arazorik sortzen. Bai, ordea, informazio hori ahalmen handiagoko aplikazio batean erabili nahi denean, galdera-erantzun sistema batean esaterako. Horrelakoetarako beharrezkoa da ebaztea entitate-izenaren agerpenak (*Armstrong*) bietako zein pertsonari (*Lance Armstrong* edo *Neil Armstrongi*) egiten dion erreferentzia, alegia, entitate-izenaren agerpena desanbiguatzea beharrezkoa da. Entitate-izenen desanbiguazioa, beraz, arreta berezia behar duen ataza dela esan daiteke.

Testuinguru eleanitz batean kokatzean, ordea, non entitate-izenak hizkuntza desberdinetan ager daitezkeen, hizkuntza desberdinetako espresioak

elkarren artean lotzeko identifikazio-, sailkapen- eta desanbiguazio-atazak ez dira nahikoak. Entitate-izen bat bi hizkuntza desberdinetan nola adierazten den jakitea beharrezkoa da lotura hori gauzatu ahal izateko. Informazio hori eskuratzeko bideetako bat, hain zuzen ere, entitate-izenen itzulpen-ataza da: abiapuntu-hizkuntzako entitate-izenentzat xede-hizkuntzako adierazpena eskuratzea helburu duen ataza. Gainera lotura eleanitza eginda, batzuetan, entitate-izen bat hobeto sailkatzea edo desanbiguatzea posible izango da.

Entitate-izenak inolako tratamendu berezirik aplikatu gabe testuak itzultzeko itzulpen automatikoko estrategia orokorrekin itzultzean, *Escuela de Derecho de Harvard* bezalako espresioak ingelesez *School of the Right of Harvard* itzuli ohi dira. Oro har, itzulpen-estrategia okerra dela ezin da esan, bai ordea, entitate-izenak itzultzeko estrategia egokia ez dela. Hori dela eta, espresio horiek modu berezian tratatzen dituen entitate-izenen itzulpen-estrategia definitzea beharrezkoa da, itzulpen-automatikoko sistemak horiek ondo ebatzi ditzan.

Entitate-izenen inguruko identifikazio-, sailkapen-, itzulpen- eta desanbiguazio-problema nagusi horiek ebazteko estrategiak erabiliz, atal honen hasieran aipatutako ahalmen handiagoko aplikazioen portaera hobetu daiteke. Horren adibide galdera-erantzun eleanitzeko aplikazioa izan daiteke; non, hiru ataza horien ebazpen automatikoak aplikazioari jarraian aipatzen diren hobekuntzak eskaintzen dizkion:

- Galderan zein erantzun-sortan, entitate-izenak identifikatuta eta sailkatuta izateak, galdetzen den entitate-izenaren sailkapenaren arabera, erantzuna bilatu beharreko dokumentu sorta murrizten lagun dezake, bilaketa-estrategia optimizatzen lagunduz.
- Erantzun sortan, hizkuntza desberdinetako dokumentuak egonik, galdegaia den entitate-izenaren beste hizkuntzetako espresioa ezagutzeak, erantzun eleanitz eta osoagoa eman ahal izatea ahalbidetzen du.
- Azkenik, galderan zein erantzun sortan espresio batek entitate-izen bat baino gehiago adieraz ditzakeenean, agerpen horiek erreferentziatzen duten entitate-izen egokia ezagutzea galderaren erantzun posibleen artean adiera egokia ez duten testuak emaitza multzotik kentzeko oso lagungarria gerta daiteke.

Tesi-lan honetan euskarazko entitate-izenen inguruko hiru problema horiek automatizatzen egin dugu lan, hain zuzen ere.

I.1.2 Gaiaren kokapena

Atal honetan, lehenik eta behin Hizkuntzaren Prozesamenduaren alorrean lantzen diren bi hurbilpen nagusiak aztertuko ditugu (I.1.2.1 atala). Ondoren, orain arte *IXA taldea*n[1] egindako beste lanekin duen lotura, eta tesilanean erabilitako baliabide nagusien deskribapenarekin batera gure lana zein testuingurutan kokatzen den deskribatuko dugu (I.1.2.2 atala).

I.1.2.1 Hizkuntzaren Prozesamenduko bi hurbilpenak

Hizkuntzaren Prozesamenduan bi hurbilpen nagusi (Oronoz, 2009) lantzen dira oro har:

1. Hizkuntza-ezagutzan oinarritutako teknikak edo teknika sinbolikoak: adituek formalismoren baten bidez definitutako erregela gramatikalak darabiltzate hauek, oro har.

2. Corpusetan oinarritutako teknikak edo teknika enpirikoak: ikasketa automatiko bidezko metodoak edo teknika estatistikoak baliatzen dituzte hauek, eskuarki.

Hizkuntza-ezagutzan oinarritutako teknikek erregeletan edo bestelako adierazpide formaletan kodetzen dituzte modu esplizituan hizkuntzari buruzko ezagutza (Dale, 2000). Teknika sinboliko horiek oso erabiliak izan dira azken 40 urteetan, baina zenbait arazo sortzen dituzte. Hasteko, hizkuntza-ezagutzarekiko menpekoak dira; lantzen den hizkuntzaren arabera aldatu egin beharko dira erregelak, alegia. Gainera, mundu errealeko arazoei aplikatzerakoan *hedagarritasunaren arazoa* agertzen da. Izan ere, oso zaila da hizkuntzaren fenomeno bakoitza, bere osotasunean, erregela formal batzuen bidez adieraztea. Horrez gain, hizkuntzaren ezagutza handia behar da hizkuntzaren fenomenoak deskribatuko dituzten erregela formalak idazteko, hizkuntza-ezagutzan oinarritutako tekniketan, tratatzen den fenomenoan adituek eskuz garatu behar izaten dituzte.

Corpusetan oinarritutako teknikek hizkuntza-ezagutzan oinarritutako tekniken aldean, hazkunde izugarria izan dute azken 20 urteetan (Oronoz, 2009), corpusen eta beharrezko baliabide informatikoen sorrerak bultzatuta eta hizkuntza-ezagutzan oinarritutako tekniketan aurretik aipatutako arazoak izan direla medio. Corpusetan oinarritutako teknikak estatistika eta

[1] `http://ixa.si.ehu.es`

ikasketa automatikoko estrategiak erabili ohi dituzte. Azken horiek hiru multzo nagusitan banatzen dira: gainbegiratuak, ez-gainbegiratuak eta erdi-gainbegiratuak. Ikasketa gainbegiratuan oinarritutako metodoek ikasketa-prozesua burutu aurretik, aurrez etiketatuko dokumentu bilduma eskuragarri behar dute, eta berau erabiltzen dute sistema trebatzeko. Ondoren, adibide berriak datozenean modu automatikoan tratatzeko erabiltzen dute ikasitako eredua. Ikasketa ez-gainbegiratuan oinarritutako metodoek berriz, ez dute aurrez etiketatutako dokumenturik behar. Ikasketa itsu-itsuan burutzen dute. Azkenik, ikasketa erdi-gainbegiratuan oinarritutako metodoek ikasketa-prozesua burutu aurretik, aurrez etiketatuko dokumentu multzo txiki bat erabiltzen dute. Dokumentu multzo txiki hori etiketatu gabeko dokumentu multzo handi batekin konbinatuz eredua trebatzen da.

I.1.2.2 Testuingurua

*IXA taldea*k hogei urte baino gehiago daramatza Hizkuntzaren Prozesamenduan lanean. Arlo zabal horren barruan, euskararen ikerketa aplikatua izan da taldearen xede nagusia, eta helburu horrekin, orain arte, morfologia, sintaxia eta semantika landu dira, batez ere.

Hizkuntzalarien eta informatikarien elkarlanari esker, euskarako baliabide eta tresna ugari sortu dira. Hala nola, Euskararen Datu-Base Lexikala (EDBL) (Aldezabal *et al.*, 2001), MORFEUS analizatzaile morfologikoa (Aduriz *et al.*, 1998) eta MORFEUSen oinarritutako Xuxen euskarako zuzentzaile ortografikoa (Agirre *et al.*, 1992; Alegria *et al.*, 2008), zenbait analizatzaile sintaktiko (Aduriz *et al.*, 2006a), corpusak (Aduriz *et al.*, 2006b), hiztegi elektronikoak (Arregi *et al.*, 2003; Díaz de Ilarraza *et al.*, 2007), sare semantikoa (Agirre *et al.*, 2006; Pociello, 2008) eta beste hainbat.

Horien guztien artean, *Eustagger* (Ezeiza *et al.*, 1998) Euskararako etiketatzaile/desanbiguatzaile morfosintaktikoa izan da tesi-lan honetan guztian zehar oinarrizko informazio iturri. Izan ere, txostenean zehar ikusiko dugun bezala, euskarazko entitate-izenen inguruko problemak ebazteko abiapuntuko datuak tresna horrekin aurreprozesatzea ezinbestekoa izan da kasu askotan. Euskara hizkuntza eranskaria izanik, testuetako osagaien kategoria/azpikategoria eta bereziki deklinabide-kasuen inguruko informazioa oso erabilgarria gertatzen baita edozein prozesamendu automatiko burutzeko garaian.

Tresna horrekin batera, *IXA taldea*n bildutako corpusak ere oso erabilgarriak gertatu dira. Izan ere, entitate-izenen inguruko hainbat problematan

corpusetan oinarritutako teknikak erabili nahi izan ditugu. Bereziki, ikasketa automatikoko metodo erdi- eta ez-gainbegiratuak erabili nahi izan ditugu. Euskara baliabide urriko hizkuntza izanik, ikasketa automatikoko metodo gainbegiratuak aplikatu ahal izateko euskarazko baliabide kopurua, mugatua gehienetan, txikiegia izan ohi baita.

Hizkuntza-ezagutzan oinarritutako teknikekin eta ikasketa automatikoko metodo gainbegiratuekin entitate-izenen identifikazio eta sailkapena ebazteko saiakerak egin baditugu ere, ikasketa automatikoko metodo ez-gainbegiratu eta erdi-gainbegiratuak izan dira entitate-izenen itzulpena eta desanbiguazioa ebazteko erabilitako metodoak. Horien konbinazioak eskaintzen dituen hobekuntzak ere aztergai izan ditugu lan honetan.

Tesi-lan hau, beraz, *IXA taldea*n egindako lan horien testuinguruan ulertu beharra dago. Hala ere, badira *IXA taldea*n garatu ez eta erabili diren bestelako baliabideak ere. Erabileran oinarrituz, aipagarrienak Wikipedia eta Hermes corpus eleanitzak dira.

Hermes corpusa Hermes proiektuan[2] bilaketa eleanitz eta erauzketa semantikorako hemeroteka elektronikoak sortzea helburu duen proiektuaren testuinguruan osatutako corpusa da. Corpusean urte bereko euskara, katalan, gaztelera eta ingelesezko egunkarietako berriak biltzen dira. Berrietako testuak gordinean egon ezik, automatikoki etiketatuta ere gordetzen dira. Etiketatze automatikoan analisi morfosintaktikoa, testuen sailkapena, izendun entitateen identifikazioa eta sailkapena, eta dokumentuen arteko loturen erauzketa burutzen dira, besteak beste. Euskarazko testuen etiketatzea *IXA taldea*n burutu da.

Azken urteetan HPko ataza askotan erabiltzen ari den Wikipedia, bestetik, tesi-lan honetan askotan erabili den beste baliabidea izan da. Wikipedia entziklopedia eleanitza, web oinarriduna eta eduki askekoa da, eta bertako sarrerak mundu osoko boluntarioek idazten dituzte. Sarrera bakoitzaren identifikadore unibokoa titulua da. Titulua sarrerak deskribatzen duen kontzeptuaren forma usuena izan ohi da, eta usuenak ez diren kontzeptuen aldaerak edo formak berbideratze- eta desanbiguazio-orrien bitartez lotzen dira sarrera nagusira. Wikipedia baliabide eleanitza izanik, sarrera bera deskribatzen duten hizkuntza desberdinetako sarrerak topa daitezke, eta horiek esteken bitartez lotuta egon ohi dira. Hizkuntza bereko sarreren arteko aipamenak egitean ere, esteken bitartez esplizitu egiten dira aipamen horiek.

Wikipediak eskaintzen duen eleaniztasuna, eta kontzeptuak eta bere al-

[2] http://nlp.uned.es/hermes/

daerak erlazionatzeko dituen irizpideak, oso interesgarriak dira entitate-izenen tratamenduan informazio-iturri gisa erabiltzeko. Bereziki, entitate-izenen itzulpen eta desanbiguazio atazak ebazteko.

Laburbilduz, euskarazko entitate-izenen tratamendu automatikoa burutzean datza tesi-lan hau, baita horretarako beharrezkoak diren tresnen inplementazioan ere: euskarazko entitate-izenen identifikazioa eta sailkapena, espresio horien itzulpena, eta azkenik desanbiguazioa. *IXA talde*ko hainbat ikerketa lanetan itzulpen-automatikoan (Labaka, 2010), semantika alorreko hitz-adieren desanbiguazioa (Lopez de Lacalle, 2009) eta terminoen erauzketaren (Gurrutxaga *et al.*, 2005) inguruko lan sakonak aurki badaitezke ere, bakoitzean entitate-izenen inguruko azterketa berezirik ez da egin orain arte. Eta behar hori betetzeko sortzen da tesi-lan hau, hain zuzen ere.

I.2 Helburuak

Tesi-lan honen helburu nagusia, beraz, euskarazko entitate-izenak automatikoki lantzea da. Helburu nagusi hori lortzeko metodologikoki hiru irizpide hartu dira kontuan:

- baliabide urriko hizkuntza izanik, baliabideen berrerabilpena eta metodo ez-gainbegiratu edota erdi-gainbegiratuei lehentasuna eman.
- hizkuntza-ezagutzan eta ikasketa automatikoko teknikak entitate-izenen atazak ebazteko ustiatu eta ahal denean konbinatu. Hala, metodo sofistikatuak baino metodo sinpleen konbinazioak eta egokitzapenak ustiatuko dira.
- entitate-izenen atazak automatizatzean euskararen ezaugarri morfosintaktikoek duten eragina aztertu.

Irizpide metodologiko horiek jarraituz tesiaren helburu orokorraren baitan hiru eginkizun nagusi landuko dira:

1. *Euskarazko entitate-izenen identifikazio eta sailkapena.*
   Euskarazko testuetan agertzen diren entitate-izenak automatikoki identifikatu eta sailkatzeko tresnaren garapenerako bidean, hizkuntza-ezagutzan zein ikasketa automatikoko teknikak erabiliz eta konbinatuz, ingelesarentzat garatutako tresnen pareko tresna garatzea da ataza honen helburu nagusia. Zeregin horretan, teknika egokienak aztertu eta

horien konbinazioen portaerak aztertzeko tarteko helburua ere landuko da. Euskararen kasuan baliabide mugatuen eta morfologia aberatsaren problemei aurre egiteko bidea bilatuko da.

2. *Euskarazko entitate-izenen itzulpena.*
   Itzulpen-automatikoko zein galdera-erantzun eleanitzeko aplikazioetarako lagungarri gertatzen diren entitate-izenen aipamen eleanitzak automatikoki sortzeko estrategia da eginkizun honen funtsa. Euskarazko entitate-izenak izanik abiapuntua eta, hasiera batean, gaztelera itzuli nahi den xede-hizkuntza, hizkuntza-ezagutzan oinarritutako eta teknika erdi-gainbegiratuekin hurbilpen desberdinak egin eta horien portaerak aztertu nahi dira. Hurbilpen bakoitzerako beharrezkoak diren baliabideak eta emaitzak aztergai izango dira. Azkenik, teknika erdi-gainbegiratuekin egindako ekarpena beste hizkuntza bikote batzuetara hedatzeko ahalmena ere aztertu nahi da.

3. *Euskarazko entitate-izenen desanbiguazioa.*
   Euskarazko testuetan agertzen diren entitate-izenen agerpen anbiguoak automatikoki desanbiguatzea da eginkizun zehatz honetan ebatzi beharrekoa. Edozein desanbiguazio-atazatan bezala, desanbiguazioa automatikoki burutu ahal izateko agerpenaren testuinguruaz gain ezagutza-base bat sortzea beharrezkoa izango da, non espresio anbiguo baten adiera posibleak deskribatuko diren. Euskarazko entitate-izenen desanbiguaziorako ezagutza-base horren definizioan, euskarazko Wikipediaren erabilgarritasuna aztertu nahi da. Eta baliabide horren ezaugarriak ahalik eta hobekien baliatuz, euskarazko entitate-izenak automatikoki desanbiguatzeko agerpen bat eta Wikipedia sarrera bat lotzen duen prozesua definitu nahi da. Prozesu horren automatizaziorako, ingelesa bezalako beste hizkuntza batzuetarako erabilitako teknika onenak erabili, eta euskara bezalako baliabide urriko hizkuntzan baliabide mugatuekin lan egitean horien portaera aztertu nahi izan da. Wikipedia handien aurrean, txikien gainean gertatzen diren desberdintasunak ere aztertu nahi dira.

Tesi-lan honetan, beraz, euskarazko entitate-izenen identifikazioa eta sailkapena, itzulpena eta desanbiguaziorako tresna automatikoak garatzea helburu nagusiak badira ere, horien garapenerako baliabide mugatuak daudenean estrategia desberdinek duten portaera aztertu eta konparatu nahi izan da. Konparaketa horiek posible izateko, eginkizun bakoitzean, estrategia

desberdinak erabili dira; euskararako baliagarritasuna aztertzeaz gain, baliabide eta metodoen azterketa hori baliabide urriko beste hizkuntzetarako ere baliagarri izango delakoan.

I.3 Tesi-txostenaren eskema

Lehen kapitulu honen ostean, tesi-lan honetako helburu diren euskarazko entitate-izenen inguruko eginkizun bakoitzaren ildoan burututako lanen deskripzioa azalduko dugu. Eginkizun bakoitzarentzako kapitulu bat erabili dugu, hiru kapitulu nagusiko txostena osatuz.

II. kapituluan euskarazko testuetan agertzen diren entitate-izenen identifikazio eta sailkapen automatikorako burututako esperimentuak deskribatu ditugu. Jarraian, III. kapituluan, euskarazko entitate-izenetatik abiatuz, bestelako hizkuntzetara espresio horien itzulpena gauzatzeko egindako hurbilpenak zehaztu ditugu. Eta IV. kapituluan, berriz, euskarazko testuetako entitate-izenen agerpenak desanbiguatzeko espresio horien agerpenak euskarazko Wikipediarekin lotzeko egindako esperimentuak azaldu ditugu.

Bukatzeko, tesi-lan hau garatzen atera ditugun ondorioak eta egindako hausnarketak aipatu ditugu. V. kapitulu horretan ere, lan honetan abiatutako ildoei jarraiki, etorkizunean zein bide har daitezkeen aztertu dugu.

Kapitulu hauez gain, lau eranskinek osatzen dute tesi-txosten hau: A eranskinak, euskarazko entitate-izenak identifikatzeko gramatika definitzen du; EDBL kategoria-sistema zerrendatzen duen B eranskina; euskarazko entitate-izenak gaztelerara itzultzeko definitutako transformazio fonologikoen erregelak biltzen dituen C eranskina; eta D eranskinak, euskarazko entitate-izenak gaztelerara itzultzeko espresio horien osagaien ordenaziorako erregelak deskribatzen ditu.

I.4 Argitalpenak

Sarrerako kapitulu honi bukaera emateko tesi-lan honekin lotuta burututako argitalpenen zerrenda aurkezten da jarraian. Batetik, tesiarekin hertsiki lotutako argitalpenak zerrendatu ditugu; eta bestetik, tesiarekin lotura estua izan ez arren, arlo honetan egindako ikerketei dagozkienak.

I.4.1 Tesiari hertsiki lotutako argitalpenak

II. kapituluko entitate-izenen identifikazio eta sailkapen atazaren inguruko argitalpenak:

- I. Alegria, N. Ezeiza, I. Fernandez, R. Urizar. Named Entity Recognition and Classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI2003*, 2003.
- I. Alegria, O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, R. Urizar. Design and development of a named entity recognizer for an agglutinative language. *First International Joint Conference on NLP (IJC NLP-04). Workshop on Named Entity Recognition*, 2004.
- I. Alegria, O. Arregi, N. Ezeiza, I. Fernandez. Lessons from the Development of a Named Entity Recognizer. *Procesamiento del Lenguaje Natural, vol. 36, p. 25-37*, 2006.

III. kapituluko entitate-izenen itzulpen atazaren ingurukoak:

- I. Alegria, N. Ezeiza, I. Fernandez. Named Entities Translation Based on Comparable Corpora. *Multi-Word Expressions in a Multilingual Context Workshop on EACL06, p.1-8*, 2006.
- I. Alegria, N. Ezeiza, I. Fernandez. Translating Named Entities using Comparable Corpora. *Building and Using Comparable Corpora. LREC 2008 Workshop*, 2008.
- I. Fernandez, I. Alegria, N. Ezeiza. Using Wikipedia for Named Entities Translation. *SALTMIL2009 Workshop: IR-IE-LRL*, 2009.

Eta azkenik, IV. kapituluko entitate-izenen desanbiguazioren inguruko argitalpena:

- I. Fernandez, I. Alegria, N. Ezeiza. Semantic Relatedness for Named Entity Disambiguation using a small Wikipedia. *TSD 2011, LNAI 6836, p. 276-283*, 2011.

I.4.2 Bestelako argitalpenak

- O. Ansa, X. Arregi, B. Arrieta, A. Díaz de Ilarraza, N. Ezeiza, I. Fernandez, A. Garmendia, K. Gojenola, B. Laskurain, E. Martínez, M. Oronoz, A. Otegi, K. Sarasola, L. Uria. Integrating NLP Tools for Basque in Text Editors. *Workshop on International Proofing Tools and Language Technologies*, 2004.
- A. Jimenez, I. Fernandez, D. Pérez, E. Viejo, F.J. Díez, de X. G. Kortazar, M. García, V. Maojo, A. Cobo, F. del Pozo. Patient-based Literature Retrieval and Integration - A Use Case for Diabetes and Arterial Hypertension. *HEALTHINF 2011*, 2011.
- I. Fernandez, A. Jimenez, X. G. Kortazar, D. Perez. A New Method to Retrieve, Cluster And Annotate Clinical Literature Related To Electronic Health Records. *Third International Workshop on Health Document Text Mining and Information Analysis (LOUHI)*, 2011.

II. KAPITULUA

Entitate-izenen Erauzketa

Izendun entitateen erauzketa (NERC), *Message Understanding Conference* (MUC) (Chinchor, 1998) izeneko konferentzian definitu zen bezala, entitate-izen (pertsona-, erakunde- eta toki-izen), adierazpen tenporal (data eta ordu) eta zenbakizko hainbat adierazpenen (portzentajeen, balio monetarioen eta antzekoen) identifikazioan eta sailkapenean datza. Sarri ataza hau bi azpi-atazatan banatu ohi da: izendun entitateen osagaiak eta mugak zehazten dituen Izendun Entitateen Identifikazioa (NER) alde batetik, eta sailkapena ebazten duen Izendun Entitateen Sailkapen ataza (NEC) bestetik.

Izendun entitateen erauzketa Hizkuntzaren Prozesamendu (HP) alorreko aplikazio ugarietarako oso lagungarri gertatu ohi da, besteak beste:

- Informazioaren Eskurapena (IR): galdera bat emanik dokumentu-bilduma bat itzultzean datza. Helburu horretarako dokumentu-bildumak aurreprozesatu eta indexatu ohi dira, galdera bat jaso ondoren bilaketa arin eta azkar bat burutu ahal izateko. Indexazioan zein galdera- prozesaketan izendun entitateen erauzketa lagungarri izan daiteke, espresio mota hauen inguruko galderak oso maiz agertzen baitira.

- Informazioaren Erauzketa (IE): testuetatik informazio egituratua edo erdi-egituratua ateratzea du helburu, besteak beste testuan agertzen diren izendun entitateen arteko erlazioak erauztea.

- Galdera-Erantzun Sistemak (QA): dokumentu bilduma batean, hizkuntza arruntaz egindako galdera zehatz baterako erantzun zuzena itzul-

tzeko gai den sistemari deritzo. Izendun entitateen erauzketak sistema hauetan egin ohi den galderaren sailkapenerako zein dokumentuen prozesaketan oso lagungarri gerta daiteke.

- Testu-sailkapena (TC): dokumentu edo pasarte bat, bere testuan oinarrituz, kategoria batean edo gehiagotan sailkatzean datza. Sailkapen horretan izendun entitateek laguntza handia eskaini dezakete, gehienetan espresio hauek gai bati lotuta egon ohi baitira.

- Itzulpen Automatikoa (MT): abiapuntu-hizkuntzako testu batetik automatikoki xede-hizkuntzako testu baliokide bat lortzean datza. Bi sistema mota nagusitzen dira arlo honetan: ezaugarri linguistikoetan oinarritutakoak eta corpusetan oinarritutakoak. Bi kasuetan hitz anitzeko espresioak identifikatzea, izendun entitateak barne, itzulpenaren prozesuan ezinbestekoa gertatzen dira, osagaiz osagaiko itzulpen arrunta aplikatzea ekiditeko.

II.1 Izendun entitateen erauzketarako teknikak

Izendun entitateen erauzketa automatikoari buruzko lehenbiziko lana, 1991 urtean kokatzen da. Rau (1991) erakunde-izenak detektatzeko erregela multzo bat eta heuristikoetan oinarritutako sistema aurkeztu zuenean, hain zuzen ere. Baina ataza honek benetako indarra 1996 urtean hartu zuen, *Message Undertanding Conference* konferentziako 6. edizioan (MUC-6[1]) (Grishman eta Sundheim, 1996). Ordutik asko izan dira ataza honen inguruan burututako ataza partekatuak eta konferentziak, besteak beste: MUC-7 (Chinchor, 1998)[2] eta *Conference on Natural Language Learning* (CoNLL) ((Tjong Kim Sang, 2002) eta (Tjong Kim Sang eta De Meulder, 2003)).

NERC atazan ikerketa gehien izan dituen hizkuntza ingelesa izan arren, Nadeau eta Sekinek (2007) aipatzen duten bezala, badaude beste hainbat hizkuntza aurretik aipatutako lehiaketetan definitutako baliabideei esker ikerketa zabala izan dutenak. Ingelesaz gain, NERC ataza landuen duten hizkuntzak alemana, gaztelera eta nederlandera dira, CoNLL-2002 eta CoNLL-2003 ataza partekatuetan definitutako baliabide eta lehien ondorioz.

[1] `http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html`

[2] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`

MUC-6 eta MUC-7 ataza partekatuetan, izendun entitate bezala entitate-izenak, espresio tenporalak eta zenbakizko espresioak hartzen dira. Izendun entitateen sailkapenean, hiru espresio mota horiek zazpi kategoria aurredefinituetan banatzen dira, eta horien artean egokiena aukeratzea da atazaren helburua. Zazpi kategoria horiek ondokoak dira:

- *ENAMEX* edo entitate-izenak: pertsona- (PER), toki- (LOC) eta erakunde- (ORG) izenak.
- *TIMEX* edo espresio tenporalak: data (DATE) eta denborazko espresioak (TIME).
- *NUMEX* edo zenbakizko espresioak: diru (MONEY) eta portzentajeak (PERCENTAGE).

Hiru multzo nagusien artean, *ENAMEX* edo entitate-izenak dira identifikatzen eta sailkatzen zailenak. Zailtasun hori dela eta, entitate-izenak dira izendun entitateen artean mota landuena.

CoNLL konferentzian *ENAMEX* motako entitateez gain *Miscellaneous* mota erabiltzen da, mota horrek, *ENAMEX* motatik at geratzen diren izen bereziak taldekatzea du helburu.

Ataza definitu zenetik NERC sistema gehienek erabilitako irizpideak aipatutako horiek izan arren, hainbat aplikaziotarako sailkapen xeheagoen beharrak entitate-izenen sailkapen finagoak burutzen dituzten sistemak garatzera bultzatu du. Ale larriko kategoria sisteman ez bezala, sailkapen xeheko kategoria sistemetan inolako estandarizaziorik ez da egin. Toki-izenen kasuan esaterako, GeoNames[3] bezalako baliabide handia existitu arren, eta bere kategoria-sisteman oinarritzen diren lanak aurki badaitezke ere, ez da sailkapen estandar gisa kontsideratu. Estandarizazio falta hori dela eta, Fleischman (2001) eta Lee eta Leek (2005) beraien lanetan toki-izenak kategoria xeheagoetan (hiri, herrialde, estatu, etab.) sailkatzeko sailkapen-hierarkia propioa erabiltzen dute, mota hauetako sistemen arteko konparaketak ia ezinekoak bihurtuz. Fleischman eta Hovyk (2002) berriz, pertsona-izenak politikari, animatzaile eta antzeko sailkapen xeheagoetan sailkatzeko saiakera egin zuten.

Aipatutakoez gain, badaude oso sailkapen xehea egiten dituzten sistemak ere. Esaterako, Sekine eta Nobatak (2004) egunkarietan agertu ohi diren

[3] `http://es.wikipedia.org/wiki/GeoNames`

izendun entitate mota ohikoenak identifikatu eta bereizteko ahaleginean, 200 izendun entitate mota identifikatzeko sistema diseinatu zuten.

Sailkapena ale xehekoa zein larrikoa izan, ataza aurrera eramateko informazio-iturriek garrantzia handia dute tresna sendo bat garatzerako garaian. McDonalden (1996) lanaren arabera, bi informazio-iturri erabili behar dira izendun entitateak identifikatu eta sailkatzeko garaian: barne-ebidentzia eta kanpo-ebidentzia. Egilearen arabera, barne-ebidentzia espresioak berak eta osatzen duten osagaiek eskaintzen duten informazioa da. Kanpo-ebidentzia, aldiz, espresioaren testuinguruak adierazten duena, alegia, inguruko hitzek ematen duten informazioa da.

Baliabideei dagokienez, izendun entitateak identifikatu eta sailkatzeko probleman erabili diren kanpo-iturrietatik eskuratutako baliabide nagusiak hitz eragileen (*trigger-word*) zerrendak eta izendun entitateen zerrendak (*gazetteer*ak aurrerantzean) dira. Hasierako lanetan zerrenda horiek eskuz sortutakoak baziren ere, lan berriagoetan *WordNet*en (Fellbaum, 1998) zein Wikipedian oinarritutako zerrendak erabili izan dira, Magnini *et al.* (2002) egileen eta Toral eta Muñozen (2006) lanetan deskribatzen den bezala, hurrenez hurren. Bi baliabide horiek bere osotasunean, eta ez soilik zerrendak sortzeko, erabili dituzten lanak ere aurkeztu dira, esaterako Kazama eta Torisawa (2007), Farkas *et al.* (2007) eta Vincze *et al.* (2011) egileek deskribatzen dituzten hurbilpenak, hain zuzen ere.

NERC ataza ebazteko estrategia desberdin ugari aurkeztu izan dira. Baina oinarrian hiru estrategia desberdin identifika daitezke (Kozareva, 2009): erregeletan oinarritutako estrategia, non ezaugarri linguistikoak diren oinarri; adibideetan oinarritutako estrategia, non etiketatutako adibideetatik sortutako eredu estatistikoak erabiltzen diren; eta, azkenik, bi metodologiak konbinatzen dituzten estrategiak.

Jarraian, estrategia bakoitzari azpiatal bat eskaini diogu, bakoitzaren inguruko hurbilpen interesgarrienak aurkezteko.

II.1.1 Erregeletan oinarritutako NERC sistemak

Erregeletan oinarritutakoa NERC ataza ebazteko sistema aitzindariek jarraitzen duten estrategia dela esan daiteke. Are gehiago, MUC-6 eta MUC-7 konferentzietan aurkeztutako sistema gehienak erregeletan oinarritutako sistemak dira. Domeinu bati eta, batez ere, hizkuntza bati lotutako adituek definitutako erregela multzo batek eta izendun entitateak biltzen dituzten zerrenda luzeek osatu ohi dituzte mota honetako sistemak. Arkitektura aldetik,

NERC sistema horiek oso antzekoak diren arren, jarraian MUCen lehiaketetan eta ondorengo urteetan aurkeztutako sistema onenen deskribapen laburra ematen da.

FASTUS sistema (Hobbs *et al.*, 1995) MUC-6n terrorismo domeinuko testuetan informazioaren erauzketarako sistema gisa aurkezten da. Izendun entitateen erauzketa informazioaren erregela sintaktikoak eta predikatu-argumentu patroiak erabiliz ebazten da, eta errore gehienak erakunde- eta toki-izenak erauztean gertatzen direla aipatzen dute.

LASIE I (Gaizauskas *et al.*, 1995) MUC-6n aurkeztutako sistema da ere. Sistema horrek kanpo- zein barne-ebidentzia erabiltzen ditu izendun entitateak identifikatu eta sailkatzeko. Batetik, ingelesezko *gazetteer*ak erabiltzen ditu: 150.000 sarrerako toki-izenen zerrenda, 2.600 erakunde-izeneko zerrenda eta, azkenik, 160 pertsona-izeneko zerrenda. Kanpo-ebidentziari dagokionez, sistemak izendun entitateen inguruan agertu ohi diren ingelesezko hitzen zerrendak (hitz eragileen zerrendak aurrerantzean) erabiltzen ditu. Izendun entitateen identifikaziorako sistemak erregeletan, bi zerrenda mota horietan eta informazio sintaktikoan oinarritzen dira, hain zuzen ere. Zehazki 206 dira eskuz definitutako erregelak. Sistemak MUC-6 ebaluazioan % 91ko asmatze-tasa lortu zuen. Errore gehienen jatorria erakunde- eta pertsona-izenen identifikazioa da kasu honetan.

Aurrekoaren bertsio hobetua den LASIE II (Humphreys *et al.*, 1998) tresna, MUC-7 edizioan aurkeztu zen. Oinarrian, LASIE I tresnaren erroreen azterketa egin ondoren, erroreak ebazteko 400 patroi berri gehituz garatutako sistema da. Erregelak gehitu arren, LASIE II-k MUC-7 edizioan ez ditu emaitzak hobetzea lortzen, besteak beste, erregelak definitzeko erabilitako testuak ez direlako ebaluazio-testuetako domeinu berekoak.

2002 urtean ingelesezko izendun entitateak identifikatu eta sailkatzeko erregelak eta baliabide semantikoak konbinatzen dituen sistema aurkezten da (Magnini *et al.*, 2002). Hizkuntzaren ezaugarri linguistikoak biltzen dituzten 200 erregelez gain, sistemak izen berezi zein hitz eragileen identifikaziorako *WordNet* baliabide lexikalean oinarritutako predikatu multzo bat erabiltzen du. Hizkuntzarekiko independenteak diren *WordNet* predikatuen erabilera sistema eleanitzetarako urrats gisa aurkezten dute, ingelesezko *WordNet*en bertsio baliokidea duten hizkuntzen sistemen garapenerako batik bat. Bestetik, *WordNet*en oinarritutako predikatuek izen berezien eta hitz eragileen zerrenden sorkuntza eta mantentzea errazten dute. Portaera aldetik, sistemak % 85eko asmatze-tasa du, errore-tasa handiena pertsona-izenen identifikazioan gertatzen da.

Azkenik, Maynard *et al.* (2003) egileek beraien lanean erregeletan oinarritutako ingelesezko NERC sistema bat Cebuera[4] hizkuntzara moldatzeko saiakera aurkezten dute. Egokitzapenerako, ingeleserako tokenizatzailea eta esaldiz esaldi banatzeko prozesua moldatzeaz gain, etiketatze morfosintaktikorako ingelesezko lexikoia Cebuerako lexikoiarekin ordezkatu, izen berezien zerrendak moldatu, eta izendun entitateak mugatzeko erregelak berridazten dira. Moldaketa guztia 10 egunetan burututa % 77,5eko asmatze-tasa duen sistema lortzen da.

Edozein kasutan, erregeletan oinarritutako sistemak erregelak definitzen dituen adituaren jakintzaren menpe egoteaz gain, erregelen mantentzea oso garestia da eta denboran zehar eguneratuta mantentzen zailak dira. Domeinu aldaketa ere ez da gardena horrelako sistemetan, erregelak eta batez ere zerrendak, domeinuaren araberakoak izan ohi direlako (LASIE IIren kasuan ikusi den bezala). Hori dela eta, azken urteetako joera ondoko atalean deskribatzen den estrategian oinarritutako sistemak garatzea izan da.

II.1.2 Adibideetan oinarritutako NERC sistemak

NERC sistema baten funtsa aldez aurretik ezagutzen ez diren izendun entitateak erauztea dela esan daiteke. Hasierako sistema gehienak erregeletan oinarritutakoak izan baziren ere, denborarekin adibideetan oinarritutako sistemak nagusitzen joan dira. Joera hau MUC-7 konferentzian argi geratzen da, aurkeztutako sistema guztien artean 8 erregeletan oinarritutakoak eta 16, berriz, adibideetan oinarritutako sistemak baitira.

Adibideetan oinarritutako NERC sistemek, izendun entitateen adibide positibo eta negatiboen multzo batetik abiatuz, atributuak erauzi eta adibide berriak identifikatu eta sailkatzeko ezagutzaren inferentzia dute helburu. Hortaz, adibideetan oinarritutako estrategian, izendun entitateak mugatuta eta sailkatuta duen testu-bilduma batetik abiatuta, ikasketa automatikoko metodoak aplikatuz atributuetan oinarritutako ereduak ikasten dira, ondoren etiketatu gabeko testuak prozesatzeko gai diren sistemak eraikitzeko. Adibide sailkatuen beharra dela eta, sistema gainbegiratuak (*supervised*) direla esaten da.

Adibideetan oinarritutako oinarrizko sistema batek testu berri batean izendun entitate gisa ikasketan erabilitako etiketatutako adibideetan ager-

[4] Cebuera austronesiar familiako hizkuntza da. Filipinetan hitz egiten dute 18 milioi pertsona inguruk, eta bere izenak Filipinetako Cebu uhartean du jatorria.

tzen diren izendun entitateen hitz berak mugatzen ditu soilik. Oinarrizko sistemaren asmatze-tasa ikasketa-corpusaren eta etiketatu beharreko testu berrien (ebaluazio-corpusaren) arteko lexikoi komunaren menpekoa izango da. Konkretuki, asmatze-tasa bi corpusen artean errepikapenik gabe bietan agertzen den hitz kopuruaren portzentajeak definitzen du, alegia, lexikoi-transferentzia tasak. Palmer eta Dayen (1997) lanean MUC-6 corpusaren lexikoi-transferentzia tasa % 21ekoa dela esaten da. Toki-izenen lexikoan % 42ko errepikapena ematen da, baina erakunde- eta pertsona-izenen lexikoietan soilik % 17 eta % 13ko errepikapena ematen da, hurrenez hurren. MUC-7 ataza partekatuko corpusaren gaineko oinarrizko sistemaren emaitzak Mikheev *et al.* (1999) egileen lanean eskuragarri daude: estalduran[5] toki, erakunde- eta pertsona-izen kategorietan % 76, % 49 eta % 26ko emaitzekin, hurrenez hurren, eta doitasunean, berriz,% 70 eta % 90 arteko emaitzak aurkezten dira.

Bikel *et al.* (1999) egileen lanean, ingeleserako MUC-6 eta MUC-7 ataza partekatuetako corpusak erabiliz eta gaztelerarako, berriz, MET-1 (*1st Multilingual Entity Task* (Merchant *et al.*, 1996)) atazakoak erabiliz Markoven Eredu Ezkutuen (HMM) metodoarekin garatutako sistemaren ebaluazioa aurkezten da. Ingelesezko ebaluazioan sistemak % 96ko estaldura eta % 93ko doitasuna badu ere, gaztelerazko corpusarekin ebaluatzean, berriz, emaitzak 3 puntutan jaisten dira estaldura zein doitasunean. Lan horretan, ikasketa-corpusaren tamainaren garrantzia islatzeko asmoz, sistemaren portaera corpus tamaina desberdinekin entrenatuta agertzen da.

2002 urtetik aurrera, CoNLLeren barruan izendun entitateen inguruko ataza partekatuan landutako corpus etiketatuei esker, adibideetan oinarritutako estrategiak indarra hartzen du. Izan ere, corpus horiekin estrategia hori izendun entitateen identifikazioa eta sailkapenean aplikagarri bihurtzen da, corpus horiek norberak garatzeko beharrik gabe. Zehazki, 2002 urteko[6] edizioan ebaluazioko corpusen hizkuntzak gaztelera eta nederlandera dira, eta 2003 urtean[7] berriz, ingelesa eta alemana. Bi edizioetara aurkezten diren tresnak izendun entitateak mugatu eta LOC, ORG, PER eta MISCELLANEOUS (MISC) kategorien artean sailkatu behar izan dituzte.

2002 urteko edizioan aurkeztutako sistema askok NERC problema NER eta NEC atazetan banatzen dituzte. Are gehiago horietako batzuk ataza

[5] Ebaluazio-neurriak II.2.4 atalean deskribatuta daude

[6] `http://www.cnts.ua.ac.be/conll2002/ner/`

[7] `http://www.cnts.ua.ac.be/conll2003/ner/`

bakoitzean algoritmo eta atributu desberdinak erabiltzen dituzte.

Testuak adierazteko atributuak eta ereduak trebatzeko ikasketa automatikoko metodo desberdinetan oinarritutako sistema ugari aurkeztu dira bi edizioetan. 2002 urtean, besteak beste, testuko hitzetatik erauzitako 9.000 atributu bitar erabiltzen dituen *Support Vector Machine* (*SVM*) metodoarekin lehenbizi identifikatu eta jarraian sailkatzeko trebatutako sistema aurkezten du McNamee eta Mayfielden (2002) lanak. Maloufen (2002) lanean, oinarrizko eredu estatistikoaren, Entropia Maximoko ereduaren eta Markoven Eredu Ezkutuen (HMM) portaeren azterketa aurkezten da, erabilitako atributu bildumarako Entropia Maximoak portaera onena lortzen duelarik. Testuko atributuez gain, *gazetteer*en informazioa baliatzen duten sistemak ere aurkeztu dira (Burger *et al.*, 2002) CONLLeko 2002ko edizioan.

Hizkuntzarekiko independentea den sistema aurkezten da Patrick *et al.* (2002) egileen lanean. Sistemak izendun entitateen identifikazioa eta sailkapena urrats anitzen ebazten du, ohiko azaleko atributuak erabili beharrean hizkuntzen ezaugarri estatistikoak erabiliz. Urrats bakoitzean ezaugarri estatistikoak eta aurreko urratsen emaitzak konbinatzeko sistemak erabaki-grafoak erabiltzen ditu. Estrategia horrekin sistemak gaztelerarako % 73,92 eta nederlanderako % 71,36 F_1 lortzen ditu, CoNLLeko 2002ko ataza partekatuan aurkeztutako 12 sistemetatik 7. eta 6. postuetan kokatuz, hurrenez hurren.

Carreras *et al.* (2002) egileen lanean, sakonera finkoko erabaki-zuhaitzei *AdaBoost* aplikatuz eraikitako *Abionet* izeneko sistema aurkezten da. Gaztelerarako % 81,39 eta nederlanderarako % 77,05 F_1arekin bi hizkuntzetan irabazlea da CoNLL 2002ko edizioan. Sistemak lehenbizi izendun entitateak identifikatzeko eta ondoren sailkatzeko atributu ugari erabiltzen ditu: hitz barneko ezaugarriak, informazio morfosintaktikoa nederlanderaren kasuan, hitz eragileen zerrendak gaztelerako sisteman, aurreko izendun entitateen kategoriak, etab., oro har testuingurua deskribatzeko atributu ugari erabiltzen ditu. *Abionet* sistema euskarazko entitate-izenak identifikatu eta sailkatzeko sistemaren garapenean erabili diren metodoetako bat da, hain zuzen ere (ikus II.4 atala).

2003 urteko edizioan, Entropia Maximoan oinarritutako sistema asko aurkezten dira, eredua bera bakarrik (Bender *et al.*, 2003), zein beste eredu batzuekin konbinatuz eraikitako sistemak (Klein *et al.*, 2003) oso emaitza onak lortuz. Edizio horretako sistema onena, hain zuzen ere, lau sailkatzaile, Entropia Maximoa, sailkatzaile lineal sendo bat, transformazioan oinarritutako eredua eta HMM ereduak, bakoitza atributu desberdinekin eraikita konbina-

tzen dituen sistema da garaile (Florian *et al.*, 2003). Sistemak ingeleserako % 88,76 eta alemanerako % 72,41eko F_1a lortzen du.

CoNLL-2002 zein CoNLL-2003 konferentzietan aurkeztutako sistemen portaerak aztertuta, orokorrean algoritmo desberdinen konbinazioarekin algoritmo bakarra erabiltzean baino sistema sendoagoak lortzen direla ondoriozta daiteke. Atributu mota desberdinak eta *gazetteer*ak erabiltzeak ere sistemaren emaitzak hobetu izan ditu. Sistema hauek guztiak, ordea, ikasketa gainbegiratuko algoritmoak dira, alegia, sarrerako adibide sailkatuen multzo handia behar dute sistema trebatu eta portaera ona lortu ahal izateko.

Sarrerako corpus etiketatu horren dependentzia hain handia ez duten metodoak ere erabili izan dira NERC problema ebazteko, ikasketa erdigainbegiratu izenarekin ezagutzen den estrategia jarraituz, hain zuzen ere. Mota horretako sistemak garatzeko *Bootstrapping* bezala ezaguna den teknika oso erabilia izan da, ez soilik NERC atazan, baizik eta HPko hainbat alorretan. Teknika horrek, sarrerako adibide sailkatu gutxi batzuen gainean ikasi eta ikasitakoa sarrerako testuen gainean aplikatuz, testuinguru antzekoan agertzen diren izendun entitateak mugatzen ditu. Mugatutako espresioak, hasierako adibideei gehituz, ikasketa-fasea berriz aplikatzen da, prozesu hau iteratiboki aplikatuz gelditze baldintza bete arte. Nadeau eta Sekinek (2007) egindako lanean sistema mota horien errepaso zabala aurki daiteke.

II.1.3 NERC sistema hibridoak

Sistema desberdinen konbinazioak hizkuntzaren prozesamenduko hainbat arlotan portaera ona erakutsi duen estrategia da, besteak beste, itzulpen automatikoan (Nomoto, 2004; Matusov *et al.*, 2006), testu sailkapenean (Larkey eta Croft, 1996), hitzen adiera-desanbiguazioan (Pedersen, 2000), zein informazio erauzketarako sistemetan (Banko eta Etzion, 2008). Euskararen gaineko hainbat atazetan ere sistema hibridoak garatu izan dira, esaterako *Eustagger* (Ezeiza *et al.*, 1998), euskarazko lematizatzaile/etiketatzailea, edota euskarazko azaleko sintaxiaren tratamendua (Arrieta, 2010).

NERC atazari dagokionean, sistema hibrido gisa erregelak eta ikasketa automatikoko teknikak konbinatzen dituzten sistemak ezagutzen dira. Sistema hauek banakako ereduen ezaugarri onenak erabiliz, emaitza hobeak lortzen dituzten sistemak eraikitzea dute helburu. Izendun entitateen arloan Mikheev *et al.* (1999) egileek egindako lanean aurkezten duten entropia maximoa eta eskuzko erregelak konbinatuz sortutako sistema, horren adibide da.

Sistemak eskuzko erregelen bitartez izendun entitateak identifikatzen ditu eta, testuinguruak informazio nahikoa eskaintzen badu, sailkapena ere burutzen du. Sailkatu gabe geratu diren NEak Entropia Maximoarekin (Mikheev *et al.*, 1998) sailkatzen ditu. Prozesu hau bi aldiz errepikatzen dute bigarren itzulian erregelak erlaxatuz.

Erregelak eta entropia maximoarekin batera Markoven Eredu Ezkutuekin (HMM) ikasitako eredua konbinatzen dituen ingelesezko NERC sistema aurkezten dute Srihari *et al.* (2001) egileek beraien lanean. Konkretuki sistemak lehen urrats batean, *TIMEX* eta *NUMEX* motako entitateak eraucteko egoera finituko transduktore bidez inplementatutako erregela bilduma aplikatzen du, eta jarraian aplikatzen diren moduluak *ENAMEX* motako entitateak ebazten dituzte. Konkretuki, pertsona- eta toki-izenak tratatzeko Entropia Maximoan oinarrituta ikasitako eredua aplikatzen du lehen. Ondoren, eskuz definitutako erregela multzo bat aplikatzen du eta hirugarren, HMM eredua aplikatzen du *ENAMEX* espresioak identifikatu eta sailkatzeko. Sistemak azken urrats bezala sailkapen xeheagoa lortzeko, Entropia Maximoko eredu bat aplikatzen du. Sistemaren emaitzak, garaiko erregela zein teknika estatistikoak independenteki erabiliz garatutako sistemak baino portaera hobea du, MUC-7ko datuen gainean ebaluatzean % 93,3ko F_1 (ikus II.2.4 atala) lortuz.

Arévalo *et al.* (2002) egileek gaztelerarako NERC sistema hibridoa aurkezten dute. Lan horretan bi entitate mota bereizten dira: ahulak eta sendoak. Entitate sendoak, bi motetatik entitate sinpleenak dira. Oinarrian izen bereziek osatzen dute talde hau. Ahulak, berriz, espresio konplexuagoak dira, kasurik sinpleenean izen arrunt eta entitate sendo batez osatutakoak. Esaterako, *Eusko Jaurlaritza* entitate sendo bat da, eta *Eusko Jaurlaritzako lehendakaria* berriz ahula. Aipatutako lanean entitate sendoak identifikatzeko ikasketa automatikoa eta ahulen erauzketarako erregeletan oinarritutako estrategia nola konbinatu deskribatzen da.

Borthwick *et al.* (1998) egileek NERC sistema desberdinen irteerak Entropia Maximoan oinarritutako eredua baten bidez konbinatuz, banakako sistemen emaitzak hobetzen dituzten sistema hibridoak deskribatzen dituzte, % 92 ko F_1a lortuz konbinazio onenaren kasuan. Sistema mota horien guztien azterketa sakona Kozarevaren (2009) tesi-lanean aurki daiteke.

Badira sistema berriagoak ere, 2007 urtean kokatzen diren Kazama eta Torisawa (2007) eta Farkas *et al.* (2007) egileen lanak kasu, Wikipedia edota Webeko corpusak ustiatuz NERC sistemen portaera hobetzeko saiakerak aurkezten dituztenak. Lehenbizikoan, esaterako, metodo estatistikoak entre-

natzeko Wikipediako informazioa erabiltzen dute. Deskribatzen duten sistemak lehenbizi sarrerako testuan Wikipedian sarrera duten hitz sekuentziak identifikatzen ditu. Ondoren, identifikatutako sekuentzia bakoitzeko eta dagokion Wikipedia sarrerako lehen esaldia erabiliz, sailkapen bat eskuratzen du. Hitz sekuentzia eta lortutako sailkapen hori dira, hain zuzen ere, eredu estatistikorako informazio-iturri. Farkas *et al.* (2007) egileen lanean berriz, Florian *et al.* (2003) egileen NERC sistemaren emaitzak hobetzeko Web-eko eta Wikipediako ezagutza ustiatzen dituzten hiru heuristiko aurkezten dira.

Gogoratu, helburuetan esan dugun bezala, gure helburua euskara entitate-izenak identifikatu eta sailkatzeko ikasketa automatikoa eta erregelak konbinatuz sistema hibrido bat garatzea dela, beti euskararen ezaugarri morfosintaktikoak kontuan hartuz. Alegia, ez dugu metodo zehatzetan ikertu nahi, baizik eta baliabide eta denbora mugatuarekin sistema bat garatu nahi dugu.

II.2 Ingurune esperimentala

NERC sistema bat eraikitzeko bide eta informazio-iturri desberdin ugari daude. Ondorengo lerroetan euskararako aukeratu eta erabili direnak aurkezten dira. Arestian aipatu bezala, izendun entitateak hiru multzo nagusitan bereizi izan dira: *ENAMEX* edo entitate-izenak, *TIMEX* edo denborazko espresioak eta *NUMEX* edo zenbakizko espresioak. Tesi-lan honetan, lehen multzokoak bakarrik tratatuko dira, entitate-izenak alegia, data eta zenbakizko adierazpen gehienak (TIMEX eta NUMEX taldeetako mota gehienak), *IXA talde*ko beste lanetan ebatzi baitira (Ezeiza *et al.*, 1998; Alegria *et al.*, 1999; Soraluze *et al.*, 2011). Hortaz, tesi-lan honetan euskarazko NERC sistemari buruz hitz egitean, euskarazko entitate-izenak (NE aurrerantzean) identifikatu eta sailkatzeko sistemaz arituko gara, eta ez izendun entitateei buruz, data eta zenbakizko adierazpenen tratamendua lan honen esparrutik kanpo geratzen baitira.

Aurrerantzean, beraz, NER laburdura entitate-izenen identifikazioa adierazteko erabiliko dugu. Berdin NEC eta NERC laburdurekin, entitate-izenen sailkapena eta entitate-izenen bi atazak erreferentziatzeko erabiliko dira, hurrenez hurren, eta ez izendun entitateen ingurukoak aipatzeko.

II.2.1 Baliabideak

McDonalden (1996) arabera, bi informazio-iturri mota bereizi behar dira entitate-hautagaiak bilatzeko eta sailkatzeko garaian: alde batetik, entitate-espresioak berak eta bera osatzen duten elementuek barneratzen duten informazioa; eta bestetik, entitatearen agerpenaren testuinguruak eskainitakoa, barne- eta kanpo-ebidentzia, hurrenez hurren.

NE berak eta bere hautagaiek eskaintzen duten informazioa ahalik eta gehien ustiatzeko asmoarekin, *Eustagger* (Ezeiza *et al.*, 1998), euskararako lematizatzaile/etiketatzailea erabili da. Tresna horrek oinarrian euskarazko testuen analisi eta desanbiguazio morfosintaktikoa burutzen du. Hortaz, horri esker, NEak jorratzeko hitzen azaleko informazioarekin lan egiteaz gain, informazio linguistiko zehatza baliatu ahal izan dugu. Informazio mota hori, aurrerago ikusiko den bezala, oso erabilgarria gertatzen da NER zein NEC atazetan. Are gehiago euskara bezalako hizkuntza eranskari batekin lan egin behar denean.

Bi izan dira kanpo-ebidentzia gisa erabili ditugun baliabideak: *trigger word* edo hitz eragileen zerrendak eta *gazetteer* edo NEen zerrendak. Esan bezala, hitz eragileen zerrendek entitateen testuinguruan agertu ohi diren hitzen informazioa gordetzen dute. Hitz horiek testuan zehar NEak identifikatzeko lagungarri dira, horien mugetan NEak agertu ohi baitira. Are gehiago, hitz eragileak NEen sailkapen zehatzetan erabiltzen dira. Lotura horri esker, identifikazio atazan ere erabiltzeaz gain sailkapen atazan lagungarri dira. Esaterako *jokalari* hitza, pertsona-izenen inguruan agertu ohi da, *Markel Susaeta jokalariak Eibarren hasi zuen bere karrera...* esaldian ikus daitekeen bezala. Hortaz, NE kategoria bakoitzeko (PER, LOC, ORG), hitz eragileen zerrenda bat erabili dugu.

Hitz eragileen zerrenda horiek sortzeko ekonomiari buruzko corpus txiki batetik abiatu gara. Bertan sailkapen-kategoria bakoitzeko bi hizkuntzalarik eskuz identifikatutako hitz eragileekin zerrendak osatu dira: 29 hitz eragile PER kategorian, 71 LOC kategorian; eta azkenik 11 ORG kategorian.

Zerrenda txikiegiak izanik, Magnini *et al.* (2002) egileek hitz eragileen zerrendak sortzeko *WordNet* (*WN*) ontologia erabiltzen duten bezala, zerrendak aberasteko asmoz Pociellok (2008) bere tesi-lanean deskribatzen duen Euskal WordNet (*EusWN*), euskarazko bertsioa erabili da. Aberasketa-prozesu horretan abiapuntua Magnini *et al.* (2002) egileen hitz eragileen zerrendak izan dira. Konkretuki, NEC atazako hiru sailkapen kategorietarako ingelesezko hitz eragileen zerrendak publikatuta daude, hitz eragileak dago-

kien *WN synset*arekin batera errepresentatuta. *Synset* batek edozein hizkuntzatako *WN*ean kontzeptu berdina errepresentatzen duenez, ingelesezko hitz eragileen zerrendetako hitzen *synset*ak *Eus WN*en kontsultatuz euskarazko hitz eragileen zerrenda baliokideak lortzen dira. Prozesu horrekin, II.1 taulan ikus daitekeen bezala, sailkapen guztietan aberasketa nabarmena lortzen bada ere, pertsona-izenen zerrenda da gehien hazten dena, 29 hitz izatetik 3.825 hitz eragile izatera pasatuz.

NE mota	**eskuz**	***Eus WN* aberasketarekin**
PER	29	3.825
LOC	71	895
ORG	11	594

II.1 Taula: Euskarazko hitz eragileen zerrenden tamainak

Gazetteer zerrendak, bestetik, maiztasun handiko NEen sailkapenak zehazten dituzten zerrendak dira. Alegia, kategoria zehatz bateko entitate-izenak biltzen dituzten zerrendak dira, eta sailkapen-atazan erabili ohi dira batez ere. Euskarazko NEC sistema eraikitzeko, sailkapen mota bakoitzeko *gazetteer* zerrenda bat erabili da.

Gazetteer horiek *Euskaldunon Egunkaria* egunkaritik eskuratu dira, eta PER kategoriako zerrenda Eusko Jaurlaritzak luzatutako erroldarekin osatu da. Zerrenda hauetarako, hitz eragileen zerrendak aberasteko *Eus WN* erabiliz egindako aberasketa prozesu bera aplikatu da. Hala ere, azken aberasketa hau hitz eragileetan lortutakoa baino txikiagoa izan da, II.2 taulan ikus daitekeen bezala.

NE mota	**egunkaritik**	***Eus WN* aberasketarekin**
PER	121.503[1]	121.677
LOC	1.444	1.547
ORG	973	1.295

[1]Pertsona-izenen zerrenda Eusko Jaurlaritzak luzatutako erroldarekin osatu da

II.2 Taula: Euskarazko *gazetteer* zerrenden tamainak

Zerrenda horietan PER sailkapen-zerrenda handiena bada ere, aberasketa handiena ORG kategorian ematen da 322 instantzia berrirekin. Aberasketa horretan, *Antzerki Grekoko Korua* edo *Estatu Batuetako Senatu* bezalako

NE internazionalagoak gehitzen dira. Esan bezala, hitz eragileetan *EusWN* aberasketa *gazetteer*etan baino handiagoa da. Bi aberasketen emaitzak alderatuz ondoriozta daiteke, beraz, *EusWN* oso lagungarri izan daitekeela NEen testuingurua definitzeko eta ez horrenbeste NEaren sailkapena ebazteko garaian, euskararen kasuan behinik behin. Izan ere, *WN* zein *EusWN*en NEen inguruko informazio gutxi dago landuta, eta NERC atazan bertako izen arruntak eta adjektiboak dira benetan lagungarri gertatzen direnak. Hori bera da, aberasketa nabarmenena ORG kategorian ematearen arrazoia, tokieta pertsona-izenetan ez bezala, oso ohikoa baita erakunde-izenetan mota horretako osagaiak agertzea.

Aurretik aipatu bezala, euskarazko NERC tresna garatzeko bi bide desberdin jorratu dira: lehenbizi erregeletan oinarritutako sistema bat garatu da, eta ondoren adibideetan oinarritutako teknikak aplikatuz beste hainbat hurbilpen egin dira. Teknika bakoitzak baliabide desberdinen beharra du.

Lehenengoak, erregeletan oinarritutako sistemak, erregelak berak definitzeko tresna bat beharrezkoa du. Tesi-lan honetan *Xerox Finite State Transducer* (*XFST*) helburu orokorreko egoera finituko automatak sortzeko tresna erabili da (Beesley eta Karttunen, 2003). Adierazpen erregular multzo bat egoera finituko automata ezagutzaile zein itzultzaile bihurtzeaz gain, hauen arteko eragiketa desberdinak aplikatzeko aukera ematen du, besteak beste, automaten arteko bildurak, konposaketak, etab. Sortutako automatak edo transduktoreak formatu bitarrean zein testu-fitxategietan gorde daitezke, bere edizioa ahalbidetuz. Sarrera gisa ezaugarri edo transformazio multzo bat definitzen duten adierazpen erregularrak konpilatuz, sarrerako kate batek ezaugarriak betetzen dituen ebatzi edo katearen bihurketa burutzeko automata eraikitzeko gai da *XFST*. Tresna aproposa da, hortaz, NER atazako erregelak definitzeko. Izan ere, tresna hori ingelesezko datak identifikatzeko erabili izan da (Karttunen *et al.*, 1996). Horretarako, II.2.1 adibidean aurkezten den bezala, datetan agertu ohi diren elementuak definituz, ondoren elementu horien arteko ordenak eta elkarrekintzak definitzen dituen patroia eta azkenik, patroia betetzean, mugatzeko erregela definitzen dute.

II.2.1 Adibidea Ingeleseko datak mugatzeko *XFST* automataren definizioa.

```
1. 1To9 = [1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9]

2. 0To9 = [%0 | 1To9]

3. SP = [", "]

4. Day = ["Monday" | "Tuesday" | ... | "Saturday" | "Sunday"]
```

```
5. Month = ["January" | "February" | ... | "November" | "December"]
6. Date = [1To9 | [1 | 2] 0To9 | 3 [%0 | 1]]
7. Year = 1To9 (0To9 (0To9 (0To9)))
8. DateExpression = Day | (Day SP) Month " " Date (SP Year)
9. DateExpression @-> %[ ...  %]
```

Egileek lehenengo 5 definizioekin datetan agertu ohi diren oinarrizko osagaiak zehazten dituzte (zenbakiak, asteko egunak, hilabeteak, etab.). Eguna definitzeko berriz, oinarrizko osagaiekin murriztapenak definitzen dituzte (II.2.1 adibideko 6. lerroa), egun bat 1etik 31ra arteko balioa izan dezakeela mugatzeko murriztapena, alegia. Antzera urtea definitzeko (II.2.1 adibideko 7. lerroa), non lehenengo zenbakia 0 ez den beste edozein zenbaki izatea mugatzen duen espresioa definitzen duten. 8. lerroko espresioak data baten agerpen posible guztiak definitzen ditu: eguna, hilabetea, egun-zenbakia, etab. Horrela, azken lerroko erregelarekin, patroi hori betetzen duen espresio batekin topo egitean, kakoen artean mugatzeko erregela definitzen dute, dagokion sarrerako testutik ondoko irteera lortuz: *Today is [Wednesday, August 28, 1996] because yesterday was [Tuesday] and it was [August 27].*

Kasuistika konplexuagoa bada ere, entitate-izenak mugatzeko ere *XFST* erabilgarri izan daiteke, beraz.

Adibideetan oinarritutako sistemek, berriz, erregela horiek automatikoki sortzeko eta ebaluatzeko adibide etiketatuak behar dituzte. Datu horien deskribapena egiten da ondoko atalean, hain zuzen ere.

II.2.2 Corpus etiketatuak

Adibideetan oinarritutako sistemetan ebaluaziorako adibide sailkatuak beharrezkoak dira eta, aukeratutako algoritmo edo teknika gainbegiratua bada, algoritmoak ikas dezan ere. Ataza zehatz baterako etiketatutako testu multzoak lortzea ez da lan erraza izaten, are gutxiago euskara bezalako baliabide murritzeko hizkuntzarekin lan egitean.

Lan honen hasieran NERC problemarako euskarazko testu etiketatu eskuragarririk ez zegoenez, corpusa guk eraiki genuen bi helburu nagusirekin: batetik ikasketa automatikoko (IA) tekniketan oinarritutako euskarazko NERC sistema garatu ahal izateko; eta bestetik, garatutako sistemak ebaluatu ahal izateko.

Corpus etiketatua eraikitzen hasterako, erregeletan oinarritutako euskarazko NER sistema garatuta zegoen. Horrek, corpus etiketatua modu erdiautomatikoan sortzea ahalbidetu zigun. Lehenbizi testuei ezaugarri linguistikoetan oinarritutako tresna aplikatu zitzaien, eta ondoren aditu batek berrikuspena eta zuzenketa burutu zuen. Prozesu hau *Euskaldunon Egunkaria*ko 2000 urteko gai desberdinetako 383 albistez osatutako bildumari aplikatu zitzaion. Adituaren berrikuspen eta zuzenketaren ondoren, 62.187 hitz eta 4.748 NEek osatutako corpusa lortu zen.

Corpus hau % 75 eta % 25 proportzioa aplikatuz ikasketa- eta ebaluazio-corpusetan banandu da, hurrenez hurren. Zatiketa horrekin 46.227 hitz eta 3.817 NE geratzen dira ikasketarako eta 15.960 hitz eta 931 NE[8] ebaluaziorako.

Corpuseko etiketatze formatuari dagokionez, *CoNLL* konferentzian definitutako BIO (Tjong Kim Sang, 2002) formatua erabili da. Errepresentazio horrek NER atazan NEak mugatzeko ondoko etiketak erabiltzen ditu:

- **B** (*Begin*): NEaren hasiera adierazteko.
- **I** (*Intermediate*): NEaren tarteko osagaiak identifikatzeko.
- **O** (*Out*): NEaren parte ez diren osagaiak adierazteko.

NEC atazako kategoriak definitzeko, berriz, MUCeko ohiko kategoriak erabili dira:

- **PER**: pertsona kategoriako entitateetarako.
- **LOC**: tokia adierazten duten espresioetarako.
- **ORG**: erakunde-izenetarako: enpresa, erakunde zein instituzioak sartzen dira.
- **OTH**: aurreko hiruak ez diren bestelako NEak sailkatzeko etiketa. Adibidez, liburuen edo filmen izenburuak.

Corpusa osatzen duen berri bakoitzeko fitxategi bat dago, non lerro bakoitzean hitz bat eta bere informazioa gordetzen den. Aurretik aipatu den bezala, eta NERC sistemaren diseinuko atalean azaltzen den bezala (ikus II.3.2 atala), NEak identifikatu eta sailkatzeko euskarazko NERC sistemak *Eustagger*ek itzulitako informazio morfosintaktikoa erabiltzen du.

II.2.2 Adibidea Corpus etiketatuaren formatua.

[8]Ebaluazio-corpus bera erabili da kapitulu honetako esperimentu eta tresna guztien ebaluaziorako.

Efektu	*efektu*	*IZE_ARR*	*EZZ*	*O*
bereziak	*berezi*	*ADJ_IZO*	*ABS*	*O*
ez	*ez*	*PRT*	*EZZ*	*O*
zuela	*ukan*	*ADT*	*EZZ*	*O*
inolako	*inolako*	*ADJ_IZL*	*EZZ*	*O*
arazorik	*arazo*	*IZE_ARR*	*PAR*	*O*
sortu	*sortu*	*ADI_SIN*	*EZZ*	*O*
azaldu	*azaldu*	*ADI_SIN*	*EZZ*	*O*
zuten	**edun*	*ADL*	*EZZ*	*O*
egunean	*egun*	*IZE_ARR*	*INE*	*O*
zehar	*zehar*	*ADB_ADOARR*	*EZZ*	*O*
Eusko_Jaurlaritzak	*Eusko_Jaurlaritza*	*IZE_LIB*	*ERG*	*B-ORG*
,	*,*	-	-	*O*
Nafarroako	*nafarroa*	*IZE_LIB*	*GEL*	*B-ORG*
Gobernuak	*gobernu*	*IZE_ARR*	*ABS*	*I-ORG*
,	*,*	-	-	*O*
Iberdrolak	*Iberdrola*	*IZE_LIB*	*ERG*	*B-ORG*
,	*,*	-	-	*O*
Bilbao	*Bilbao*	*IZE_IZB*	*EZZ*	*B-ORG*
Bizkaia	*Bizkaia*	*IZE_LIB*	*EZZ*	*I-ORG*
Ur	*ur*	*IZE_ARR*	*EZZ*	*I-ORG*
Partzuergoak	*partzuergo*	*IZE_ARR*	*ABS*	*I-ORG*
eta	*eta*	*LOT_JNT*	*EZZ*	*O*
beste	*beste*	*DET_DZG*	*EZZ*	*O*
hainbat	*hainbat*	*DET_DZG*	*EZZ*	*O*
erakunde	*erakunde*	*IZE_ARR*	*EZZ*	*O*
eta	*eta*	*LOT_JNT*	*EZZ*	*O*
enpresak	*enpresa*	*IZE_ARR*	*ABS*	*O*
.	*.*	-	-	*O*

II.2.2 adibidean, ebaluazio-corpuseko berri bateko testu etiketatu zati bat ikus daiteke, non hitz bakoitza 5 zutabetan deskribatzen den: forma, lema, kategoria/azpikategoria, deklinabide-kasua eta, azkenik, dagokion BIO etiketa. Azken zutabean B eta I kasuetan NEC atazako zein kategoria dagokion (PER/LOC/ORG/OTH) ere gordetzen da. Bertan ikus daitekeenez, NEen parte ez diren osagai guztiek O etiketa dutela eta, aldiz, entitate-osagaiak direnak B edo I etiketak, entitatearen lehen osagaia edo bestelako NEaren osagaia den arabera, hurrenez hurren. I osagaiaren aurreko osagaia, I edo B etiketa izango da beti, kontrako kasuan, etiketatze okerra dela kontsideratuko da. Izan ere, NEaren tarteko osagai baten aurrean, beste tarteko bat edo hasierako NEaren osagaiak besterik ezin baitute egon. 3. zutabeko kategoria/azpikategorien deskribapena, EDBLko kategoria sistema deskriba-

tzen duen eranskinean eskuragarri dago (ikus B eranskina). 4. zutabeko deklinabide-kasuen esanahia aldiz, jarraian aipatzen da:

- EZZ edo - karakterea: kasu markarik ez.
- ABS: absolutiboa.
- PAR: partitiboa.
- INE: inesiboa.
- ERG: ergatiboa.
- GEL: leku/denborazko genitiboa.

Jarraian aurkezten den metodoen atalean (ikus II.2.3 atala) ikusiko den bezala, esperimentuetarako erabilitako hainbat IAko metodo aplikatzeko *Weka* (Hall *et al.*, 2009) paketeko inplementazioak erabili ditugu. Eta pakete horretako metodoak erabili ahal izateko, datuak formatu berezi batean egon behar dira, *arff* formatuan, hain zuzen ere. Hori dela eta, II.2.2 adibideko formatua IAko hainbat esperimentuetarako *arff* formatura bihurtu behar izan dugu.

II.2.3 adibidean, NER ataza ebazteko *arff* formatuan sarrerako testuetako hitzak nola adierazi diren ikus daiteke. *arff* formatuan, bi atal nagusi bereiz daitezke: sailkapen problemaren instantzia bakoitzeko osagaiak adierazteko erabiliko diren atributuen deskribapenaren atala batetik; eta bestetik, datuen atala. Bi atalen arteko bereizketa *@DATA* lerroaren bitartez egiten da. Atributuen deskribapena zein osagaien deskribapena, bakoitza lerro batean adierazten da. Lehen ataleko lerro bakoitzean atributu bati dagokion informazioa zehazten da, eta bigarrenean, aldiz, lerro bakoitzak adibide baten informazioa bilduko du. Adibideak, lehen atalean deskribatutako atributuen balioen bitartez eta ordena berean deskribatzen dira.

II.2.3 Adibidea NER problema ebazteko *arff* formatua.

```
@RELATION     eu_NER
@ATTRIBUTE    word       REAL
@ATTRIBUTE    lemma      REAL
@ATTRIBUTE    pos        IZE_ARR,ADJ_IZO,ADI_SIN,ADL,SIG_DEK
@ATTRIBUTE    up         0,1
@ATTRIBUTE    cap        0,1
@ATTRIBUTE    up_lem     0,1
```

```
@ATTRIBUTE   cap_lem   0,1
@ATTRIBUTE   class     B,I,O
@DATA
2165         3942      IZE_ARR 1 0 0 0 O
11282        5420      ADJ_IZO 0 0 0 0 O
6445         2985      ADI_SIN 0 0 0 0 O
6742         34        ADL 0 0 0 0 O
1484         788       SIG_DEK 1 1 1 1 B
274          253       IZE_ARR 0 0 0 0 O
36           38        - 0 0 0 0 O
```

Esan bezala, lehen ataleko lerroak atributuak deskribatzeko erabiltzen dira. @ATTRIBUTE etiketarekin hasten dira, ondoren atributuaren izena eta azkenik bere balioak definitzen dira. Adibidean ondoko atributuak eta murriztapenak definitu dira: forma (*word*) eta lema (*lemma*) zenbaki errealekin adieraziko direla definitzen da, kategoria/azpikategoria (*pos*) atributuak zerrendatzen diren balioetako bat besterik ezingo du izan, forma zein lema letra larriz (*up, up_lem*) hasten diren eta forma zein lema osorik letra larriz (*cap, cap_lem*) idatzita dauden adierazteko atributu bitarrak definitzen dira, eta azkenik ebatzi beharreko problemaren balio posibleak, NER atazarako adibidea izanik B,I,O kategoriak, hain zuzen ere.

Formak eta lemak zenbaki errealekin adieraztea beharrezkoa da, *arff* formatuak zerrendatu ezin daitekeen baliorik ez baitu onartzen. Hortaz, *arff* formatuan, ikasketa-corpusa osatzen duten hitz eta lema guztiak zerrendatu ordez, horiekin hiztegi bat sortu ohi da hitz/lema bakoitzerako identifikadore bat definituz, eta identifikadore hori erabiliko da, hain zuzen ere, *arff* formatuan forma/lemaren agerpenak errepresentatzeko.

Datuen atalean lerro bakoitzeko testuko hitz baten informazioa definitzen da, atributu bakoitzari dagokion balioa zehaztuz, aurreko atalean deskribatutako ordena berean eta tabulazioz banatuta. II.2.3 adibideko testua ikasketa-corpuseko adibidea izanik, sailkapen-problemako kategoria ezaguna da eta hortaz balioa adierazita dago. Adibide berrien kasuan, sailkapen-balioa ezezaguna denez, azken atributuaren balioa *?* ikurrarekin definituko da, balioa ezezaguna dela adierazteko. Ikur hori, atributu baten balioa ezezaguna adierazteko ere erabili ohi dira.

II.2.3 Ikasketa automatikoko metodoak

Ikasketa automatikoan oinarritutako metodoak hiru multzo nagusi hauetan banatu ohi dira:

- Ikasketa gainbegiratuan oinarritutako metodoak: ikasketa-prozesua burutu aurretik aurrez etiketatuko dokumentu bilduma bat eskuragarri dago, eta berau erabiltzen da sistema trebatzeko eta sailkatzailea eraikitzeko, ondoren, adibide berriak datozenean modu automatikoan sailkatzeko.

- Ikasketa ez-gainbegiratuan oinarritutako metodoak: ez da erabiltzen aurrez etiketatutako dokumenturik ezta aldez aurreko sailkapenik. Ikasketa itsu-itsuan burutzen da.

- Ikasketa erdi-gainbegiratuan oinarritutako metodoak: ikasketa-prozesua burutu aurretik aurrez etiketatuko dokumentu multzo txiki bat eskuragarri dago, eta berau etiketatu gabeko dokumentu multzo handi batekin erabiltzen da sistema trebatzeko eta sailkatzailea eraikitzeko, ondoren adibide berriak datozenean modu automatikoan sailkatzeko.

NER eta NEC, biak sailkapen ataza gisa modelatu dira. NERen kasuan, testuko osagai bat entitate baten parte den edo ez erabaki behar da (ikus BIO formatua II.2.2 atalean), eta NECen berriz N kategorien artean egokiena aukeratu behar da. Planteamendu horrekin, eta garaiko bibliografia aztertu ondoren ikasketa-corpusaren beharra egon arren, euskarazko NERC sistema garatzeko ikasketa gainbegiratuko metodoak erabiltzea erabaki zen. Metodo askoren azterketa egin ondoren eta bakoitzak NERCen agertzen dituen alde positiboak eta negatiboak baloratu ostean, ondorengo hauekin esperimentatzea erabaki zen:

- *Naïve Bayes* (*NB*): sarrerako instantziak atributu askorekin errepresentatzen direnerako interesgarria, atributu asko eraginkortasunez konbinatzeko duen gaitasuna dela eta (Mitchell, 1997).

- Erabaki-zuhaitzak: atributuetan balio galduak dauden kasuetarako egokiak. Gainera, sailkatzailearen zuhaitz-egiturak sailkapenaren bidea jarraitzea posible egiten du. HP arloko ataza askotan emaitza onak eman ditu.

- *Support Vector Machine* (*SVM*): atributu multzo itzelekin portaera ona izan ohi du. Ikasketa-fasea motela izan arren, testean erantzun azkarra eman ohi du. Esperimentuak egin ziren garaian, NERCen emaitzarik onenetakoak ematen zituen (McNamee eta Mayfield, 2002).

- *AdaBoost*: atributu-espazio handietan ondo moldatzen den algoritmoa da. HP ataza askotan emaitza onak erakutsi ditu, eta UPC TALP taldeak emaitza bikainak lortu zituen NERC problema algoritmo horrekin ebaztean (Carreras *et al.*, 2002).

IAko metodoak inplementatzen eta eskuragarri jartzen dituzten tresna ugari aurki daitezke. Oinarrian algoritmoak berdinak izan arren, inplementazio bakoitzak bere formatu propioa erabiltzen du. Ahalik eta formatu kopuru gutxien erabiltzeko asmoz, *Weka* paketea aukeratu da tesi-lan honetako IAko esperimentu gehienak burutzeko. Gehienak eta ez denak, Adaboosten kasuan NERC ebazteko TALP taldekoen inplementazio egokitua erabili baita. Gainerakoetan, IAan egoki agertu diren Wekako sailkatzaile sinpleak erabili ditugu.

Jarraian, euskarazko NERC sistema eraikitzeko aipatutako lau algoritmoak aurkezten dira.

II.2.3.1 *Naïve Bayes*

Algoritmo estokastiko sinple hau sarrerako adibideak deskribatzen dituzten atributu guztiak beraien artean independenteak direneko hipotesitik abiatzen da. Sailkapen-problema gehienetan atributu askok beraien artean gordetzen dituzten erlazioak direla eta, independentzia baldintza hori betetzen ez den arren, hipotesi horrek suposatzen duen sinplifikazioak sailkatzaile eraginkorrak lortzen laguntzen du. HP alorreko ataza desberdinetan portaera egokia erakutsi izan du, besteak beste testuinguruaren araberako zuzenketa ortografikoan, etiketatze morfosintaktikoan, eta dokumentuen sailkapenean, Arregi eta Fernández ek (2002) egindako lanean ikus daitekeen bezala.

Grafoen bidez errepresentatu ohi dira eredu estokastikoek kalkulatzen dituzten sarrerako atributuen dependentzia probabilistikoak. Grafoko adabegi bakoitzak zorizko aldagai bat adierazten du eta probabilitate-banaketa bat du esleituta. Aurretik aipatu den bezala, algoritmo honek atributu asko eraginkortasunez konbinatzeko gaitasuna du (Mitchell, 1997) eta portaera egokia erakutsi du sarrerako datuen multzoa oso handia denean (StatSoft, 2007).

Adibide baten klasea iragartzeko adibidearen probabilitatea maximizatzen duen klasea aukeratzen da. Horretarako, Bayes-en teorematik eratorritako formula sinple bat erabiltzen da, non atributu guztiei dagozkien balioak emanik $(a_1, a_2, \ldots a_n)$ emaitza-atributuaren klase probableena (V_{nb}) aukeratzen duen:

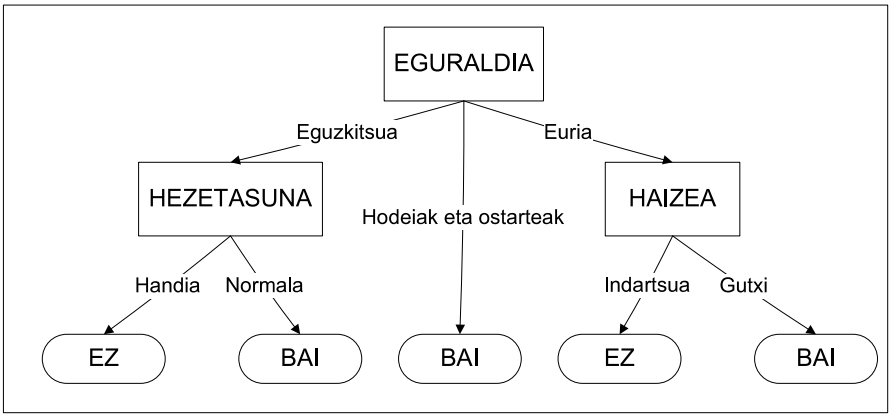
$$V_{nb} = argmax_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

non v_j emaitza-atributuak har dezakeen j. balioa den.

Askotan, besteak baino metodo sinpleagoa delako, oinarrizko neurri (*baseline*) gisa erabiltzen da metodo hau.

II.2.3.2 Erabaki-zuhaitzak

Estrategia jalean oinarritutako algoritmoak dira erabaki-zuhaitzak. Izenak adierazten duen bezala, algoritmo hauek sortzen dituzten ereduak zuhaitz itxurakoak dira. Erabaki-zuhaitzak sortzeko prozesua modu errekurtsiboan azal daiteke (Quinlan, 1993). Lehendabizi, atributu bat aukeratu behar da erro-adabegian kokatzeko, eta bere balio posible bakoitzeko adar bat egingo da. Gero, prozesua errepikatzen da errekurtsiboki, adar bakoitzerako, baina adar bakoitzeko baldintzak bete dituzten adibideekin soilik. Adabegi-ume bakoitzeko adibide guztiek sailkapen bera dutenean amaituko da prozesua, kasu horretan ezingo baita adabegia gehiagotan banatu; adabegi hori, beraz, hostoa izango da. Erabaki beharreko gauza bakarra, eskema honetan, zera da: une bakoitzean aukeratu beharreko atributua. Unean uneko atributuaren aukeraketak, ordea, berebiziko garrantzia dauka, behin atributu hori erabili eta gero ez baita gerora hartuko diren erabakietan atzera egingo.



II.1 Irudia: Tenisean jolastu Bai/Ez sailkapenerako erabaki-zuhaitza

II.1 irudian eguraldiaren araberako hiru atributuren (eguraldia, hezetasuna eta haizea) menpe tenisean jolasteko aukera egongo den (BAI) edo ez (EZ) sailkatzeko erabaki-zuhaitza ikus daiteke. Erabaki-zuhaitz horrek,

hiru atributu horien informazioaren arabera, edozein instantzia berri sailkatu ahal izango du. Erabaki-zuhaitzetan oso garrantzitsuak dira hasierako aukeraketak; erro-adabegiko atributuak berak zuhaitzaren alde batera edo bestera eramaten duelako, eta erabaki hori txarra bada, ez baitu sailkapen egokia egingo. Eskema honen abantailetako bat, ordea, zera da: eskuratutako ezagutza, sortutako eredua, alegia, adierazpide ulergarri batean hierarkia bat osatuz jar daitekeela eta hierarkia horretatik erregelak erauz daitezkeela. Hala, gaian aditua denak erregela horiek interpreta ditzake. Desabantailen artean, zenbakizko atributuak erabiltzeko duen problema aipa daiteke. Zenbakizko atributuak erabiltzeko moldaketaren bat egin behar da eskema hau erabili ahal izateko, moldaketa ohikoena diskretizazioa izanik.

Algoritmo mota honetan sarreren errepresentaziorako erabiltzen den atributu kopuruak eragin zuzena du sailkatzailea eraikitzeko behar den denboran, zenbat eta atributu gehiago, orduan eta zuhaitz-egitura konplexuagoa eraiki behar baita eta atributu-aukeraketa zailtzen baita. Adibide berrien sailkapenerako denboran ere eragina du, zenbat eta konplexuagoa izan orduan eta denbora gehiago behar baita zuhaitz konplexua aztertzeko. Horregatik, sarritan erabaki-zuhaitzekin batera inausteko teknikak aplikatzen dira, zuhaitza optimizatzeko asmoz.

HPko ia ataza guztietan erabili izan dira erabaki-zuhaitzak. Emaitza onak lortu izan dira ahotsaren ezagutzan, morfosintaxiaren etiketatzean, desanbiguazio semantikoan, laburpenen sorkuntzan, dokumentuen sailkapenean, itzulpen automatikoan eta analisi sintaktikoan, Witten eta Frankek (2005) eta Arrietak (2010) beraien lanetan aditzera ematen duten bezala.

Tesi-lan honetan aurkezten diren esperimentuak burutzeko erabaki-zuhaitzen familiako *C4.5* algoritmoa aukeratu da. Algoritmoa, erabaki-zuhaitzak eraikitzeko *ID3* (Quinlan, 1993) algoritmoan oinarritzen da. Zuhaitzaren erregela baliokideak kalkulatzen dira, eta balioztatze gurutzatua aplikatuz erregelak inausiz joango da, sailkapen errore minimora iritsi arte. Errorea minimizatzen duen erregela multzoa izango da, hain zuzen ere, sailkatzailea osatuko duena.

II.2.3.3 *Support Vector Machines*

Vladimir Vapnik eta AT&T laborategiko taldeak garatutako gainbegiratutako algoritmo multzo batek osatzen du *Support Vector Machines* edo Sostengu-Bektore bidezko Makinak izenarekin ezagutzen den metodoa (Cortes eta Vapnik, 1995).

Oinarrian, *SVM*k entrenamenduko instantziak n dimentsioko espazioan puntuen bidez irudikatzen ditu, eta espazio bat bilatzen saiatzen da non kategoriak ahalik eta bananduen geratzen diren. Horrela, adibide berri bat sailkatu nahi denean, espazioan puntu baten bidez irudikatzen du, eta puntu horren eta ikasitako kategorien gertutasunean oinarrituz sailkapena burutzen du.

Banaketako espazioari buruz hitz egitean, dimentsio altuko (n) hiperplano bati edo multzo bati buruz ari gara, sailkapen- zein erregresio-problemetan erabil daitekeena. Kategorien arteko banaketa onena eta kategoria desberdinen puntuen arteko distantzia edo marjina handiena uzten duen hiperplanoa edo hiperplano multzoa bilatzea da *SVM*ren helburu nagusia. Hiperplanotik gertuen dauden instantziei *support vector* (sostengu-bektore) deritze. Gutxienez, sostengu-bektore bat dago kategoria bakoitzeko.

Sarrerako lagineko instantziak espazioan irudikatu ahal izateko, n dimentsiotako bektoreak bailira definitzen dira, non bektoreko dimentsio bakoitzak atributu bat errepresentatzen duen. Instantzia guztiak espazioan irudikatuta, *SVM*k kategoria bateko puntuak alde batean eta beste kategoriakoak beste aldean utziko dituen n dimentsiotako hiperplanoa(k) bilatzen d(it)u. II.2 irudian bi kategoria desberdinetako puntuak banatzeko *SVM*k topatzen duen marjina handieneko hiperplanoa ikus daiteke.

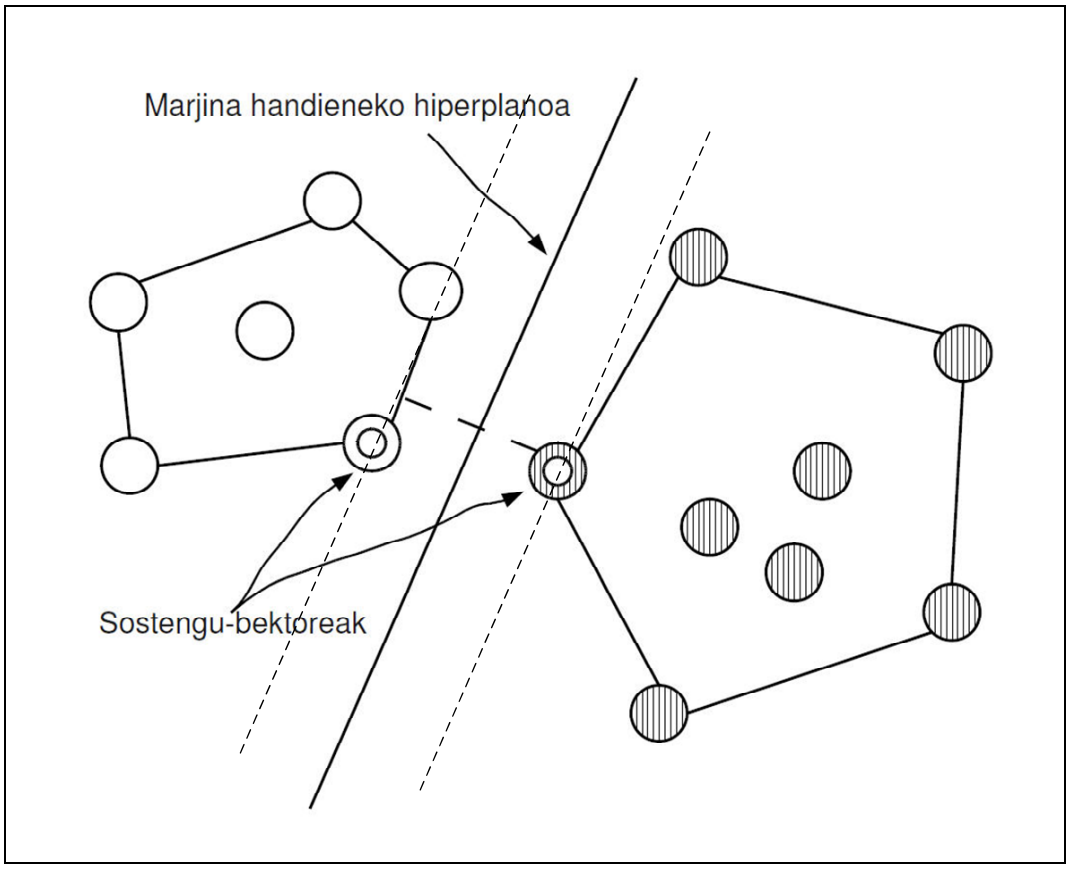
Problema ebazteko sailkatzaile lineala nahikoa ez denean, *kernel* izeneko funtzioa erabiltzen da. Funtzio horiek sarrerako atributuen espazioa egokitu eta espazio berrien planoen bilaketa burutzen dute, funtsean.

SVM algoritmoaren ikasketa-fasea entrenamenduko instantzia kopuru handiekin motela da, aldiz portaera bereziki ona dauka atributu askoko atazetan, baita linealki banagarriak ez diren problemetan ere. Azkenik esan, kategorien banaketarako erabaki-muga konplexuak eta finak eskuratzen dituenez, sortzen dituen sailkatzaile automatikoek emaitza onak eman ohi dituztela. Informazio zehatzagoa Daumèk (2004) egindako lanean aurki daiteke.

II.2.3.4 *AdaBoost*

Boosting algoritmoa, eta zehazkiago *AdaBoost* (*Adaptative Boosting*) algoritmoa, meta-sailkatzailea dela esaten da, jada inplementatuta dauden beste sailkatzaile[9] batzuen erabileran oinarritzen baita.

[9]Edozein algoritmok balio dezake, adibidez aurretik aipatu ditugun algoritmoekin eraikitako sailkatzaileak erabilgarri lirateke.



II.2 Irudia: *SVM* algoritmoaren margina handieneko hiperplano banaketa eta sostengu-bektoreak

*AdaBoost*ek, pisu bidezko gehiengoaren bozketaren eskema erabiliz, hainbat sailkatzaile edo erregela sinple konbinatzen ditu sailkatzaile sendo eta bakar bat sortzeko. Sailkatzaile horiek sekuentzialki ikasten dira pisudun adibideetatik abiatua. Pisu horiek dinamikoki egokitzen dira ikasketa-iterazio bakoitzean, aurretik ikasitako erregelen arabera.

Hortaz, *Boosting* prozesuak aurreko sailkatzaileek asmatzen ez dituzten adibideei garrantzi handiagoa emanez entrenatzen ditu sailkatzaileak, sailkatzaile bakoitzak indibidualki egiten dituen erroreak murrizten dituen meta-sailkatzaile bat lortzeko helburuarekin. Horretarako, adibide eta kategoria pisuen matrize bat erabiltzen du adibide bakoitza zein sailkatzailerekin ebatzi duen ikasteko (pisua maximizatzea helburu izanik). Asmatzeak saritzeko

eta erroreak penalizatzeko estrategiarekin eraikitzen den matrize horretan islatuta geratzen da sailkatzaile bakoitzaren portaera zein adibide multzorekin den ona edo txarra. Estrategia horri sailkatzailearen espezializazioa deritzo.

Kategoria ezezaguna duen adibide bat sailkatu nahi denean, *AdaBoost*ek entrenamenduan erabilitako sailkatzaile guztiak aplikatzen ditu eta bozketa haztatua egiten du ikasketan sortutako pisuen matrizean oinarrituz. Horrela, entrenamendu-fasean antzeko adibideak gaizki sailkatu dituzten sailkatzaileek 0 pisua izango dute eta portaera egokia aurkeztu dutenek pisu altuak. Pisu altu horiek lortzen dituzten sailkatzaileen antzeko adibideen erabakietan oinarrituz, bozka gehien lortzen dituen kategoria izango da algoritmoak adibide berriarentzat itzuliko duen sailkapena.

Ikasketa-algoritmo horien eta HPan gehien erabiltzen diren algoritmoei buruz gehiago jakin nahi izanez gero, Witten eta Frankek (2005) eta Manning eta Schutzek (2003) idatzitako liburuetan informazio zabala aurki daiteke.

AdaBoost teknika hau izendun entitateen alorrean emaitza onak lortuz erabili izan da (Carreras *et al.*, 2003). Lan horretan zehazki, erabaki-zuhaitz txikietan oinarritutako *AdaBoost* teknika azaltzen da. *Abionet* deitzen den inplementazio hori bera da, hain zuzen ere, gure esperimentuetarako aukeratu duguna.

II.2.4 Ebaluazio-neurriak

NERC problema, aurretik aipatu den bezala, bi ataza nagusitan banatu ohi da: identifikazio eta sailkapen atazetan, alegia. Ataza bakoitzaren portaera ondo aztertzeko asmoz, bukaeran NERC atazaren ebaluazio bateratua aurkeztuko den arren, metodo bakoitzaren identifikazioa eta sailkapenaren ebaluazio independenteak aurkeztuko dira. Horretarako IR arloko doitasun, estaldura eta F_1 ebaluazio-neurrietan oinarrituz (Baeza-Yates eta Ribeiro-Neto, 1999) oinarrituko gara.

NER atazan, BIO etiketatzean oinarrituz, neurketa desberdinak erabil daitezke:

- Estaldura, doitasuna eta F_1 BIO kategoria guztientzat.
- Estaldura, doitasuna eta F_1 B eta I kategorientzat soilik, azken finean horiek baitira entitate-hautagaiak mugatzen dituzten etiketak.
- CoNLL ataza partekatuko estaldura, doitasuna eta F_1 neurria (identifikatutako entitateena soilik), NE osoa izanik ebaluazio unitatea eta ez

B, I eta O etiketak.

Hiru aukera horietatik, metodoen portaera bi atazetan banatuta eta beste hizkuntzetan egindako ikerketekin konparaketak burutu ahal izateko, azken neurketa metodoa aukeratu da ebaluazioa aurkezteko. Horrela izanda, sailkatzaile automatikoak osagai bat zuzen etiketatu duela esango da, eskuz etiketatutako osagaiaren kategoriarekin bat datorrenean. NEC atazan, kategoria automatikoa eta eskuz esleitutakoa bat datozenean gertatuko da. NER atazan aldiz, NEa ondo etiketatu dela kontsideratuko da, bera osatzen duten osagai kateko osagai guztien eskuzko eta automatikoki esleitutako kategoriak bat datozenean. Nahikoa izango da osagai bateko desadostasuna NE osoa gaizki mugatu dela kontsideratzeko. Esaterako II.2.4 adibideko esaldian, *Bilboa Bizkaia* eta *Ur Patzuergoak* hitz kateak NE bakar bezala identifikatu ordez sailkatzaileak automatikoki bi NEtan banatu dituenez, *Ur* osagaiari *I* kategoria esleitu beharrean B kategoria esleitu dionez, sistemaren hutsegitea kontsideratuko da gainerako elementuak ondo etiketatuta dauden arren.

II.2.4 Adibidea Automatikoki etiketatutako ebaluaziorako adibidea.

Eusko_Jaurlaritzak	*B*
,	*O*
Nafarroako	*B*
Gobernuak	*I*
,	*O*
Iberdrolak	*B*
,	*O*
Bilbao	*B*
Bizkaia	*I*
Ur	***B***
Partzuergoak	*I*
eta	*O*
beste	*O*
hainbat	*O*
erakunde	*O*
eta	*O*
enpresak O	
.	*O*

Doitasunaren bidez sailkatzaile automatikoak identifikatutako osagaien zuzentasuna neurtzen da; estalduraren bidez, berriz, identifikatu/sailkatu behar diren guztietatik zenbat asmatu diren adierazten da. Bi neurri hauen arteko erlazio matematiko zuzenik ez dagoen arren, ikerketa enpirikoetan ikusi izan

da alderantziz erlazionatuta daudela: sistemak detektatutako elementuen kopurua handitzean (estaldura handitzen bada), doitasuna murrizten da, eta alderantziz. Bi neurri hauek kontuan hartzen dituen metrika da hain zuzen ere $F_{\mathcal{B}}$.

Emaitzak, beraz, neurri hauen arabera aurkeztuko dira:

- $doitasuna = \frac{sailkatzaile\ automatikoak\ zuzen\ etiketatutako\ NE}{sailkatzaile\ automatikoak\ etiketatutako\ NE\ guztiak}$
- $estaldura = \frac{sailkatzaile\ automatikoak\ zuzen\ etiketatutako\ NE}{eskuz\ etiketatutako\ NE\ kopurua}$
- $F_{\mathcal{B}} = \frac{(\mathcal{B}^2+1)*Doitasuna*Estaldura}{(\mathcal{B}^2*Doitasuna+Estaldura)}$

Normalean, $\mathcal{B} = 1$ erabiltzen da, doitasunari eta estaldurari pisu bera emanez:

$$F_1 = \frac{2 * Doitasuna * Estaldura}{(Doitasuna + Estaldura)}$$

NEC atazan, sistemak edozein NEerako kategoria bat eta bakarra itzuliko duenez, doitasun eta estaldura berdinak dira, eta ondorioz F_1 neurria ere.

II.3 Euskarazko NERC sistemaren diseinua

Euskarazko entitate-izenen erauzketarako sistemaren helburu nagusia euskarazko testuetatik pertsona-izenak, tokiak, erakundeak eta hiru sailkapen horietan sartzen ez diren espresioak erauztea da. Izendun entitateak ere badiren denborazko zein zenbakizko espresioak aurreprozesuan aplikatzen den *Eustagger* izeneko euskarazko lematizatzaile/etiketatzaileak ebazten ditu lehendik. Hori dela eta, sistemak ez du espresio mota horiek erauzteko helburua izango, soilik entitate-izenenaz arduratuko da.

Bibliografian ikusi den bezala, IAko tekniken artean, metodo gainbegiratuek sendotasuna eta emaitza onak lortzeko aukera eskaintzen dute. Metodo horien baldintzetako bat corpus etiketatuaren beharra da. Hizkuntza batzuetan corpus etiketatuak biltzea ez da hain lan zaila izaten, euskara bezalako hizkuntzetan aldiz, ia ezinezkoa da corpus etiketatuak lortzea. Bide bakarra, hortaz, baliabideak norberak sortzea da. Horregatik euskarazko NERC sistemaren diseinua burutzeko garaian bi bide aurreikusi ziren:

- Euskarazko corpus bateko entitate-izenak hutsetik eskuz etiketatu eta IAko teknikak aplikatuz euskarazko NERC tresna eraiki.

- Euskarazko corpus bateko entitate-izenak erdi-automatikoki etiketatzeko ezaugarri linguistikoetan oinarritutako NERC tresna eraiki eta, ondoren, horren irteerako NEak instantzia gisa erabiliz, eskuzko zuzenketa/berrikuspenaren ondoren, IAko tekniketan oinarritutako euskarazko NERC tresna eraiki.

Euskarazko corpusa hutsetik eskuz etiketatzeko baliabide nahikorik ez izanik, eta kostu altuko lana zela ikusirik, bigarren aukera hautatu zen lana aurrera eramateko. Lehen urratsa, hortaz, ezaugarri linguistikoetan oinarritutako euskarazko NERC sistema da. Hizkuntza-ezagutzan oinarritutako tresna horren garapenarekin, entitate-izenen etiketatze erdi-automatikoa ahalbidetzeaz gain, estrategia hori jarraitzera bultzatutako bestelako abantailak ere badaude, honako hauek batez ere:

- Ezaugarri linguistikoetan oinarritutako tresnaren garapenak NEen alorrean baliagarriak diren hizkuntzaren ezaugarriak eta portaerak ondo ezagutzeko aukera luzatzen du, aurrerago IAko tekniketan atributu egokiak definitzen lagungarri izango dena.

- Hizkuntzari hertsiki lotutako estrategiaren eta bibliografian ezagunagoak diren IAko tekniken arteko konbinazioa aztertzeko aukera emango du.

Euskarazko entitate-izenak erauzteko sistemaren garapena, hortaz, lau urratsetan banatuta dago:

1. Ezaugarri linguistikoetan oinarritutako tresnaren garapena.

2. Euskarazko entitate-izenak etiketatuta dituen corpusa eraiki, lehenengo urratseko tresna aplikatuz eta jarraian eskuzko berrikuspena burutu.

3. Bigarren urratsean sortutako corpusarekin IAko metodo desberdinak aplikatu, guztietatik euskararako portaera onena duen sailkatzailea aukeratzeko.

4. IAko metodo desberdinetatik lortutako ereduen emaitzak konbinatu tresna sendoagoa lortzeko.

II.3.1 *Eihera*, ezaugarri linguistikoetan oinarritutako tresna

Eihera, euskarazko entitate-izenak erauzteko ezaugarri linguistikoetan oinarritutako tresna da. Tresna honek NEen erauzketa bi ataza nagusi eta sekuentzialetan burutzen du: identifikazioa eta sailkapena, hurrenez hurren.

Identifikaziorako erregelez osatutako gramatika bat erabiltzen du. Erregela horiek testuen aurreprozesuan *Eustagger*en bitartez lortutako informazio morfosintaktikoan oinarritzen dira.

Sailkapenerako, berriz, barne- zein kanpo-ebidentziak baliatzen dituen heuristikoa aplikatzen du. Heuristiko horrek ere informazio morfosintaktikoa baliatzen du sailkapen posibleen artean egokiena hautatzeko.

Ondorengo ataletan bi ataza horien deskribapen zehatza aurki daiteke.

II.3.1.1 Identifikazioa / NER

*Eihera*k oro har informazio linguistikoa erabiltzen du NEen erauzketarako, bereziki identifikazio atazan. Tesi-lan honetan informazio linguistikoari buruz hitz egitean, gehienetan, informazio morfosintaktikoari buruz ari gara. Informazio horretan oinarrituta, hain zuzen ere, euskarazko NEek testuetan jarraitu ohi dituzten patroi morfosintaktikoak definitu dira. Patroien definizio-prozesuan bi hizkuntzalariren parte-hartze zuzena egon da.

Lehen urratsa, gramatikan parte hartzen duten osagai nagusiak identifikatzea eta horien ezaugarriak zehaztea izan da. NER atazako gramatikak bi informazio mota nagusi dituela esan daiteke: NEak berak eta hitz eragileak.

Euskarazko NEen osagai gehienak, ingelesa eta gaztelera hizkuntzetan gertatzen den bezala, letra larriz idatzita egon ohi dira. Hori dela eta, letra larriak dira NE osoaren ezaugarri nagusiena eta intuitiboena. Baina ez da kontuan hartu beharreko ezaugarri bakarra. *Eihera*ren identifikazio gramatikak NEaren osagaien kategoria, azpikategoria eta deklinabide-kasuak ere erabiltzen ditu NEen espresioak testuetan mugatzeko garaian.

NEen osagaien kategoria/azpikategoria guztien artean patroiak definitzeko erabiltzen direnak, ondoren izendatzen diren etiketekin identifika daitezke: IZE_ARR (izen arruntak), IZE_IZB (pertsona-izen bereziak), IZE_LIB (leku-izen bereziak), SIG (akronimoak) eta BST (partikula bereziak[10]).

NER gramatika osatzen duten patroiak deskribatzean ikusiko den bezala, NEen osagai guztiak kategoria morfosintaktiko horietako bati egokituko zaiz-

[10]BST etiketa gazteleraz idatzitako partikula batzuei egokitzen zaie, hala nola, preposizio zein determinatzaileei (*de*, *la*, etab.). Adib: Santiago *de* Compostela.

kio, kontrako kasuan, gramatikak ez baititu entitate osagaitzat joko. Osagai horien letra larriei erreparatuz, BST etiketako elementuak ezik, gainerako osagaiek letra larriz hasi beharko dute NEaren osagai kontsidera daitezen, BST horiek ere letra larriz idatzita ager daitezkeelarik.

Identifikazio-patroien definiziorako osagaien deklinabide-kasuaren araberako bereizketa ere egin da: deklinatu gabeko elementuak batetik; genitibo kasuan deklinatuak bestetik; eta azkenik genitiboa ez den beste edozein deklinabide-kasua dutenak bereiziz.

Gramatikan NEak mugatzeko bi patroi nagusi aurki daitezke (ikus A eranskinean gramatika osoa): osagai bakarreko entitateak (*Europan*) mugatzeko patroia eta osagai bat baino gehiagoko entitateak (*Europako Banku Zentralean*) identifikatzeko patroia. Lehenengo motako patroia betetzen duten entitateek, IZE_IZB, IZE_LIB edo SIG kategoria/azpikategoriakoak izan beharko dute, eta deklinatuta dauden kasuetan, edozein deklinabide-kasu onartuko dute.

Osagai bat baino gehiagoz osatutako NEek, bigarren motako patroia betetzen dutenek alegia, egitura konplexuagoa dute. Hasteko, entitatearen azken osagaia deklinabidearen ikuspuntutik gainerako osagaien portaera desberdina du, ez baitu inolako deklinabide murriztapenik bete behar. Gainerako osagaiek, berriz, genitibo kasua besterik ez dute onartzen. *Eustagger*ek esleitutako kategoria/azpikategoriari dagokionez, aurreko patroian aipatutakoez gain, IZE_ARR, BST eta ADJ[11] motakoak ere onartzen dira bigarren patroi honetan.

Bigarren patroia betetzen duten bi entitate ikus daitezke II.3.1 adibidean. Lehenengoak hasierako osagaia genitiboan du, erdiko izenak ez du deklinabide-kasurik eta azken adjektiboak edozein kasu onartzen duenez (adibidean inesiboa) patroia betetzen du. Bigarren NEean, berriz, lehenengo osagaia deklinatu gabekoa da, erdian letra xehez idatzitako BSTrekin etiketatutako bi hitz daude eta azken osagaia ergatiboan deklinatuta dago, azken horretan deklinabide-kasuan murriztapenik ez dagoenez, patroia betetzen du.

II.3.1 Adibidea NER gramatikako 2. patroia betetzen dituzten adibideak.

- *Europako* (IZE_LIB+GEN[12] *Banku* (IZE_ARR) *Zentralean* (ADJ+INE)
- *Alex* (IZE_IZB) *de* (BST) *la* (BST) *Iglesiak* (IZE_IZB+ERG)

Gramatika aplikatzean patroiekin bat egiten duten osagai sekuentziarik luzeenak mugatuko dira NE gisa.

[11] Adjektiboa

[12] ikus B eranskinean deklinabide etiketen esanahia

NEak mugatzeaz gain, hitz eragileak ere identifikatzeko eta mugatzeko gramatikako erregelak definitu dira. Hitz eragileak, NEen osagaiak ez bezala, letra xehez idatzita egon daitezke eta zehazki NEaren aurreko zein atzeko posizioan ager daitezke. Gramatikak osagai bat aurreko zein ondorengo hitz eragiletzat jotzeko, izen (IZE) kategoriakoa beharko du izan, kategoria horretako hitzak NEak kontestualizatzeko informazio gehien eskaintzen duten unitateak baitira. NEaren aurretik agertzen diren hitzak hitz eragile kontsideratzeko, kategoriaren murriztapenaz gain, deklinabidea ere murriztua dute: deklinatu gabeko hitzak dira edo genitibo kasuan deklinatuta daude. Beste deklinabide-kasuekin deklinatutako hitzak baztertzen dira, azken hauen izaera dela eta, semantikoki NEarekin zerikusirik ez dutela kontsideratzen baita. Esaterako *Jokalariekin Anoetara joan zen* esaldian, *jokalari* hitza ez da *Anoeta* NEaren hitz eragile kontsideratzen deklinabideak NE horrekin loturarik ez du definitzen eta. Aldiz, *Markel Susaeta jokalaria zelaira sartu zen* esaldian, *jokalari* hitzak *Markel Susaeta* pertsona-izenaren deskribatzaile kontsidera daiteke, eta hortaz gramatikak hitz eragile kontsideratuko du.

*Eihera*ren identifikazio-modulua, hortaz, patroi morfosintaktikoak biltzen dituen gramatika batek osatzen du. Gramatika hori garatzeko *XFST* tresna erabili da. Tresna horrek NEen egitura definitzeko patroiak zehazteko aukera eskaintzeaz gain, patroi horiek mugatzeko edo identifikatzeko erregelak formalizatzeko lana errazten du.

Esaterako, osagai bakarreko NEak mota desberdineko osagaiekin eratuta egon daitezke. *XFST* tresnak guztiak patroi bakarrean biltzeko aukera eskaitzen du:

```
define PAT1 [TokenSIGLA | TokenIZENB | TokenIZENLB |
             TokenMAIUSK | TokenIZEARR ];
```

PAT1 aldagaiak beraz, sigla, izen berezi, toki-izen berezi edo izen arrunt motako hitzak definitzen ditu. Patroi hori erabiliz, kakoen artean mota horretako edozein NEren agerpena mugatzea nahiko bagenu, *XFST*ko ondoko erregela definituko genuke:

```
define MugaMuga [PAT1 @-> "[" ... "]"];
```

A eranskinean *XFST* sintaxia jarraituz, euskarazko NER ataza ebazteko gramatika osatzen duten patroi eta erregelak aurki daitezke. Guztira NEentzat eta hitz eragileentzat zazpi definizio-patroi eta mugatzeko hiru erregela nagusi idatzi dira.

Zaila da denboraren estimazio bat egitea, baina tresnak corpusa erdiautomatikoki etiketatzeko eskaintzen duen aukeragatik bakarrik tresnaren garapena merezi duela uste dugu. Gainera, finkatutako helburuetako bat betetzen lagundu du, garapen azkarra izateaz gain, aurrerago ikusiko dugun bezala, sistema hibridoa sortzeko bidean oso erabilgarri izan baita. Zer esanik ez, euskarazko NER ataza ebazteko atributu esanguratsuen identifikazioan oso lagungarri izan da, IAko metodoak aplikatzean berrerabili den informazioa, identifikatzen, hain zuzen ere.

II.3.1.2 Sailkapena / NEC

*Eihera*ren sailkapen-modulua, kanpo-ebidentzia zein informazio morfosintaktikoa konbinatzen dituen heuristiko batek definitzen du. Kanpo-iturriak hitz eragileen zerrendek eta *EusWN* aberasketarik gabeko *gazetteer*ek osatzen dituzte (ikus II.2.1 atala). Bi kasuetan sailkapen kategoria bakoitzeko zerrenda bat erabiltzen da. Sailkapen kategoria posibleak ondokoak dira:

- Pertsona-izena (PER)
- Toki-izena (LOC)
- Erakunde-izena (ORG)
- Bestelakoak: film izenburuak, liburu izenburuak eta antzeko espresioak (OTH)

Entitate bakoitzeko *Eihera*k jarraitzen duen heuristikoaren urrats sekuentzialak deskribatzen dira jarraian:

1. NEa *gazetteer* zerrendaren batean agertzen den begiratzen da, eta hala denean, *gazetteer* horri dagokion kategoria itzuliz bukatzen da sailkapen-prozesua. Kontrako kasuan, 2. urratsera jotzen du.

2. NEaren elementuak banan-banan eskuinetik ezkerrera eskuratu eta dagokien deklinabide-kasua eta etiketa morfosintaktikoa[13] aztertzen dira. Orokorrean, bi iturri horien balioen arabera sailkapen kategoriei pisuak esleitzen zaizkie. Genitibo kasuan deklinatutako osagairen bat egotekotan, tratamendu berezia aplikatzen da: genitibo kasuko osagaitik abiatuta ezkerrerago dauden osagaiak, bera barne, ez dira pisuak

[13]Etiketa morfosintaktikoa esatean kategoria/azpikategoria informazioz osatutako etiketa esan nahi da.

esleitzeko erabiliko. Adibide batekin argitzen da urrats hori jarraian. Suposatu sarrerako entitatea *Europako Banku Zentralean* dela. Heuristikoak *Banku* eta *Zentralean* osagaien informazioa erabiliko du pisuak estimatzeko, *Europako* osagaia baztertuz ezkerrean eta genitibo kasuan deklinatuta baitago. Hori horrela egiten da euskarak burua bukaeran duen hizkuntza baita. Heuristiko honi esker NE konplexuen sailkapena hobetzen da.

Kategoria/Azpikategoria | Deklinabidea	**PER**	**LOC**	**ORG**
Izen arrunta (IZE_ARR)	-	-	+1
Leku-izen berezia (IZE_LIB)	-	+2	+1
Izen berezia (IZE_IZB)	+1	-	+1
Sigla (SIG)	-	-	+2
Ergatiboan deklinatuta (-k)	+2	-	+2
Inesiboan deklinatuta (-n)	-	+2	+2
Leku-denborazko deklinabide-kasuak (-ra,-tik,-raino, etab.)	-	+2	+2

II.3 Taula: Kategoria/azpikategoria eta deklinabide-kasuen araberako pisuak.

Osagaien kategoria/azpikategoria eta deklinabide-kasuen araberako sailkapen bakoitzeko pisuen definiziorako, adituen laguntzarekin definitutako eta erabilitako pisuen balioak II.3 taulan erakusten dira. Bertan ikus daitekeen bezala, ORG motako NEak, mota guztietako osagaiak onartzen dituztenez, edozein dela kategoria/azpikategoria zein deklinabide-kasua sailkapen horren pisua beti handitzea erabaki da. Pertsona-izenen kasuan aldiz, IZE_IZB motako osagaiak eta ergatibo deklinabide marka duten osagaiek soilik handitzen dute kategoriaren pisua, eta toki-izenen kasuan berriz, IZE_LIB eta inesibo zein leku-denborazko deklinabide-kasuko osagaiek.

3. NEarekin batera hitz eragileren bat identifikatu bada, hitz eragileen zerrendak kontsultatuz zein entitate kategoriaren inguruan agertu ohi den aztertzen da, eta kategoria horren pisua bi puntutan handitzen da, azken erabakian informazio horrek eragina izan dezan. NEaren ondorengo hitz eragilea erabiliko da, baldin eta soilik baldin, NEaren azken osagaia genitibo kasuan deklinatuta ez badago; hala denean, hitz hori ez da haztaketarako kontuan izango. Esaterako, *Markel Susaeta jokalariak* espresioan *jokalariak* hitz eragilea haztaketan erabiliko da.

Ez ordea *Eibarko alkatea* espresioko *alkatea* hitz eragilea, *Eibar* NEa genitibo kasuan deklinatua baitago.

4. Aurreko pausuetan kategoriei esleitutako pisuak aztertzen dira azken urrats honetan. Baliorik altuena lortu duen kategoria da, hain zuzen ere, itzultzen dena. Berdinketa gertatzean, sailkapen kategorien artean II.4 taulan agertzen diren ebazpen irizpideak aplikatzen dira[14].

Berdinketako kategoriak	**Kategoria hautatua**
ORG eta LOC	LOC
PER eta LOC	LOC
PER eta ORG	PER
PER, ORG eta LOC	PER

II.4 Taula: Kategorien pisuen arteko berdinketak ebazteko irizpideak.

II.3.2 Ikasketa automatikoan oinarritutako tresna

Hizkuntza-ezagutzan oinarritutako tresnaren garapenerako egin den bezala, IAko metodoetan oinarritutako tresnak garatzeko garaian NERC problema NER eta NEC atazetan banatu dugu. Jarraian, ataza bakoitza ebazteko egindako lanak deskribatzen dira.

II.3.2.1 Identifikazioa / NER

IA aplikatzean algoritmo egokia aukeratzea bezain garrantzitsua da modelizatu nahi den problemaren informazioa adierazteko erabiliko diren atributuen definizioa eta aukeraketa. NER atazarako testuen errepresentaziorako zein atributu erabili aztertzen hastean, ez gara hutsetik abiatu. *Eihera* tresna garatzean antzemandako portaerak zein identifikaziorako gramatika garatzeko erabili diren ezaugarri morfosintaktikoak erreferentzia garrantzitsuak izan dira.

Informazio hori abiapuntu gisa harturik, euskarazko testuetako hitzak deskribatzeko oinarrizko atributuak definitu dira, unitatea hitza izanik. Testuko hitz bakoitza, hortaz, ondoko atributuen bidez adierazten da:

[14] Irizpideak erregela guztiak definitzeko erabilitako adibide multzoaren azterketatik ondorioztatu dira.

- hitzaren forma
- hitzaren lema
- kategoria/azpikategoria morfosintaktikoa
- deklinabide-kasua
- hitza letra larriz hasten da (bai/ez)
- lema letra larriz hasten da (bai/ez)
- hitza osorik letra larriz (bai/ez)
- lema osorik letra larriz (bai/ez)

2003 urteko CoNLL edizioan parte hartu zuten hainbat sistemek (Tjong Kim Sang eta De Meulder, 2003) kanpo-ebidentzia erabiltzen zuten hitz/lemen inguruko letra larriaren informazioa erauzteko. Gure sisteman ordea, informazio hori *Eustagger*ek eskaintzen duen lemaren informaziotik erauztea erabaki da, letra larrian informazio hori lemaren baitan islatzen baita. Hala, kanpo-ezagutzara jotzeko beharra ekiditen da eta beste hainbat pausotarako beharrezkoa den *Eustagger*en informazioa berrerabiltzen da.

Saio desberdinak egin dira hitz bakoitza adierazteko leiho tamainarik egokiena [-3,+3] dela erabakitzeko. Leiho tamaina hori izango da, hain zuzen ere, tesi-lan honetan aurkeztuko diren NEReko esperimentu guztietan erabiliko dena. Hitz bakoitzeko beraz, 56 atributu erabili dira: leihoko 7 hitzetako bakoitzeko 8 atributu.

Ezaugarrien edo atributuen aukeraketak emaitzak hobe ditzake, dokumentu-sailkapeneko Arregi eta Fernández ek (2002) egindako lanean erakusten den bezala. Dokumentu-sailkapenean bezala, formak edota lemak aukeratuta entitateen arloan esperimentatzea interesgarria izan daitekeela kontsideratuz, hitz edo lema atributu-aukeraketa hori aurrera eraman da IA-ko NER modulua hobetzeko asmoz. Azken batean, lema batek forma askoren informazioa bil dezake (bereziki euskara bezalako hizkuntza eranskarietan), baina bestalde bestelako informazio aberatsa galtzen du (deklinabide-kasua, mugatasuna eta numeroa izenetan, denbora marka aditzen kasuan, etab.). Emaitzen atalean (ikus II.4 atala) forma/lema atributu-aukeraketaren portaera aztertuko da.

Atributuen aukeraketaz gain, sailkatzaileak hobetzeko adibideen hautaketa beste bide posible bat da. Brighton eta Mellishen (2002) lanean aurkezten den bezala, sarrerako instantzien errepresentazioan osagai zaratatsuak

egoteak sailkatzailearen erantzun-denborak handitzeaz gain, sailkatzailearen asmatze-tasan ere eragin negatiboa du. Bestetik, EACLn aurkeztutako esperimentuetan Mikheev *et al.* (1999) egileek argi erakusten dute sailkapen-klaseen ikasketarako erabilitako instantzia kopuru ez-orekatua kaltegarria dela, batez ere *Naïve Bayes* bezalako algoritmoetan.

Ikasketa-corpusa aztertzean, oso maiz agertzen diren eta NER atazarako hain interesgarriak ez diren O motako osagaiek corpuseko kategorien arteko desoreka eragin dezaketela kontsideratu da. Izan ere, testuetan NEen osagai diren baino NEen osagai ez diren hitz gehiago egon ohi baitira. NERen helburua NEak (B eta I motako osagaiak) identifikatzea izanik, ebaluazioan ongi mugatutako NEak neurtzen dira, B I etiketa-kate zuzenen kopurua, alegia, eta O osagaiak ebaluaziotik kanpo geratzen dira. Desoreka hori ebazteko asmoz, NEen osagaiak ez diren hitzak ezabatzeko heuristiko sinple bat diseinatu da. Oinarrian heuristiko horrek, izenak, juntagailuak eta BST motako elementuak salbu, letra larriz hasten ez diren hitz guztiak ezabatzen ditu.

Testutik IAko algoritmoen atributuen errepresentaziora bihurtzeko prozesuak bi momentu desberdinetan aplika dezake instantzia aukeraketa horietarako heuristikoa:

- lehenengo hitz baztergarriak ezabatu eta ondoren atributu erauzketarako leihoa aplikatu.
- lehenengo atributuak erauzteko leihoa aplikatu eta ondoren ezabaketa burutu.

Zalantzarik gabe bigarren aukera da egokiena inguruko hitzen semantika mantentzen baitu. Azken batean, leihoa aplikatu aurretik hitzak ezabatzeak jatorrizko testuan testuingurukoak diren hitzak desagerraraztea zein testuingurukoak ez diren osagaiak testuingurutzat hartzea eragin baitezake, trebatzean zarata sortuz.

Heuristiko sinple hori aplikatzean O motako osagaien ezabaketarekin ikasketa-corpusaren tamaina erdira jaisten dela ikusi dugu, NE kopuru totalaren % 0,4 soilik galduz (ikus II.5 taula). Galera hori bereziki NEen parte diren, izenak, juntagailuak eta BST motako elementuak ez diren eta letra xehez idatzitako osagaietan du jatorria. Esaterako, irizpide horren arabera Joan Mari Irigoienen *Lur bat haratago* liburuaren izenburuko azken bi osagaiak ezabatuko lirateke corpuseko NE osoaren galera eraginez. Ezabaketa-heuristikoa ebaluazio-corpusean aplikatzean murrizketari dagokionez tamainan proportzio bera mantendu da. Ikasketa- eta ebaluazio-corpusaren ta-

maina zehatzak instantzia-aukeraketarekin (murriztua) eta aukeraketa gabe (osoa) II.5 taulan laburbiltzen dira.

	osoa	**murriztua**
Ikasketa-corpusa	46.227	23.054
Test-corpusa	15.960	8.022

II.5 Taula: Corpusen hitz kopuruak instantzia-aukeraketa aplikatuta (murriztua) eta gabe (osoa)

Emaitzen atalean (ikus II.4.1.1 atala) murrizketa horrek emaitzetan izandako eragina azalduko da, sailkatzaileen ikasketa- eta erantzun-denborekin batera.

NER sistema horien portaera hobetzeko erabilitako azken estrategia metodoen konbinaketa izan da. Aurretik aipatutako esperimentu guztiak erabiliz, konbinazio mota desberdinak aplikatu dira: metodo bera atributu eta ikasketa-corpus desberdinekin[15]; eta algoritmo desberdinekin ikasitako sailkatzaileen arteko konbinazioak burutu dira. Emaitza horiek II.4.1.2 atalean azaltzen dira.

II.3.2.2 Sailkapena / NEC

Euskarazko NEen sailkapenerako ikasketa automatikoan oinarritutako sistema garatzeko, ingurune esperimentaleko metodoen azpi-atalean (ikus II.2.3 atala) deskribatutako algoritmo guztiekin esperimentuak egin dira. Hala ere, atributu eta instantzien aukeraketa zein metodo konbinazioetarako algoritmo guztiak erabili beharrean, emaitza onena lortu duten algoritmoak hautatu dira, NER atazan ere estrategia horrek ondo funtzionatzen duela ikusi baita.

Sarrerako adibideen deskribapenerako atributuak definitzeko, *Eihera*ren sailkapen moduluko heuristikoa hartu da oinarri gisa. Hala, NER atazarekin bat datozen hainbat atributu erabiltzea erabaki da, osagaien barne-ebidentziak edo ñabardurak islatzen dituzten atributuak erabiliz, hain zuzen ere:

- hitzaren forma
- hitzaren lema

[15] corpus osoa edo instantzia aukeratuekin

- kategoria/azpikategoria morfosintaktikoa
- deklinabide-kasua

Ataza honetan helburua identifikatutako NEen sailkapen kategoria den heinean, ikasketarako oinarrizko unitateak testuko hitzak beharrean NEak izango dira.

Hortaz, sarrerako instantziak, NEen osagai guztien atributuekin adierazi-ko dira. NER atazan egin den bezala, testuinguruko informazioa ez galtzeko, ikasketa leihoa finkatu da. Leiho horretako unitateek ez dute zertan NEak izan, azken batean testuinguruko informazioa NEak ez diren osagaiek ere gordetzen baitute. Leihoaren tamaina mugatzeko saiakera desberdinak egin ostean, emaitza onenak ematen dituen [-2,+2] leihoa aukeratu da sailkatzai-leak trebatzeko.

NERen ez bezala, ataza honetan *Eihera*ren sailkapen heuristikoan ustia-tutako kanpo-iturriak ere erabiltzea erabaki da, *Eihera*n erakutsitako eragin positiboan oinarrituz. *Gazetteer* eta hitz eragileen zerrenden agerpenak isla-tzen dituzten atributuak ere definitu dira, bat zerrenda mota eta kategoria bakoitzeko. Esperimentu desberdinak egin dira, *EusWN*ekin aberastutako zerrendak erabilita eta erabili gabe ere.

Aurretik aipatu den bezala, batzuetan nahikoa da atributuen aukeraketa ona egitea sistemaren emaitzak hobetzeko. Bide horretan, NEC atazarako sailkatzaileak entrenatzeko garaian, kanpo-informazioaren atributuekin saia-kera desberdinak egin dira beraien eragina ere neurtzeko asmoz.

Metodo konbinaketei dagokienez, NER atazan bezalaxe esperimentu des-berdinak egin dira hiruko konbinazioak eginez:

- Metodo bera hitz eragile, *gazetteer* edota *EusWN* informazio-iturrien konbinazio desberdinak erabiliz eraikitako sailkatzaileak konbinatu.
- Metodo desberdinak hitz eragile, *gazetteer* edota *EusWN* informazio-iturrien konbinazio antzekoekin ikasitako sailkatzaileak konbinatu.
- Metodo desberdinak hitz eragile, *gazetteer* edota *EusWN* informazio-iturrien konbinazio desberdinekin trebatutako ereduak konbinatu.

II.4 Emaitzak

Atal honetan euskarazko testuetan NEak identifikatu eta sailkatzeko egindako esperimentuen emaitzak aurkezten dira. Ebaluaziorako, baliabideen atalean (ikus II.2.2 atala) definitutako 931 entitate dituen ebaluazio-corpusa erabili da. Bertan testuetan agertzen diren pertsona- (PER), toki- (LOC), erakunde-izenak (ORG) eta hiru sailkapen horien barne ez dauden film-izen eta bestelako entitate-izenak multzokatzen dituen bestelakoak (OTH) kategoriako NEak identifikatu (NER) eta sailkatzeko (NEC) sistemek duten ahalmena neurtu da.

NER atazan BIO etiketatze sisteman oinarrituz (ikusi II.2.2 atala) *Eihera*k zein IAko metodoek identifikatutako NEen zuzentasuna neurtzen da, NE bat ondo identifikatu dela kontsideratzeko bere osagai guztien BI etiketek ondo esleituta egon beharko dute. Nahikoa izango da osagai bat gaizki etiketatuta egotea NE horrentzat sistemak kale egin duela kontsideratzeko.

Bestalde, NEC atazan NEen osagaiak baino NE osoak dira sailkatu beharreko unitateak. Sistemaren asmatzea NEari dagokion PER, LOC, ORG edo OTH kategoria egokia esleitzean emango da, eta hutsegitea kontrako kasuetan.

NEC ataza ebaluatzean, egindako esperimentuen asmatze-tasak ondo aztertu ahal izateko, ongi mugatutako NEetatik abiatuko da sistema, alegia NEC ebaluazioan NER atazako emaitza automatikoak erabili beharrean eskuz mugatutako NEak sailkatuko dituzte sistemek.

Azkenik, kate osoaren ebaluazioa NERC azpi-atalean azalduko dugu, orduan bai sistema NER moduluko irteeran oinarrituko da NEC urratsa aplikatzeko, lehenengoan sortutako erroreak hedatuz. Sistemen asmatze-tasak okerragoak izango dira, normala den bezala, baina ebaluazio horrek sistemaren portaera erreala islatuko du, azken batean NERC sistema aplikazio batean erabiltzean, bi atazak sekuentzialki aplikatuko baitira. Eszenatoki erreal baterako, beraz, sistemen asmatze-tasak NERC azpi-atalean deskribatuko direnak izango dira.

Laburbilduz, emaitzen atala hiru puntu nagusitan banatu dugu: lehenbizi NER atazako esperimentuen ebaluazioak aurkezten dira, jarraian NEC atazarenak eta, azkenik, bi atazak sekuentzialki aplikatzean lortutako emaitzak, NERC sistema osoaren emaitzak, alegia.

II.4.1 NER

NER ataza ebazteko ezaugarri linguistikoetan oinarritutako tresna eta IA-ko metodoetan oinarritutako hurbilpenen emaitzak aurkezten dira jarraian. Lehenbizi metodo bakoitzak lortzen dituen banakako emaitzak aurkezten dira eta ondoren, emaitzak hobetzeko asmoz, sistema hauek konbinatuz sortutako sistema berrien emaitzak aztertuko ditugu.

II.4.1.1 Sistemen ebaluazioa banaka

II.4.1.1.1 Eihera

*Eihera*k, II.3.1.1 atalean deskribatu den bezala, NEak identifikatzeko informazio morfosintaktikoan oinarraitutako gramatika bat aplikatzen du. Gramatikak, NEak identifikatzeaz gain, espresio horien inguruan agertzen diren hitz eragileak mugatzen baditu ere, NER ebaluazioan azken horien identifikazioaren egokitasuna ez da neurtuko, hori ez baita sistemaren helburua. Hitz eragileak identifikazioan eta bereziki NEC atazan lagungarri gerta daitezkeelako identifikatzen dira. Hortaz, *Eihera*ren NER moduluaren ebaluaziorako, 931 NEko ebaluazio-corpusean aplikatzean, horietatik zenbat mugatzen dituen ondo neurtu da. Konkretuki, hizkuntza-ezagutzan oinarritutako tresnaren NER moduluak 951 entitate identifikatzen ditu, horietako 805 izanik zuzenak. Emaitzak neurrien atalean (ikus II.2.4 atala) deskribatutako estaldura, doitasun eta F_1 neurrien arabera ikus daitezke II.6 taulan.

doitasuna	**estaldura**	**F_1**
% 84,65	% 86,10	% 85,37

II.6 Taula: *Eihera*ren NER emaitzak

Erroreen jatorria aztertu nahian, ebaluazio-corpuseko 100 NE erroredun eskuz landu dira. II.7 taulan eskuzko azterketa horretan identifikatutako errore iturri nagusiak laburbiltzen dira. *Eihera*k identifikazio atazan (ikus II.3.1.1 atala) NEak mugatzeko hitzak letra larriz idatzita dauden adierazten duen atributua garrantzitsuenetakoa da.

Testuetan letra larrien erabilera okerrak, osagaia maiuskulaz idatzita dagoenean minuskulaz egon beharrean, alegia, *Eihera*k hitz hori NEaren osagai gisa identifikatzera bultzatuko du, errore bat eraginez. Errore mota hori, hain zuzen ere, errore guztien artean % 35ekoa izan da.

arrazoia	**portzentaia**
Erroreak letra larrien erabileran	% 35
***Eustagger*en analisi okerrak**	% 29
Erroreak sarrerako formatuan	% 22
Habiaratutako entitateak	% 8
Bestelakoak	% 6

II.7 Taula: *Eihera*ren NERen errore iturriak

Bestalde, *Eustagger*en erroreek eragin zuzena dutela ikusi ahal izan dugu. Prozesu horretatik eskuratutako informazioa baita letra larrien ezaugarriarekin batera NEak identifikatzeko moduluaren funtsa. II.4.1 adibidean bezalako analisi okerrek, beraz, *Eihera*k NEak gaizki mugatzera bideratuko du, aztertutako adibideen % 29a errore horretan jatorria izanik. Adibidean konkretuki, *Eustagger*ek *Lan* hitzaren analisian *IZE_ARR* ordez *ADI_SIN* itzultzean, *Lan* ez da *Lan Ministerio*ko NEaren 1. osagai gisa identifikatuko, *Eihera*ren hutsegitea eraginez.

II.4.1 Adibidea *Eustagger*en analisi okerraren ondorioz gaizki identifikatutako NEa.

Lan	*landu*	***ADI_SIN***	*EZZ*	***O***
Ministerioak	*Ministerioa*	*IZE_IZB*	*ERG*	*B*
egindako	*egin*	*ADI_SIN*	*EZZ*	*O*
azken	*azken*	*DET_ORD*	*EZZ*	*O*
kontaketaren	*kontaketa*	*IZE_ARR*	*GEN*	*O*
arabera	*arabera*	*ADB_ADOARR*	*EZZ*	*O*

Horiez gain, habiaratutako NEen erroreak, portzentaje txikiagoa bada ere, % 8ko tasa, hain zuzen ere, eskuzko errore analisian identifikatutako errore iturburu bat izan da. Izan ere, NE laburrago bat bere baitan duen NE batek identifikazio ataza nabarmenki zailtzen du. Esaterako, *Eibarko J.D. Arrate* NEak bi NE habiaratzen ditu, *Eibar* eta *J.D. Arrate*.

Azkenik, sarrerako testuetan agertzen diren formatu-erroreek zarata handia sortzen dutela ikusi da, zarata hori askotan errore bilakatzen delarik. II.4.2 adibidean O motako osagai arrunt bat (*Galdera*), puntuazio-zeinu faltagatik B bezala mugatzera bultzatzen du, *Eihera*ren hutsegitea eraginez.

II.4.2 Adibidea Errore tipografikoaren ondorioz gaizki identifikatutako NEa.

Ene	*ene*	*IOR_PERARR*	*EZZ*	*O*
Jainkoa	*Jainko*	*IZE_ARR*	*ABS*	*B*
Galdera	***galdera***	***IZE_ARR***	***EZZ***	***B***
zaila	*zail*	*ADJ_IZO*	*ABS*	*O*
da	*izan*	*ADT_A1*	*EZZ*	*O*

II.4.1.1.2 Ikasketa automatikoan oinarritutako tresnak

NER ataza ebazteko IAan oinarritutako metodoen esperimentuetarako II.2.3 atalean deskribatutako *Naïve Bayes*, *C4.5*, *SVM* eta *Abionet* metodoen emaitzekin batera, Whitelaw eta Patricken (2003) lanean deskribatzen den hizkuntzarekiko independentea den NERC sistemaren[16] emaitzak aurkezten dira. Tresna hau hizkuntzarekiko independentea izanik, euskarazko corpusaren gainean zuzenean aplikatu ahal izan da inolako moldaketarik egin gabe. Bere portaera, beste metodoekin alderatuz okerragoa da II.8 taulan ikus daitekeen bezala. Hala ere, metodo-konbinazioetarako interesgarria izan daiteke, erabiltzen duen estrategia zeharo desberdina da eta.

Metodo horien guztien ebaluazioak II.8 taulan laburbiltzen dira. *Abionet*i dagokionez bi hurbilpen egin dira. Horren arrazoia, tresnaren inplementazioan du jatorria. *Abionet*ek, bere jatorrizko inplementazioan, sailkapen-instantziak deskribatzeko erabilitako atributuez gain, aurreko osagaien BIO etiketak erabiltzen ditu uneko osagaiaren etiketa ondorioztatzeko. Portaera hori ordea, Wekan ezin denez errepikatu eta beraz gainerako IAko metodoak ezin direnez berdintasunean ebaluatu, *Abionet* bi modutan ebaluatu da: jatorrizko inplementazioarekin, aurreko osagaien BIO etiketak erabiliz (II.8 taulan *Abionet* bezala adierazita); eta aurreko osagaien BIO etiketak erabili gabe (*Abionet*-BIO taulan). *Abionet*en azken esperimentu horren emaitzak izango dira, beraz, gainerako tresnekin zuzenean konparatu daitezkeenak.

Naïve Bayes (*NB*) metodoak, HPko beste ataza batzuetan oso emaitza onak aurkeztu izan baditu ere NER atazan emaitzak ez dira onak izan. Izan ere, esperimentatutako algoritmo guztietatik portaera eskasena aurkeztutako algoritmoa izan da. Portaera eskas horren jatorria, osagaiak deskribatzeko erabilitako atributu kopuruan izan dezake jatorria. Izan ere, portaera ona erakutsi duen testu-sailkapenaren probleman erabilitako atributu kopurua, NER atazarako erabilitako kopuruarekin alderatuz, oso handia baita. Beraz, *NB* ondo trebatzeko erabilitako atributu kopurua nahikoa ez izatea izan daiteke emaitza eskasaren arrazoia.

SVM metodoak, bestetik, horrelako problemen aurrean bere portaera

[16] *Sydney* izenarekin adierazten da tauletan

	doitasuna	estaldura	F_1
Sydney	% 74,74	% 77,54	% 76,12
Eihera	% 84,65	**% 86,10**	**% 85,37**
NB	% 46,08	% 54,76	% 50,05
C4.5	% 86,74	% 82,57	% 84,60
SVM	% 85,02	% 81,93	% 83,44
***Abionet*-BIO**	% 89,22	% 85,8	**% 87,52**
Abionet	% 89,81	% 85,78	**% 87,75**

II.8 Taula: IAn oinarritutako tresnen NER moduluaren emaitzak banaka ebaluatuz

hobea izan ohi den arren ez ditu *C4.5* teknikak lortutako emaitzak hobetzen. Azpimarratzekoa da *SVM*ren portaera ona ezaugarri kopuru oso handiekin lan egitean lortu izan dela. Gure kasuan, atributu kopurua hain handia ez izatea metodo honen emaitzak *C4.5*enak ez hobetzearen arrazoia izan daiteke.

NER atazarako, erabilitako banakako metodoen artean *AdaBoost* algoritmoan oinarritutako *Abionet* da portaera onena erakutsi duen sailkatzailea. Hala izatea espero zen, CoNLLn irabazle izan baitzen, eta esperimentu hauekin bere portaera ona euskararekin mantentzen dela berretsi dugu. Gainerako IAko algoritmoetan oinarritutako sailkatzaileen emaitzak hauen azpitik geratu dira.

*Eihera*k aldiz, *Abionet*en emaitzetara iritsi ez arren, gainerako IAko algoritmoak baino portaera hobea izan du, metodo guztien artean % 86,10ekin estaldura altuena lortuz. Hizkuntza-ezagutzan oinarritutako tresna horrek espresioen egitura morfosintaktikoa (erregelen bidez) gainerako metodoek baino hobeto esplotatzen duelako izan daiteke. Eta ezaugarri horri esker, hain zuzen ere, NER emaitzak hobetzeko metodo desberdinen konbinaketei begira oso tresna interesgarria dela uste dugu, gainerako metodoei informazio desberdina eskaintzen baitiezaieke, osagarritasuna lortuz.

Euskarazko NER sistemaren portaera hobetzeko bidean, II.3.2 atalean azaldu den bezala, Arregi eta Fernándezen (2002) laneko atributu-aukeraketaren estrategia erabili dugu. Konkretuki, jarraian aurkezten diren emaitzak, osagai bakoitzaren deskribapenerako forma eta lema, bi atributuen araberako aukeraketa eginez egindako esperimentuen emaitzak dira. Alegia, orain arte aurkeztutako emaitzak, forma eta lema, bi atributuak erabiliz ikasitako ereduak aurkeztu baditugu ere, gainerako atributuekin batera bietako

bat soilik erabiliz metodoen portaera aztertu dugu jarraian. Izan ere, bi atributu horiek informazio berdina ez bada ere NER ataza ebazteko antzeko informazioa eskaintzen baitute.

Atributu-aukeraketa hori aplikatzean, gainerako IAko algoritmoek *Abionet*en portaera baino okerragoa izaten jarraitu arren, *C4.5*ek zein *SVM*ek *Eihera*ren errendimendutik gertu geratzea (bat eta bi puntu inguru azpitik) lortzen dute II.9 taulan ikus daitekeen bezala.

	formak + lemak	**formak**	**lemak**
NB	% 50,05	% 50,02	**% 62,36**
C4.5	% 84,60	% 84,59	**% 84,94**
SVM	% 83,44	**% 84,25**	% 83,38

II.9 Taula: NER atazan atributu-aukeraketarekin lortutako F_1 emaitzak

Naïve Bayes algoritmoaren kasuan, atributu-aukeraketa egokiena formaren informazioa baztertuz soilik lema erabiltzea dela ematen du. Hobekuntza batez ere estalduran islatzen da, % 46,08tik % 72,6ra igotzen delarik. Izan ere, proportzionalki *NB* da hobekuntza handiena lortzen duen metodoa. Hala ere, gainerako metodoekin konparatuz gero, sistemaren portaera besteen F_1 neurritik oso urruti jarraitzen du.

C4.5 algoritmoaren emaitzei erreparatuz, lemak soilik erabiltzea dirudi aukeraketa onena, baina kasu honetan hobekuntza oso txikia da, ia esanguratsua ez dela esan daiteke. *SVM*ren kasuan, aldiz, portaera hobea du formak bakarrik kontuan izanda, ia puntu batean hobetzen dituelarik emaitzak.

Atributu-aukeraketaz gain, sistema hobetzeko esperimentatutako adibideen aukeraketa ere ebaluatu da. Jarraian estrategia horren eragina aztertzen dugu.

Orain arte aurkeztutako emaitzak, IAko metodoak trebatzeko, ikasketa- zein ebaluazio-corpuseko B, I eta O kategorien adibide kopuru oso desberdinak erabili dira. Izan ere, testuetako hitz gehienak ez ohi dira NE bateko osagai, alegia, testu batean B edota I motako osagai baino O motako osagai gehiago aurkitu ohi dira. Kategorien arteko desoreka hori ezabatzeko asmoz, ikasketa- zein ebaluazio-corpusetik hainbat O motako osagaiak ezabatuz (ikus II.3.2.1 atala) trebatu eta ebaluatu ditugu metodoak.

II.10 eta II.11 tauletan atributuen eta instantzien aukeraketarekin trebatutako ereduen konbinazio-esperimentuen emaitzak aurkezten dira: atributu guztiak (formak+lemak) erabiliz eta instantzia-aukeraketarekin lortu-

tako emaitzak batetik; eta forma atributua baztertuz soilik lemarekin eta instantzia-aukeraketa aplikatuz lortutako emaitzak bestetik. Murrizketaren emaitzak instantzia-aukeraketa gabeko esperimentuekin konparatu ahal izateko, taulen 1. zutabean, instantzia-aukeraketa gabeko sistemen emaitzak agertzen dira, eta 2. zutabean, berriz, O motako osagaien ezabaketarekin lortutakoak.

Ikasketa-fasean instantzia kopuruak duen eragina aztertu ahal izateko *ikas. denb.* izeneko zutabeetan, ereduen **ikasketa-denbora** segundotan islatzen da.

	F_1 (osoa)	**F_1 (murriz)**	**ikas. denb. (osoa)**	**ikas. denb. (murriz)**
NB	% 50,05	% 62,20	14	7
C4.5	% 84,60	% 84,30	110	61
SVM	% 83,44	% 84,45	35.500	18.900

II.10 Taula: Instantzia-aukeraketa aplikatuta, forma+lemekin lortutako emaitzak (F_1)

	F_1 (osoa)	**F_1 (murriz)**	**ikas. denb. (osoa)**	**ikas. denb. (murriz)**
NB	% 62,36	% 69,18	12	6
C4.5	% 84,94	% 84,74	75	38
SVM	% 83,38	% 84,36	29.900	16.350

II.11 Taula: Instantzia-aukeraketa aplikatuta, lemekin (formak erabili gabe) lortutako emaitzak (F_1)

Emaitzak aztertuz, *SVM* eta *Naïve Bayes* metodoek ikasketa-corpusaren tamainaren murrizketaren aurrean ondo egokitzen direla esan daiteke. *Naïve Bayes*en kasuan, hobekuntza hori estalduraren hobetzearen eskutik dator. *SVM*ren kasuan, hobekuntza puntu batekoa da eta horrekin *C4.5* algoritmoaren emaitzetara hurbiltzea lortzen du. Fenomeno horren arrazoiak adibide kopuru txikian izan dezake jatorria. Hala ere, gehiago gerturatzea lortzen badute ere, esperimentu hauen emaitzek 2002ko CoNLL ataza partekatuko (gaztelera eta holandeserako) irabazle geratu zen *Abionet* sistemaren emaitzen % 3 azpitik egoten jarraitzen dute.

C4.5 metodoa instantzia-aukeraketarekin portaera hobetzea lortzen ez duen metodo bakarra da, izan ere diferentzia txikiarekin baina emaitzak oke-

rragoak dira. Hala ere, diferentzia hain txikia izanik ezin dugu esan galera esanguratsua denik.

Ikasketa-denborei dagokionez, metodo guztietan corpusaren tamaina erdira jaistean ikasketa-denborak ere erdira jaisten dira. Baina bereziki murrizketa hori *SVM* algoritmoan da nabarmena, aplikatzerako garaian eredu azkarrenetakoa bada ere, ikasketa-fase oso motela du eta. Kontuan hartzeko faktorea izan daiteke, beraz, etorkizunean *SVM* corpus handiagoekin erabiltzen bada.

II.4.1.2 Sistema konbinaketen ebaluazioa

Arestian aipatu bezala, NER atazako emaitzak hobetzeko asmoz, banaka ebaluatutako metodoen arteko konbinazioekin esperimentatzea erabaki da. Konbinazioan, metodo bakoitza bere aldetik aplikatzen da eta bakoitzetik lortutako emaitzak konbinatzen dira gehiengoaren bozketa aplikatuz. Alegia, NER ataza ebazteko n metodo konbinatzen ditugula esatean, metodo bakoitzak osagai bakoitzeko itzulitako BIO etiketen arteko gehiengoaren bozketa egiten dela esan nahi da, puntu gehien lortu duen etiketa itzuliz.

Berdinketak daudenean sistemak heuristiko sinple bat aplikatzen du, etiketa egokia ebazteko B, I eta O etiketen ordena mantenduz. Alegia, B eta I etiketen artean berdinketa emanez gero, B etiketa itzuliko du, eta I eta O arteko berdinketa bada, I etiketa. Heuristiko hori, osagaiek testuetan izan ohi duten izaeran oinarrituta definitu dugu.

Lehenbiziko kasuan, B eta I artean aukeratu behar denean, B aukeratzea egokiagoa izango dela suposatuz, hasierako osagai mota gabeko NErik eratzen ez dela ziurtatzeko. Erabakia lokala denez, alegia, heuristikoak ez duenez aurreko osagaiaren kategoria erabiltzen berdinketa ebazteko, aurreko osagaia O denean, uneko osagaiari I kategoria esleitzea ekiditeko (Padró eta Padró, 2006), O motako osagai baten jarraian agertzen den osagaia NE baten parte bada, B motakoa izan behar baitu. Bigarren kasuan aldiz, I eta O kategorien artean aukeratzeko, O motako osagai gehiago egon ohi direnez testuetan, eta ondorioz I motako osagaiak identifikatzeko zailagoak izanik, I kategoria hobestea erabaki da.

Konbinazio mota desberdinak ebaluatu ditugu, familia bereko metodoak konbinatuz zein familia desberdinetakoak erabiliz. Konbinazio horien emaitzak II.12 taulan ikus daitezke, metodoak konbinatuz metodoen bakarkako emaitzak hobetzen direla berresten dituzten datuak, hain zuzen ere. IA-ko algoritmo guztien artean *C4.5* algoritmoan oinarritutako sailkatzaileekin

egindako esperimentuak soilik aurkezten dira, konbinazio gehiago egin diren arren, emaitza interesgarrienak algoritmo horrekin lortu baitira.

	doitasuna	**estaldura**	**F_1**
C4.5*, *C4.5*-Lem, *Eihera	% 86,65	**% 84,71**	% 85,67
Sydney*, *C4.5*-Lem, *Eihera	% 88,71	**% 87,38**	% 88,04
Sydney*, *C4.5*, *Eihera	% 89,49	**% 87,38**	**% 88,42**
Abionet*, *Sydney*, *C4.5	**% 90,72**	% 85,78	% 88,18
Abionet*, *Sydney*, *Eihera	% 90,08	**% 87,38**	% 88,71
Abionet*, *C4.5*, *Eihera	% 90,47	% 87,27	**% 88,84**

II.12 Taula: NER atazan metodo konbinaketa eta gehiengoaren bozketa aplikatuz lortutako emaitzak

Egindako esperimentuetan, familia desberdinetako metodo/tresnak konbinatzean emaitzak nabarmenki hobetzen direla ikusi dugu, eta mota berdineko sailkatzaileak konbinatzean, aldiz, hobekuntzak txikiagoak direla. Are gehiago, izaera derberdineko hiru metodo konbinatzean (*Sydney*, *Eihera*, *C4.5*) bakarkako NER sistema onenetakoa den *Abionet* sistemaren emaitzak gainditzea lortzen da (% 88,42 vs. % 87,75). Hobekuntza horren arrazoi nagusia estalduran lortzen den igoeran du jatorria.

Sydney sistemak bakarka erabiltzean emaitzarik okerrenetarikoak lortu arren, konbinazioaren portaeran eragin handia duela esan daiteke, bereziki sistemaren doitasunean. II.12 taulako lehenengo bi lerroetan ikus daitekeen bezala, sistema konbinatuan *C4.5* erabili beharrean *Sydney* erabiltzean, sistemaren estaldura ia 3 puntutan hobetzen da. Eragin hori, sistemak erabiltzen duen informazio mota desberdinaren eskutik datorrela ematen du (Patrick *et al.*, 2002).

Konbinazioen doitasunean eragin positiboa duen beste tresna *Abionet* da, bakarka aplikatzean doitasun emaitzarik onenak lortzen dituen sistema, hain zuzen ere. 4. eta 5. konbinazioen emaitzak aztertuz gero, *Abionet* eta *Sydney*, *C4.5* eta *Eihera*rekin konbinatzean lortutako emaitzak, hurrenez hurren, lehenengoak konbinazioari doitasun hobea izaten laguntzen diola ikus daiteke, eta *Eihera*k aldiz, estaldura hobetzen laguntzen du. Izan ere, estaldurarik onenak lortzen dituzten sistema-konbinazio guztietan *Eihera* erabiltzen da, bakarkako emaitzetan ikusi ahal izan den bezala, hori bera izanik bere ezaugarririk onena.

Azkenik aipatu, espero bezala, hobekuntza handia ez bada ere, *Abionet* eta bakarka emaitzarik onenak lortzen dituzten beste bi sistemak konbina-

tzean emaitzarik onenak lortu direla, II.12 taulako azken lerroan azaltzen den bezala.

Laburbilduz, izaera desberdineko sailkatzaileak konbinatzea NER atazaren ebazpenerako sistemak hobetzeko bide egokia da. Izan ere, izaera desberdinekoak izateak, konbinazioari bakoitzaren abantailak erabiliz sistema sendoagoa eraikitzeko aukera eskaintzen duela ematen du. Izaera desberdin horien artean, *Eihera*ren ekarpena aipagarria da, eskaintzen duen informazio morfosintaktikoaren erabilpen egokiari esker, sistema konbinazioetan estalduran hobekuntza handiak eragiten dituela erakutsi baita.

II.4.2 NEC

NERerako erabilitako ikasketa- zein ebaluazio-corpus berdinak erabili badira ere, kontuan izanik NERerako hitzak (46.227 ikasketan eta 15.960 testean) direla sailkatu beharreko unitateak eta NECen aldiz NEak (3.817 ikasketan eta 931 testean), NEC atazarako datu kopurua txikiagoa dela esan daiteke. Datu kopuru txiki horrek, bereziki IAko esperimentuetan izango du eragina, ikasketarako adibide gutxi egongo baitira.

Esan behar da, atal honetan NEC bakarrik ebaluatu ahal izateko eta NEen identifikazioan gerta daitezkeen erroreak ekiditeko, *Eihera* zein IAko metodoak ebaluatzeko eskuz mugatutako ebaluazio-corpuseko 931 NEetatik abiatu garela. Aurrerago, NERC emaitzak azaltzean (ikus II.4.3 atala) NER atazan automatikoki identifikatutako NEak sailkatzearen emaitzak aurkeztuko ditugu.

Lau dira NEC moduluak sailkatu beharreko kategoriak: PER, LOC, ORG eta OTH. Azken kategoriarako bereziki adibide oso gutxi daudenez, OTH kategoriako emaitzei ez diegu garrantzi handirik emango, eta PER, LOC eta ORG kategorien azterketan zentratuko gara.

NEC moduluan, sistemak beti sailkapen bat eta bakarra itzultzeko diseinatu direnez, sistema orokorraren estaldura eta doitasun neurriak baliokide dira kasu honetan, F_1 balioa ere baliokide bihurtuz. Baliokidetza hori, ordea, ez da ematen NEen sailkapena kategoriaka ebaluatzean, aurrerago ikusiko den bezala. Hortaz, NEC sistema kategoria guztiekin ebaluatzean emaitza tauletan neurri horietako bakarra aurkeztuko da, eta kategoriaka emaitzak aurkeztean aldiz, hiru neurriak aurkeztuko dira.

Lehenbizi NEC ebazteko metodo bakoitzaren portaera banaka aztertzeko sistemen bakarkako ebaluazioak aurkezten dira. Eta, NERerako egin dugun bezala, ondoren, metodo horien konbinaketa desberdinen emaitzak aurkezten

dira.

II.4.2.1 Sistemen ebaluazioa banaka

II.4.2.1.1 Eihera

Eihera tresna hasiera batean NER ataza soilik ebazteko diseinatu bazen ere, aurrerago NEC ataza ebazteko heuristiko sinple bat gehitu zitzaion. Heuristikoak erabiltzen duen informazioa, oinarrian, NER atazan erabiltzen den berdina da, baina oraingoan diseinu atalean erakutsi den bezala (ikus II.3.1.2 atala), informazioa NEC atazarako interpretatzen da. Heuristikoa hain sinplea izanik, II.13 taulan ikus daitekeen bezala, emaitza egokiak lortzen direla esan daiteke.

	NER	**NEC**
Eihera	% 85,37	% 72,93

II.13 Taula: *Eihera*ren NEC emaitzak (F_1)

Are gehiago, sofistikatuagoak diren sailkapen-metodo batzuekin lortutako emaitzekin alderatuz gero, aurrerago erakutsiko den bezala, heuristiko hau ez dela horien portaeratik hain urruti geratzen esan daiteke. Konkretuki 931 NEko ebaluazio-corpusean, 679 entitate ondo sailkatzen ditu *Eihera*k NEC atazan % 72,93ko zehaztasunarekin.

II.4.2.1.2 Ikasketa automatikoan oinarritutako tresnak

NEC ataza ebazteko IAko metodoen artean NER atazako esperimentuetan ere erabilitako *C4.5*, *SVM* eta *Abionet* metodoak aurkeztuko ditugu. Metodo horiek, II.3.2.2 atalean aztertu den bezala, NEen osagaien barne-ebidentziaz gain NE baten kategoria egokia aukeratzeko *gazetteer* eta hitz eragileen zerrendak erabiltzen dute. Lehenbiziko esperimentuetarako bereziki *Euskaldunon Egunkari*tik eta errolda-zerrenda batekin osatutako zerrendak erabili dira (ikus II.2.1 atala). Bigarren esperimentazio fasean, berriz, *EusWN* baliabidean oinarritutako aberasketarekin osatutako zerrendekin ebaluatu ditugu metodoak. Bi esperimentu multzoen emaitzak II.14 eta II.15 tauletan ikus daitezke, hurrenez hurren.

Emaitzak aztertuz, erraz ikus daiteke *C4.5* algoritmoa dela NEC ataza ebazteko banakako sistemetan onena *EusWN*ko informazioa erabili gabe

	$\mathbf{F_1}$
C4.5	**% 75,93**
SVM	% 74,43
Abionet	**% 72,93**
Eihera	**% 72,93**

II.14 Taula: IAko metodoen NEC emaitzak

	$\mathbf{F_1}$
$\mathbf{C4.5_{EusWN}}$	% 76,69
$\mathbf{SVM_{EusWN}}$	% 74,43
$\mathbf{Abionet_{EusWN}}$	**% 71,75**

II.15 Taula: IAko metodoen NEC emaitzak *EusWN* aberasketarekin

(II.14 taula) zein informazio hori erabilitako (II.15 taula) esperimentuetan. Zehaztasun onena berak agertzen du eta hurrengo onena *SVM* da, 1,5 puntura. NER atazan, algoritmo onena izan ez den arren, onenetarikoen artean kokatu dela kontuan hartuz, *C4.5* metodoa NEen tratamendurako izaera homogeneoena aurkezten duen algoritmoa dela esan daiteke.

*Abionet*i dagokionez, NER atazan oso portaera onekoa izan bada ere, emaitzetan NECerako ez dirudi oso egokia denik, *Eihera*k berak heuristiko sinple batekin horren emaitzak parekatzen baititu. Azpimarratzekoa da ere, *EusWN*en informazioak esperimentu hauetarako laguntza handirik ez duela ematen, *C4.5* baita informazio horrekin emaitzak pixka bat hobetzea lortzen duen metodo bakarra. Hala ere, izaera desberdineko metodo eta informazioiturriak izanik, beraien osagarritasuna egiaztatzeko asmoz metodo konbinatuetan horiek ere erabiliko dira.

	NE kop	**doitasuna**	**estaldura**	$\mathbf{F_1}$
PER	293	% 76,60	**% 81,57**	% 79,01
LOC	315	**% 85,96**	% 79,68	**% 82,70**
ORG	293	% 68,54	% 75,08	**% 71,66**
OTH	30	% 66,67	% 13,33	% 22,22

II.16 Taula: $C4.5_{EusWN}$en NEC emaitzak kategorien arabera

II.16 taulan NE mota bakoitzeko ebaluazio-corpuseko adibide kopurua

eta banakako NEC sistema onenaren ($C4.5_{EusWN}$) emaitzak aurkezten dira. Esan, OTH kategoria, ebaluazio- zein ikasketa-corpusean, gainerako sailkapenekin alderatuz, oso adibide gutxiko kategoria dela, eta hori emaitzetan islatzen dela. Azken batean hain adibide gutxirekin, IAko metodoek kategoria horretarako ezin dute eredua behar den bezala trebatu. OTH kategoriako adibide kopurua ezik, gainerako kategorien kopuruak antzekoak dira.

Hiru kategoria nagusien (PER, LOC, ORG) emaitzak aztertuz, hiruen artean NEC atazan kategoria zailena ORG dela ikus daiteke, eta egoera hori IAko metodo guztietan ematen da. Konkretuki, F_1 neurriari erreparatuz gero ORGen emaitzak LOC kategoriaren % 10 azpitik geratzen dira bataz beste esperimentu guztietan. Sistemen doitasuna aztertuz gero, LOC kategoriak lortzen ditu emaitzarik onenak, eta estaldurari erreparatuz, PER kategoria gailentzen da.

*Abionet*en kasuan, *EusWN* informazioaren aurrean sistemaren portaeran okertzearen iturria aztertu nahian kategoriaka analisi sakonagoa egin da. II.17 taulan aurkezten dira bi esperimentuen emaitzak kategoriaren arabera.

	NE kop	**doitasuna**	**estaldura**	**F_1**
PER	293	% 77,35	% 75,77	% 76,55
LOC	315	% 75,07	% 80,32	% 77,61
ORG	293	**% 68,71**	**% 68,94**	**% 68,82**
OTH	30	% 15,38	% 6,67	% 9,30
PER (EusWN)	293	% 77,43	% 76,11	% 76,76
LOC (EusWN)	315	% 74,70	% 79,68	% 77,11
ORG (EusWN)	293	**% 65,10**	**% 66,21**	**% 65,65**
OTH (EusWN)	30	% 0,00	% 0,00	% 0,00

II.17 Taula: *Abionet* eta $Abionet_{EusWN}$en NEC emaitzak kategoriaren arabera

Ikus daitekeenez, PER eta LOC kategorien sailkapenak gutxi gora-behera mantentzen badira ere, ORG kategoriaren sailkapenaren jaitsierak eragiten du sistemaren bataz besteko portaera okertzea.

Atributuei dagokienez, *C4.5* zein *Abionet* metodoak *gazetteer* zein hitz eragile zerrendarik gabe trebatutako esperimentuak burutu ditugu, kanpoebidentziak NEC atazan duen eragina aztertzeko asmoz[17]. Emaitzak II.18 taulan laburbiltzen dira.

[17]Hitz eragileak baztertu direnean *-trigger* agertzen da taulan eta *gazetteer*ak baztertzean, berriz, *-gazetteer*

	$\mathbf{F_1}$
C4.5	% 75,93
C4.5-trigger	% 75,83
C4.5-trigger-gazetteer	% 61,86
Abionet	% 72,93
***Abionet*-trigger**	% 72,28
Abionet-trigger-gazetteer	% 70,67

II.18 Taula: NEC atazan atributu-aukeraketarekin lortutako emaitzak

Hitz eragileen zerrendak erabili gabe emaitzak asko okertzen ez badira ere, *gazetteer*en informazio ezak sistemei kalte handiagoa egiten diela ikusi ahal izan dugu. *C4.5* kasuan, bi iturriak kentzean 14 puntuko jaitsiera ematen da. *Abionet*en kasuan, aldiz, *gazetteer*ak erabili gabe sisteman portaera okertzen badu ere, 2 puntuko galera besterik ez da ematen. *Abionet* beraz, kanpo-ebidentziekiko independenteagoa dela ematen du, eta *C4.5*ek aldiz, NEC ebazteko kanpo-ebidentzia asko baliatzen duela.

*Gazetteer*en informazioa erabiltzen ez dituzten sailkatzaileen azterketa kategoriaka egitean, LOC kategoriaren emaitzetan galera txikiagoa dagoela ikusten da. *C4.5* metodoaren emaitzak II.19 taulan aurkezten dira. Bertan ikus daitekeenez, PER eta ORG kategorien kasuan, % 10eko jaitsiera gertatzen da. Esan beharra dago, hala ere, beste metodo batzuetan portaera hori errepikatzen bada ere, galera ez dela hain nabarmena.

	doitasuna	**estaldura**	$\mathbf{F_1}$
PER	% 75,95	% 81,91	% 78,82
LOC	% 86,96	% 76,19	% 81,22
ORG	% 66,97	% 76,11	**% 71,25**
OTH	% 66,67	% 13,33	% 22,22
PER-gazetteer	% 67,44	% 69,28	% 68,35
LOC-gazetteer	% 77,97	% 73,01	% 75,41
ORG-gazetteer	% 57,56	% 65,19	**% 61,02**
OTH-gazetteer	% 50,00	% 3,33	% 6,25

II.19 Taula: *C4.5 gazetteer*ekin eta gabe erabiltzean NEC emaitzak kategorien arabera

Hortaz, hitz eragileak NEC atazan lagungarriak baina ez nahitaezkoak izan arren, *gazetteer*ak oso beharrezkoak direla esan daiteke. Hala ere, az-

ken informazio mota horren gabeziaren eragina sistemaren sendotasunaren araberakoa izango da. Zerrenda horiek, esaterako, metodoek egiten duten informazioaren erabilera desberdinaren ondorioz, *C4.5* metodoarentzat *Abionet*entzat baino garrantzitsuagoak direla esan daiteke.

II.4.2.2 Sistema-konbinaketen ebaluazioa

NEC atazako sistema hobetzeko bidean, sailkatzaile desberdinen konbinazioak burutu ditugu gehiengoaren bozketa bidez. Konbinazio horietarako hiru bide jarraitu ditugu, konbinazio mota desberdinak neurtzeko asmoz:

- Metodo bera baina informazio-iturri desberdinekin sortutako sailkatzaileen konbinazioak.
- IAko metodo desberdinak konbinatuz: *Abionet*, *SVM*, *C4.5* eta *NB* metodoekin eratutako sailkatzaileak.
- Hurbilpen desberdinekin sortutako sailkatzaileen konbinazioak: IAko metodoak *Eihera*rekin konbinatuz.

NER atazan bezala, NEC atazan sailkatzailerik onenak sailkatzaile onenen konbinazioak burutzean lortzen direla ikusiko dugu. Ataza honetan ordea, hobekuntza ez da hain nabarmena, beharbada, ikasketa-corpusaren tamaina txikia dela eta.

NEC sistemen konbinazioa, NERen antzera, bakarkako sistemen emaitzen gehiengoaren bozketarekin burutu dugu: sistema bakoitza bere aldetik aplikatu eta emaitza guztiekin gehiengoaren bozketa egiten da, sistema guztiek pisu bera izanik, kategoria garailea itzultzen da sistema konbinatuaren emaitza bezala. Berdinketa kasuetan, sistemak PER, ORG, LOC eta OTH lehentasun ordena mantenduz ebatziko du zein den itzuli beharreko kategoria. Esaterako, ORG eta LOC arteko berdinketaren aurrean, sistemak ORG itzuliko du.

Lehenbiziko esperimentuan, informazio-iturri desberdinetan oinarritutako *C4.5* ereduak konbinatzean, hiru horietatik onena den $C4.5_{EusWN}$ sistemaren emaitzak lortzen dira inolako hobekuntzarik gabe, II.20 taulako 1. lerroan ikus daitekeen bezala. Antzeko zerbait gertatzen da $C4.5_{EusWN}$ IAko *SVM* eta *Abionet* metodoekin konbinatzean, berriz ere $C4.5_{EusWN}$ sistema onenaren emaitzak mantentzea besterik ez da lortzen. Beraz, IAko metodoen goi-muga dela ematen du.

	$\mathbf{F_1}$
C4.5$_{\mathbf{EusWN}}$, ***C4.5*-trigger**, ***C4.5***	% 76,69
C4.5$_{\mathbf{EusWN}}$, ***SVM*, *Abionet***	% 76,69
C4.5$_{\mathbf{EusWN}}$, *Abionet*, *Eihera*	% 78,73

II.20 Taula: NEC atazan metodo konbinazioarekin lortutako emaitzak

Muga hori gainditzeko asmoz, hurbilpen desberdinetako metodoak konbinatu dira eta II.20 taulan aurkezten da *Abionet*, *C4.5* eta *Eihera*ren arteko konbinazioa.

*Eihera*k banakako sistemetan emaitzarik okerrenetakoak lortzen baditu ere, sistema konbinatuak sortzeko erabiltzean, $C4.5_{EusWN}$ sistemaren bakarkako emaitzak hobetzea lortzen da. Hortaz, informazio mota eta estrategia desberdinak konbinatzen dituzten sistemek emaitzak hobetzen dituzten hipotesia berresten da. Konbinazio gehiago egin ditugun arren, emaitza aipagarrienak besterik ez ditu islatzen II.20 taulak.

Emaitzarik onenak *Abionet*, *Eihera* eta $C4.5_{EusWN}$ aberasketarekin konbinatzean lortzen direla ikusirik, *Eihera*ren aportazioa neurtu nahi izan da. Horretarako sistema *Abionet* eta $C4.5_{EusWN}$ konbinazioarekin ebaluatu da, batez beste *Eihera* erabiltzen duen sistemarekiko % 2ko galera ematen delarik, alegia IAko metodoen goi-mugara jaisten da sistema.

	doitasuna	**estaldura**	$\mathbf{F_1}$
***PER* (C4.5$_{\mathbf{EusWN}}$)**	% 76,60	% 81,57	% 79,01
***LOC* (C4.5$_{\mathbf{EusWN}}$)**	% 85,96	% 79,68	% 82,7
***ORG* (C4.5$_{\mathbf{EusWN}}$)**	% 68,54	% 75,08	% 71,66
***OTH* (C4.5$_{\mathbf{EusWN}}$)**	% 66,67	% 13,33	% 22,22
***PER* (C4.5$_{\mathbf{EusWN}}$, *Abionet, Eihera*)**	**% 82,03**	**% 82,59**	% 82,31
***LOC* (C4.5$_{\mathbf{EusWN}}$, *Abionet, Eihera*)**	% 80,84	**% 85,71**	% 83,2
***ORG* (C4.5$_{\mathbf{EusWN}}$, *Abionet, Eihera*)**	**% 73,99**	% 74,74	% 74,36
***OTH* (C4.5$_{\mathbf{EusWN}}$, *Abionet, Eihera*)**	% 33,33	% 6,67	% 11,11

II.21 Taula: NECeko bakarkako eta sistema konbinatu onenen emaitzak kategoriaren arabera

Konbinazioen eragina kategoriaka emaitzetan aztertzeko II.21 taulan agertzen diren bakarkako sistema onenaren ($C4.5_{EusWN}$) emaitzak eta sistema konbinatu onenarenak ($C4.5_{EusWN}$, *Abionet, Eihera*) aztertu ditugu.

Estaldurari erreparatuz, PER eta LOC kategorien kasuan sistema konbinatuak sinpleak baino portaera hobea aurkezten du, ez ordea ORG kategorian, estaldura, gutxi bada ere, jaitsi egiten baita.

Doitasunean, berriz, sistema konbinatuak PER eta ORG kategorien sailkapenean % 5eko hobekuntza lortzen du. LOC motako adibideak sailkatzean, ordea, emaitzak okertu egiten dira, doitasuna % 5ean jaitsiz. Kategoria horrentzat sistemak lortzen duen estaldura hobekuntzari esker (% 6), konpentsatu egiten da; bataz bestean, F_1 neurrian hobekuntza txikia lortuz.

Oro har, sistema sinpletik konbinatura pasatzean, hobekuntzarik onena PER kategorian gertatzen da. Hobekuntza horren arrazoi nagusia, kategorien arteko berdinketa kasuetan, PER hobesten den kategoria izatea dela ematen du.

II.4.3 NERC

Orain arte NER eta NEC sistemak aparteko bi ataza gisa aurkeztu badira ere, NEak erauzteko sistema bat eraikitzeko biak dira beharrezkoak. Atal honetan NER eta NEC atazak modu sekuentzialean aplikatzean, metodo zein konbinazio desberdinak erabiliz sistemak erakusten duen portaera aztertzen da.

Kontuan hartu behar da NEC modulua NER atazaren emaitzen gainean aplikatzean, ongi identifikatutako entitateak sailkatzeaz gain, identifikazioan oker markatu diren NEak ere sailkatuko dituela. Oker mugatutakoetan NE izan gabe NE gisa mugatu diren egiturak, gehiegizko osagaiak edo osagaiak faltan dituzten espresioak, etab. sartzen dira. Hortaz, sistemaren estaldura zein doitasun neurriak jaistea esperokoa da.

Bi NERC esperimentu mota nagusi bereiz daitezke: NER zein NEC atazetan metodo bakarra aplikatuz eraikitako sistemak, eta metodo konbinazioekin lortutakoak.

II.22 taulak islatzen duen bezala, *Eihera*ren informazioa ez da nahikoa NERC sistema eraginkor bat sortzeko. Beste hurbilpenekin konparatuz gero gutxienez 2-3 puntuko aldea du F_1 neurrian. *Abionet*ekin eta *C4.5*ekin modu independentean lortutako sailkatzailerik onenak aplikatzean, *Eihera*ren emaitzak hobetzen badira ere, oraindik ere ez dira nahikoa onak NERC sistema baterako. NER metodo onena (*Abionet*) NEC metodo onenarekin (*C4.5*) konbinatuz, emaitzak hobetzen dira, aurrekoekin alderatuz 3 puntuko igoera lortzen delarik.

Emaitzak hobetzen jarraitzeko asmoarekin sistemaren konplexutasuna

NERC = NER - NEC	**doitasuna**	**estaldura**	F_1
Eihera* - *Eihera	% 62,14	% 63,48	% 62,80
Abionet* - *Abionet	% 66,62	% 63,90	% 65,24
***C4.5* - C4.5$_{\mathbf{EusWN}}$**	% 67,97	% 64,98	% 66,44
***Abionet* - C4.5$_{\mathbf{EusWN}}$**	% 71,19	% 68,20	% 69,62
***Abionet*,*C4.5*,*Eihera* - *Abionet*,*Eihera*,C4.5$_{\mathbf{EusWN}}$**	% 72,50	% 70,24	**% 71,35**

II.22 Taula: NERC emaitzak

handitu da. Horretarako ataza bakoitzean portaera onena agertu duten sailkatzaile konbinatuak sekuentzialki aplikatu dira, *Abionet*,*C4.5*,*Eihera* eta *Abionet*,*Eihera*,$C4.5_{EusWN}$ identifikazio eta sailkapenerako, hurrenez hurren. Konbinazio horrekin lortutako emaitzak *Eihera* % 8an eta gainerakoak, *Abionet* - $C4.5_{EusWN}$ ezik, % 5ean gainditzea lortu da. *Abionet* - $C4.5_{EusWN}$ kasuan, hobekuntza % 1,5ekoa da.

Oro har, NERC ataza ebazteko, NER eta NEC ataza sekuentzialak ebazteko hurbilpen desberdinetako sistema sinpleak konbinatzean % 70eko F_1a gainditzea lortu dugu. Beraz, sistema sinpleen konbinaketarekin NERC sistema onargarria sortzea bideragarria dela esan daiteke.

	doitasuna	**estaldura**	F_1
PER	% 77,90	% 73,38	% 75,57
LOC	% 75,46	% 78,09	% 76,75
ORG	% 66,43	% 62,80	% 64,56

II.23 Taula: NERC sistema onenaren emaitzak kategorien arabera

Abionet, *C4.5*, *Eihera* - *Abionet*, *Eihera*, $C4.5_{EusWN}$ konbinazio onenetik lortzen diren emaitzak kategoriaka aztertzen baditugu (ikus II.23 taula), hiru kategoria usuenen artean ORG kategoria gainerako kategorien % 11 azpitik geratzen dela ikus daiteke. Esperimentu guztietan hiruetan sailkatzen kategoria zailena ORG dela konfirmatzen da.

Emaitzak aztertuz erraz ikus daiteke emaitzarik onenak hizkuntza-ezagutzan oinarritutako tresnak parte hartzen duen konbinazioekin lortzen direla. Beraz, informazio hori bakarrik nahikoa izan ez arren, oso aberasgarria eta, neurri batean, IAko metodoen osagarria dela ondoriozta daiteke.

Hemendik aurrera *Abionet*, *C4.5*, *Eihera - Abionet*, *Eihera*, $C4.5_{EusWN}$ sistema ***Eihera+*** izenarekin erreferentziatuko da.

II.4.4 Beste hizkuntzetako emaitzak

Euskararako NERC beste hizkuntza batzuekin konparatu nahian, *Abionet* sistema erabiliko dugu erreferentzia gisa. Alde batetik, sistema horren euskararekiko portaera ezaguna delako, tesi-lan honetan erakutsitako emaitzak batik bat, eta bestetik CoNLL-2002 eta CoNLL-2003 edizioetan lehenengo postuen artean geratu izanak erreferentzia ona dela adierazten baitu. Zehazki, 2002ko edizioan gaztelera zein nederlandar hizkuntzetan irabazlea izan zen eta 2003ko edizioan, berriz, 5. postuan sailkatu zen ingelesa zein alemanerako. *Eihera* eta *Abionet* sistemen portaera F_1 neurriaren arabera neurtzean, % 2,5eko aldea dagoela ikusi dugu, *Abionet* (% 65,24) *Eihera* (% 62,8) baino hobea delarik (ikus II.22 taula).

	Abionet	***Konbinatua***
Ingelesa	% 85,00	% 90,30
Gaztelera	% 81,39	-
Nederlandera	% 77,05	-
Alemana	% 69,15	% 74,17
Euskara	% 65,24	% **71,35**

II.24 Taula: *Abionet* NERC emaitzak (F_1) hizkuntza desberdinetarako

II.24 taulan *Abionet*en NERC emaitzak hainbat hizkuntzatarako (Carreras *et al.*, 2002, 2003) eta euskararako islatzen dira. Horrez gain, euskararako gure metodo konbinazio onena eta ingelesa eta alemanerako CoNLL 2003 edizioan aurkeztutako sistemen konbinazioekin[18] lortutakoak ere gehitu dira. Ikus daitekeenez, guztietan hizkuntza, corpus eta ezaugarri desberdinak erabiltzen direla kontuan hartuz, ezaugarri morfologiko konplexuak dituen alemanieratik gertu dago euskara, eta besteetan gertatzen den bezala metodo konbinazioekin bakarkako sistemen emaitzak hobetzen dira.

Euskararen emaitzak morfologia aberatsa duen alemanarekin konparatuta, 3 puntu azpitik geratzen dira. Hala ere, kontuan hartu behar da, CoNLLen luzatutako corpusak guk erabilitako corpusa baino handiagoa izanik (ikus II.25 taula), sistemak hobeto trebatzeko aukera eskaintzen duela, eta

[18] `http://www.clips.ua.ac.be/conll2003/pdf/14247tjo.sh.pdf`

gure corpus tamaina txikiak aldiz, lan hori zaildu besterik ez duela egiten. Beraz, euskararako esperimentuak corpus handiagoekin errepikatzea komenigarria litzateke.

	Ikasketa-corpusa	*Test-corpusa*
Ingelesa	23.499	5.648
Gaztelera	18.797	3558
Alemana	11.851	3.673
Euskara	3.817	931

II.25 Taula: Hizkuntza desberdinetarako corpus tamainak

Hala ere, euskararako NERC sistema sortzeko egindako esperimentu guztien artean emaitzarik onenak, NERerako *Abionet*, *C4.5* eta *Eihera* eta NECen *Abionet*, *Eihera* eta $C4.5_{EusWN}$ sekuentzialki aplikatzean (*Eihera+* tresna) lortu ditugu. Sistema horrek *Abionet* corpus berarekin trebatu eta ebaluatzean lortzen duen F_1 neurria % 5 inguruan gainditzen du. Hortaz, *Abionet* erreferentzia ontzat jotzen badugu, euskararako NERC atazan lortutako emaitzak ingelesa ez diren beste hizkuntza batzuekiko konparagarriak direla esan dezakegu.

II.5 Ondorioak

Kapitulu honetan euskararako NERC sistema baten diseinua eta garapena aurkeztu da. Euskara hizkuntza eranskaria izanik, garatutako sistema aplikatu aurretik ezinbestekoa da aurreprozesu linguistikoa aplikatzea hainbat atributu erauzi ahal izateko.

Lanaren hasieran finkatutako helburuak bete ditugula esan daiteke. Izan ere, metodo sinpleak eta baliabide urriak erabiliz euskarako entitate-izenak mugatu eta sailkatzen dituen tresna eraginkorra martxan da. Egindako hipotesi nagusia, alegia, hizkuntza-ezagutzan eta ikasketa automatikoan oinarritutako metodoak konbinatzeak emaitza onak lortuko zituela egiaztatu dugula azpimarratzekoa da ere.

Euskararako NERC atazarako garatutako hizkuntza-ezagutzan oinarritutako *Eihera* tresnaren erabilgarritasuna ondo frogatuta geratu da. Izan ere, horrelako tresna batek corpus baten etiketatze-lana arintzeko, guztiz eskuzkoa izan beharrean modu erdi-automatikoan egiteko bide ona dela ikusi

dugu. Gainera, bere diseinuan egiten den azterketa sakonak ataza ebazteko atributu gakoak eta interesgarriak identifikatzen laguntzen du, bereziki IAko metodoekin lan egiteko garaian oso lagungarriak gertatzen dira. Azkenik, aipatzekoa da ere, tresna horri esker sistema hibrido eraginkor bat sortzea posible izan dugula. Gainera, *Eihera* IAko metodoekin konbinatzeko informazio-iturri aberats eta egokia gertatu da, NEen identifikazio zein sailkapenean.

Esperimentuetan erabilitako atributu guztien artean atributu linguistikoen garrantzia azpimarratzekoa da. NER atazarako lemen informazioa nahikoa bada ere, NEC atazan emaitza onak lortu ahal izateko deklinabide eta kasu markak erabiltzea beharrezkoa dela ikusi da.

NECen oso lagungarria gertatu den beste atributua *gazetteer*en erabilpenetatik eskuratutakoa da. Izan ere, ataza horretako esperimentuetan NEC sistema on baterako *gazetteer*ak ia ezinbestekoak direla ikusi ahal izan dugu.

NEC atazan hain garrantzitsuak diren *gazetteer*ak, baita hitz eragileen zerrendak automatikoki aberastu eta sistemaren portaera hobetzeko asmoz, *EusWN* baliabidea erabili dugu. Konkretuki, osagaien *synset*ak adierazita dituzten Magnini *et al.* (2002) egileen laneko ingelesezko zerrendak abiapuntutzat hartuz, *synset* horien euskal formak *EusWN*etik eskuratuz eta *gazetteer*/hitz eragileen zerrendetan gehituz aberasketa prozesu automatikoa burutu dugu. Zerrenda aberastuek ekarri duten hobekuntza nabarmenena C4.5 algoritmoarekin eman da, ia puntu bateko aldea lortuz F_1 neurrian. Gainerako metodoetan, aldiz, ez da hobekuntza nabarmenik igarri.

NER zein NEC atazetan egindako instantzia-hautaketari dagokionez ikasketa-prozesua arintzeaz gain, saiakuntza horiek sistemaren emaitzak hobetzeko aukera dagoela frogatu dute.

Edozein kasutan, euskararako eskuragarri izan ditugun baliabide mugatuek NEen identifikazio eta sailkapen atazak ebazteko lana zaildu egin digute. Baliabide urriek bereziki NEC atazan eragina izan dute. Esaterako, OTH adibide kopurua hain txikia izan da, sistema ezin izan dugula prestatu horrelako NEak sailkatzeko.

Aipatzekoa da ere, ezaugarri morfosintaktikoak automatikoki eskuratzeko *Eustagger*en gertatzen diren erroreek eragin handia dutela sistemaren portaeran. Hori dela eta, etorkizunean *Eustagger*en egiten diren hobekuntzek euskarazko NERC sisteman eragin positiboa izatea espero daiteke, eta ondorioz NERC sistemaren portaera hobetzea.

Hala ere, beste hizkuntzetarako sistemak garatzeko erabilitako baliabideak baino txikiagoak izan arren, ingelesa ez diren beste hizkuntzentzat

CoNLLeko sistema onenetatik gertu dagoen euskararako *Eihera+* sistema hibridoa lortu dugula esan dezakegu, NERC F_1 neurrian % 71,35eko asmatzetasarekin ebazten duen tresna garatzea lortu baitugu.

CoNLLen aurkeztutako NERC sistemen emaitzak gainditzen dituzten sistemak aurkeztu izan dira ere (Nadeau eta Sekine, 2007). Hala ere, hobekuntzak ez dira hain handiak izan, eta horiek eraikitzeko erabili diren tekniken konplexutasuna kapitulu honetan aurkeztutako metodoen konplexutasuna baino handiagoa da. Konplexutasun horrek, NERC sistema garatzeko esfortzu handiagoa eskatzen du. Gure helburua metodo sinpleekin eta horien konbinazioarekin NERC sistema eraginkorra sortzea izan denez hasieratik, teknika konplexuago horien aplikazioa tesi-lan honen esparrutik kanpo geratu da.

Etorkizunean euskararako NERC sistema hobetzeko bidean, baliabide handiagoak sortu eta erabiltzea komenigarria ikusten dugu NER zein NEC atazetako emaitzak hobetu ahal izateko. Ez soilik ikasketa- zein ebaluazio-corpusen ikuspuntutik, sisteman erabiltzen diren kanpo-ebidentzien ikuspuntutik ere. Horregatik, *gazetteer* zerrenden aberasketarako esperimentuak egitea komenigarri ikusten dugu, batez ere IV. kapituluan landu dugun Wikipedia baliabidea ustiatuz. Izan ere, *Gazetteer*ak NEC sistemen hobekuntzarako bide ona dela frogatu dugu, informazio horrek sailkapenean duen garrantzia batez ere.

NEC hobetzeko bidean, *EusWN*en erabilera hobeto aztertzea ere interesgarri ikusten dugu, *C4.5* ez diren metodoetan ere lagungarria izan dadin.

Azkenik, landu ez diren eta sistematik at geratu diren NE sendoen identifikazioa eta sailkapen xehea, etorkizuneko lanen artean aipatzea ere ezinbestekoa da. Izan ere, horrelako NEen zehaztasun handiagoa helburu orokorragoko aplikazioetan oso lagungarria izan baitaiteke.

Eihera+ tresna, txosten honen hurrengo kapituluetan NEen itzulpen eta desanbiguazioko atazak ebazteko oinarrizko tresna gisa erabiltzeaz gain, bestelako helburuekin ere erabili da. Besteak beste, HERMES (Verdejo *et al.*, 2003) eta EPEC (Aduriz *et al.*, 2006b) proiektuetako euskarazko corpusak etiketatzeko erabili da.

Aplikazioei dagokionez, *Eihera+* euskararako galdera-erantzun sisteman egindako hurbilpenetan erabili izan da (Alegria *et al.*, 2009). Tresna bera bakarrik erabilgarri egoteaz gain, *IXA talde*ko analisi sintaktikoa egiteko Ixati[19] tresnan integratuta eta erabilgarri dago.

[19] `http://ixa2.si.ehu.es/demo.zatiak.jsp`

Eihera+ tresa *IXA talde*tik at ere erabili izan da. Itzulpen-automatikorako OpenTrad (`http://www.opentrad.org/`) proiektuaren barruan, itzulpen-memoriak publikatzeko egindako NEen garbiketa eta orokortze lanetan erabilgarria izan da. GALAN taldekoek, berriz, ikas-materialetako domeinu identifikaziorako egindako HPko hurbilpenean erabili dute (Larrañaga *et al.*, 2003).

III. KAPITULUA

Entitate-izenen Itzulpena

Pertsona-, toki- eta erakunde-izenak (entitate-izenak, alegia) edozein testu mota zein domeinutan agertu ohi diren espresioak dira, eta azken aldian informazio-erauzketarako zein bilaketa-tresnetarako ezinbesteko informazio-unitate bilakatu dira.

II. kapituluan aurkeztu den bezala, asko dira NEen identifikazio zein sailkapenerako aurkeztu diren lan eta tresnak. Baina, identifikazioa eta sailkapenaz gain, NEen inguruan indarra hartzen ari diren bestelako problemak badira ere, NEen itzulpen-ataza kasu. Izan ere, itzulpen-sistemetan, informazio-erauzketa eleanitzean eta, oro har, edozein sistema eleanitzetan NEen informazio eleanitza oso erabilgarria gertatzen da. Eta informazio hori lortzeko bide bat NEen itzulpena da, hain zuzen ere.

Kontuan hartzen bada testuetan edozein momentutan NE berriak sortu daitezkeela, asko izango dira hiztegi elebidunetan agertuko ez diren espresioak. Horrek hiztegi bilaketa baino prozesu konplexuagoa bihurtzen du NEen itzulpena.

Azken urteetan itzulpen automatikoko sistemak garatzen eta horien hobekuntzetan esfortzu handiak egin dira. NEen itzulpenak, aldiz, ez du tratamendu berezirik jasotzen horrelako sistema gehienetan. Hori dela eta, sistema gehienek, erregeletan oinarritutako itzulpen automatikoko sistemak behinik behin, gaztelerazko *Escuela de Derecho de Harvard* NEa, ingelesez *school of the right of Harvard* itzultzen dute, itzulpen zuzena izango litzatekeen *Harvard Law School* NEa itzuli beharrean (Reeder *et al.*, 2001). Arazo bera gertatzen da estatistiketan oinarritutako itzulpen automatikoko sistemetan,

terminoa aurretik ikasi ez denean behintzat.

Tesi-lan honetan euskara, gaztelera eta ingelesaren arteko NEen itzulpenerako burututako esperimentuak eta eraikitako sistemak aurkezten dira. Lehenik eta behin, existitzen diren NEen itzulpen lanen inguruko errepaso bibliografikoa egingo da, eta, ondoren, lan honetarako erabilitako ingurune esperimentala aurkezten da. Jarraian, diseinatutako esperimentuen deskribapena aurkezten da, eta, azkenik horietan lortutako emaitzak eta ondorioak.

Azpimarratu behar da aurreko kapituluan bezala gure helburua teknika sinpleak erabili eta konbinatzea dela, beti euskara bezalako hizkuntzetan aplikagarriak diren agertokietara egokituak. Hori lortze aldera, kapitulu honetan (eta hurrengoan) metodo ez-gainbegiratuei eman diegu lehentasuna.

III.1 NEen itzulpenerako teknikak

Itzulpen automatikoa aplikazio eleanitz gehienen funtsezko osagaia da. Gaur egun itzulpen automatikoa erabiltzen duten aplikazio errealen artean denbora errealean Web orriak zein posta mezuak itzultzen dituzten aplikazioak daude, besteak beste.

Giza itzulpenerako, zein hiztegietan oinarritutako itzulpen automatikorako sistemak garatzeko, hiztegiak oinarrizko itzulpen-baliabide gisa erabili izan dira. Baliabide horiek ordea, hainbat muga aurkezten dituzte, besteak beste, hiztegitik at geratzen diren hitzen itzulpena burutzeko gaitasun eza. Dalek (2007) bere aurkezpenean adierazten duen bezala, produktu-, pertsona-, toki- eta erakunde-izenen osagaiak izan dira hiztegitik at geratu ohi diren hitzak. Hiztegitik at geratzen diren hitz horiek itzultzeko estrategien artean, transliterazioa asko erabili izan den teknika da. Definizioz transliterazioa, *...prozedura hertsia eta normalizatua da, idazkera-sistema baten grafema bakoitzari beste idazkera-sistema bateko grafema bat edo grafematalde bat esleitzen diona...* (Euskaltzaindia, 2009).

Ataza honetan transliterazioak duen garrantzia dela eta, NEen itzulpentekniken azterketa bi ataletan banatu dugu: transliterazioaren inguruko lanen azterketa batetik; eta NEen itzulpen-sistemak eratzeko erabilitako baliabide motaren araberako sistemen aurkezpena bestetik. Azken horretan corpus paraleloetan, corpus konparagarrietan eta Webean zein Wikipedian oinarritutakoen bereizketa egin dugu.

III.1.1 Transliterazioan oinarritutako NEen itzulpen-sistemak

Transliterazio automatikoa azken hamarkadan itzulpen automatikoaren alorrean NEen eta terminologia teknikoaren itzulpenaren inguruan indarra hartu duen estrategia da, bereziki abiapuntu- eta xede-hizkuntzetan ahoskera antzekoa aurkezten duten osagaien itzulpena burutzeko. Izan ere, estrategia hori NEen itzulpena burutzen duten sistema gehienen oinarrizko osagaietako bat bihurtu da. Karimi *et al.* (2011) egileek transliterazio automatikoaren errepaso bibliografiko sakona egiten dute.

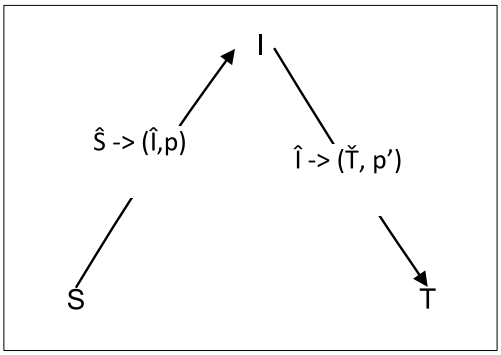
Lan horretan transliterazioa, abiapuntu-hizkuntzan idatzitako hitz bat xede-hizkuntzara transformatzeko prozesua formalki definitzen da. Transliterazio-sistemak nagusiki bi taldetan sailkatzen dira bertan: transliterazio sortzaileak eta transliterazioaren erauzketa burutzen dutenak. Lehen taldeko sistemek, itzulpen-lexikoan agertzen ez diren edo ezezagunak diren osagaien transliterazioa burutzeko algoritmoak deskribatzen dituzte. Transliterazioaren erauzketarako sistemek berriz, corpus eleanitz handietako transliterazio instantziak erabiliz itzulpen-lexikoa aberastea dute helburu.

Transliterazio automatikorako sistemek, sortzaileek zein erauzketarakoek, transliterazio-erregelak sortzeko edota lexikorako bikote berriak sortzeko corpusak edota transliterazioko hiztegiak hizkuntzen ezaugarriekin batera erabili ohi dituzte. Ezaugarriei dagokienez, ezaugarri fonetiko edota ortografikoetan oinarritutako sistemen bereizketa egin daiteke.

Transliterazioa problema fonetikoa bailitz planteatzen dituzten sistemek, hizkuntza desberdinen errepresentazio fonetikoa berdina dela suposatuz, abiapuntu- eta xede-hizkuntzen hitzak transliteratzeko fonetikan oinarritutako hitzen transformazioa aplikatzen dute. Horrela, abiapuntu-hizkuntzako hitza (S) xede-hizkuntzako formara transformatzeko (T), lehenbizi fonetikan oinarritutako transformazioak aplikatuz tarteko espresioa (I) sortzen du, eta ondoren I espresio fonetikoa xede-hizkuntzako forma eskuratzeko erregelak aplikatzen ditu, III.1 irudian agertzen den bezala.

Mota bereko sistemen arteko bereizgarri nagusiak transformazio fonetikorako erregelak ($\hat{\mathrm{S}} \rightarrow \hat{\mathrm{I}}$) eta hizkuntzen fonemak identifikatzeko ($\hat{\mathrm{I}} \rightarrow \check{T}$) estrategiak dira.

Wan eta Verspoorek (1998) ingeles-txinatar hizkuntza-bikoteko toki-izenak itzultzeko sistema aurkezten dute. Sistemak, fonemetan oinarritutako transliterazioa aplikatu aurretik, toki-izeneko osagai bakoitzarekin hiztegi elebidun bat kontsultatzen du eta, itzulpen-hautagairik lortzen ez duenean, osagaiak transliteratzeko estrategia abiarazten du, fonemetan oinarritutako



III.1 Irudia: Ezaugarri fonetikoetan oinarritutako transliterazio-eskema

transliterazioa, hain zuzen ere. Ingeleseko toki-izeneko osagai bakoitza lehenbizi silabetan banatzen dute, ondoren silaba bakoitza txinatarrez ahoskatu ahal izateko azpi-silabetan banatuz. Ingelesezko azpi-silaba bakoitzaren fonema txinatar baliokidea lortzeko sistemak bi urrats ematen ditu: lehenbizi azpi-silaba bakoitza ahoskera txinatarreko alfabeto latinoan errepresentatzeko *Pinyin* idazkera sistemara mapatzen du, eta ondoren alfabeto latinoko idazkeratik idazkera txinatarrera bihurtzen du. Sistemak transformazio guztiak erregelen bitartez burutzen ditu.

Ingeles-txinatar hizkuntza-bikoterako antzeko sistema aurkezten dute Virga eta Khudanpurek (2003) beraien lanean. Lan horretan, ingeles eta txinatar arteko ahoskera transformazioetarako erregelak erabili beharrean, GIZA++[1] (Och eta Ney, 2003) tresnarekin automatikoki erauzitako transformazioak erabiltzen dituzte.

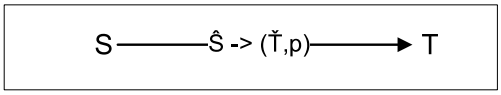
Gao *et al.* (2005) egileen lanean aurkezten den sistemak, aurreko bietan bezala ingeles-txinatar hizkuntzen arteko toki-izenak prozesatzean hiztegi bidez itzulezinak diren izen berezien osagaiak itzultzeko, fonemetan oinarritutako transliterazio estrategia aplikatzen da. Baina lan horretan transformazioak erauzteko adibideetan oinarritutako eredua eraikitzeko algoritmoa aurkezten dute.

Fonetikan oinarritutako sistema hauen desabantaila nagusietako bat hizkuntzen ahoskera deskribatzen dituzten baliabideen menpekotasuna da, bereziki nabarmena dena baliabide hori eskuragarri ez duten hizkuntza-bikoteekin lan egiten denean. Bestetik, transformazio prozesua bi urratsetan emateak,

[1]Itzulpen automatiko estatistikorako tresna honek hizkuntza desberdinetako hitz eta sekuentziak lerrokatzeko ereduak ikasteko algoritmo eta baliabideak eskaintzen ditu.

erroreen hedapenerako arriskua handitu egiten da.

Zailtasun horiek ekiditeko bidean, transformazio ortografikoa bailitz ebazten duen transliterazio-estrategia dago: abiapuntu-hizkuntzako hitzeko (S) karaktere multzoak xede-hizkuntzako hitzeko (T) karaktere multzoetan transformatzeko erregela ortografikoak aplikatzen dira III.2 irudian agertzen den bezala, transliterazioa urrats bakarrean ebatziz.



III.2 Irudia: Ezaugarri ortografikoetan oinarritutako transliterazio-eskema

Transliterazioa ezaugarri ortografikoetan oinarrituz ebazten duten sistemek nagusiki hitzen karaktereetatik erauz daitekeen informazio estatistikoan oinarritzen dira.

Jaleel eta Larkeyek (2003) beraien lanean n-gramak erabiliz ingeles-arabiera hizkuntzen arteko ezaugarri ortografikoetan oinarritutako hitzen transliterazioa aplikatzen duen sistema aurkezten dute. Ikasketa-fasean transliterazioko corpus paralelo elebidun batetik hasiz, GIZA++ tresnarekin ezaugarri linguistikoetan oinarritutako hitzen lerrokatze-eredua erauzten dute. Ondoren, transformazio-erregelak sortzen dituzte eta hauei corpuseko agerpen kopuruan oinarrituz probabilitateak esleitzen dizkiete. Egileek sistemaren portaera ebaluatzeko eskuz sortutako eredu batekin konparatzen dute, biak berrien artikuluetatik eskuratutako 815 hitz-bikoteen gainean aplikatuz. Sistemaren ebaluazio automatikoak % 69,3ko asmatze-tasa lortzen du, eskuzko sistemaren % 71,2tik hurbil agertuz.

NEen informazio ortografikoan oinarritutako pertsona-izenak transliteratzeko sistema berrienetako bat Li *et al.* (2007) egileen lanean aurkezten da. Sistema honek aspektu semantikoak gehitzen ditu transliterazio prozesuan, besteak beste sarrerako izen berezien abiapuntu-hizkuntza, generoa eta izen edo abizenaren informazioa. Esperimentuetan japoniera-txinatarra, txinatarra (alfabeto latinoan idatzitako espresioak) -txinatarra, eta ingeles-txinatarra hizkuntza-bikoteak erabiltzen dituzte, abizenak, emakumeen izenak eta gizonezkoen izenak bereiziz. Lau karaktereetako sekuentziak erabiliz abiapuntu-hizkuntza eta generoa erauzten dute. Informazio semantikoaren aberasketa horrekin guztiarekin autoreek ezaugarri ortografikoetan soilik oinarritutako sistema (Li *et al.*, 2004) hobetzea lortzen badute ere, eta corpusen deskribapen zehatz ezak beste sistemekin zuzenean alderatzea ezinezkoa

egiten badu ere, semantika erabiltzen ez duten antzeko sistemekiko (Zhang *et al.*, 2004) hobekuntza handirik lortzen ez dutela esan daiteke.

Ezaugarri ortografikoetan oinarritutako sistemak, oro har fonetikan oinarritutakoak baino portaera hobea erakutsi duten arren, ohiko ahoskeratik urruntzen diren osagaiak transliteratzeko garaian arazoak dituzte. Esaterako ingeleseko ohiko *gh*, ahoskera desberdina duen *Edinburgh* toki-izena transliteratzeko fonetikan oinarritutako sistemek ortografian oinarritutakoek baino hobeto ebatzi ohi dute. Bi sistemen abantailak aprobetxatzeko asmoz, bi estrategiak nahasten dituzten sistema hibrido eta konbinatuak ere aurki daitezke.

Ildo horretan, arabieratik ingeleserako transliteraziorako, ezaugarri fonetikoetan oinarritutako eredua eta ezaugarri ortografikoetan oinarritutako ereduetako irteerak linealki konbinatzeko saiakerak aurkeztu izan dira ere (Al-Onaizan eta Knight, 2002b, a). Ondoko formulan konbinazio lineala nola burutzen duten zehazten da:

$$P(T|S) = \alpha P_s(T|S) + (1 - \alpha)P_p(T|S)$$

Formulak, T xede-hizkuntzako hitza S abiapuntu-hizkuntzako hitzaren itzulpena izateko probabilitatea kalkulatzen du. P_s probabilitatea ezaugarri ortografikoetan oinarrituta kalkulatzen da, eta P_p, berriz, ezaugarri fonetikoetan. Konbinazio linealerako α faktorea ereduen arteko lehentasunak aldatzeko erabiltzen da.

Sistema hibrido horren portaera soilik pertsona-izenen osagaiak transliteratzeko ebaluatu da, eta eredu fonetikoaren emaitzak % 11,9an hobetzea lortzen bada ere, soilik eredu ortografikoa aplikatzean lortzen diren emaitzekin alderatuz gero % 3,7an okertzen da. Beraz, kasu honetan, eredu ortografikoa egokiagoa dela ematen du.

Hainbat osagairen transliterazioa transformazio fonetikotik hurbilago eta beste batzuk berriz ortografikotik gertuago daudela argudiatuz, gertutasun horretan oinarritutako ereduen konbinazioa burutzen duten lanak ere aurki daitezke (Oh *et al.*, 2006b, a). Bi lan horietan egileek ikasketa automatikoko entropia maximoko eredua, erabaki-zuhaitzak eta memorian oinarritutako ikasketarako algoritmoak konbinatzen dituzte. Proposatutako estrategiak ingeles-korear hizkuntza-bikoterako 6.172 izeneko ikasketa-corpusarekin entrenatutako eta 1.000 bikotekin ebaluatutako sistemak, % 68,4ko asmatzetasa lortzen du. Estrategia bera aplikatuz ingeles-japoniera bikoterako % 62,3ko tasan geratzen da.

Oh eta Isahara (2007) egileek transliterazio-sistema desberdinen irteerako transliterazio-proposamenak berrordenatzeko *SVM*etan eta entropia maximoan (ME) oinarritutako metodoa proposatzen dute beraien lanean. Konkretuki zazpi transliterazio eredu sortzen dituzte, informazio ortografiko eta fonetikoan oinarritutako hiru eredu sinple eta hiru horien konbinazio linealekin (Al-Onaizan eta Knight, 2002b) sortutako beste lau eredu hibrido. Eta zazpi eredu horien irteeren ordenazioa burutzeko eta transliterazio egokiena hautatzeko, *SVM* eta *ME* metodoetan oinarritutako ordenazio-ereduak sortzen dituzte. Azken eredu horiek entrenatzeko zazpi transliterazio-ereduen probabilitateak eta Web kontaketak erabiltzen dituzte. Bi hurbilpenak ingeles-korear eta japoniera-ingelesa hizkuntza-bikoteekin entrenatu eta ebaluatu dituzte: entropia maximoan oinarritutako eredua aplikatzean, ingeles-korear bikoterako % 87,4eko eta japoniera-ingeleserako % 87,5eko asmatze-tasak lortzen dituzte; *SVM* erabiltzean, berriz, % 87,8 eta % 88,2, hurrenez hurren.

Karimik (2008) proposatzen duen ingeles-persiera bikoterako sisteman aldiz, ezaugarri ortografikoetan oinarritutako sistema desberdinak konbinatzen dira, *Naïve Bayes* eta bozketa estrategian oinarritutako konbinazioa erabiliz. Lan honetan sistema konbinatua hizkuntza-bikotearen bi norabideetan ebaluatzen da. Ingeles-persiar bikoterako ingelesetik hasiz ebaluaziorako eskuz transliteratutako 1.500 hitzetako zerrenda erabiliz % 85,5eko asmatze-tasa lortzen dute. Ebaluazio- estrategia bera erabiliz 2.010 persiar pertsona-izen eta beraien ingeles transliterazioko corpusarekin, persiar-ingeles bikoterako sistemak % 69,5era jaisten da asmatze-tasa. Edozein kasutan gehienetan eredu ortografikoak modu independentean aplikatuta lortzen diren emaitzarik onenak gainditzen dira sistemen konbinazioekin.

III.1.2 Baliabideen araberako NET sistemak

NEen itzulpeneko sistemak eratzeko erabilitako corpusei erreparatuz gero, hiru sistema mota bereiz daitezke nagusiki: corpus paraleloetan oinarritutakoak, corpus konparagarrietan oinarritutakoak eta Web-a corpus gisa erabiliz eraikitako sistemak. Hiztegietan oinarritutakoak ere aipa daitezke baina, aurretik esan bezala, horien erabilera hutsa motz geratzen da, eta hori dela eta, ez ditugu sakonki aztertuko txosten honetan.

Arlo honetako lehen ikerketak, itzulpen automatikoko lehen ikerketetan bezala, corpus paraleloetan oinarritzen dira. Corpus paraleloak ordea, ez daude eskuragarri hizkuntza askotarako, edota ezin dira domeinu askotarako

erraz eskuratu, eta NEak asko aldatzen dira domeinutik domeinura. Ondorioz, corpus konparagarrien erabilera indarra hartzen joan da. Eta azken urteetan Web-a corpus gisa erabiltzen duten ikerketa ugari aurki daitezke. Jarraian corpus-mota bakoitzaren ezaugarriak azpimarratu eta horiekin NEen itzulpen arloan egindako ikerketa aipagarrienak azaltzen dira.

III.1.2.1 Corpus Paraleloak

Corpus paraleloa bi hizkuntza edo gehiagotan idatzitako testu sorta bat da, non testu bakoitza beste hizkuntzen itzulpen zehatza den. Itzulpen zehatza izatearen ezaugarri horrek esaldi, hitz edo etiketa mailan lerrokatze automatikorako sistemak eratzeko bidea irekitzen du. Aldi berean, ezaugarri berak baliabide mota hau eraikitzea oso zaila egiten du. Errealitatean, hizkuntzen desberdintasunak direla eta, corpus paralelo gehienak ez dira itzulpen zehatzak izaten. Zarata duten corpus paralelo (*noisy parallel corpora*) izena jasotzen dute, itzulpenak izan arren, bi hizkuntzetako testuak esaldi mailan lerrokatu ezin direnean (Smith *et al.*, 2010).

1997 urtean berriekin osatutako corpus paralelo zaratatsuan oinarritutako izen bereziak itzultzeko lehen sistema aurkezten da (Collier eta Hirakawa, 1997). Japoniera-ingelesa hizkuntza-bikoteko izen bereziak itzultzeko proposatutako sistemak fonetikan oinarritutako erregelak aplikatuz Katakanako hitzen ingelesezko itzulpen posible guztiak sortzen ditu. Azkenik, programazio dinamikoko algoritmoak erabiliz, sistemak abiapuntu- eta xede-bikote guztien artean bi hizkuntzetan fonetikoki antza handiena duen bikoteko ingelesezko hautagaia aukeratzen du japonierako hitzaren itzulpen gisa. Sistemak, lerrokatutako 150 berriekin osatutako corpus batekin ebaluatzean, % 75eko doitasuna eta % 82ko estaldura du.

Huangen (2002) ingeles-txinatar bikoteko izendun entitateak itzultzeko sistemak, arreta berezia eskaintzen die pertsona-, toki- eta erakunde-izenen itzulpenei, entitate-izenen itzulpenei, alegia. Esaldi mailan lerrokatutako corpus paraleloan NERC sistema komertzial bat aplikatuz, esaldi guztietako NEak identifikatu eta sailkatzen ditu. Behin NEak erauzita, Brown *et al.* (1993) egileen ereduan oinarritutako agerkidetza eta itzulpen probabilitateetan oinarrituz itzulpen-hautagaien aukeraketa egiten du. Prozesu iteratibo horrek hasierako adibide sorta batetik hasi, eta ondorengo iterazioetan findutako sorta erabiliz, sekuentzialki corpusean oinarritutako transliterazio hiztegi bat eraikitzen du. NEen itzulpenerako lehenetariko ekarpen interesgarria izan arren, beste hizkuntza batzuetara hedatzea zaila da corpus

paraleloaren beharra dela eta.

Moorek (2003) antzeko lan bat aurkeztu du ingelesa-frantsesa hizkuntza-bikoteko NEak itzultzeko. Sistema NEak erauzteko NERC sistema independente bat erabili beharrean, NEen identifikazioa eta itzulpena ebazteko diseinatu da. Ingelesezko testuak hartuz, hitz hasieren letra larriak kontsideratuz NEak identifikatzeko heuristiko sinplea erabiltzen du, eta, ondoren, teknika estatistikoak erabiliz, abiapuntu-hizkuntzako NEen espresioei xede-hizkuntzako testu lerrokatuetan hitz-sekuentzia baliokideak bilatzen saiatzen da helburu bikoitzarekin: batetik, frantsesez identifikatutako NEak hizkuntza horretako lexiko batean gehituz lerrokatze-sisteman lagungarri izan daitezen, eta bestetik, NE bikotea itzulpenerako hiztegi elebidunean gehituz, baliabidea aberasteko. Ebaluazioa ikasketarekin zerikusirik ez duten software eskuliburuetako 192.711 testuz osatutako corpusarekin aurkezten da, sistemak % 80ko asmatze-tasa duelarik.

Lee *et al.* (2006) egileen lanean ingeles-txinatarreko NEak itzultzeko hiztegi elebidun, corpus paralelo eta transliterazio-lexikoian oinarritutako itzultzaile estatistikoa proposatzen da. Bertan proposatzen den NEen lerrokatze-prozesua Brown *et al.* (1993) egileen itzulpen automatikorako sistema estatistiko ospetsuko lerrokatze-ereduarekin konparatuz, ingeles-txinatarreko NEen lerrokatzean hobekuntza nabarmena lortzen dela erakusten da.

III.1.2.2 Corpus Konparagarriak

Corpus konparagarriak, paraleloak bezala, bi hizkuntza edo gehiagoren testu sortak biltzen dituzten baliabideak dira. Corpus konparagarrietako testuak ordea, ezin dira bata bestearen itzulpen kontsideratu. Testu konparagarriak, hizkuntza desberdinetan gai berdinak eta informazio antzekoa jorratzen dituzten testuak dira (Sproat *et al.*, 2006).

Itzulpen arloan corpus konparagarrietan oinarritutako lehen lanetako bat Tanaka eta Iwasakiren (1996) lana da. Bertan lexikoaren agerkidetzari buruzko informazioa erabiliz hitzen arteko itzulpen baliokidetzak aztertzen dira. Abiapuntu-hizkuntzan elkarrekin agertzen diren hitzak xede-hizkuntzan agerkidetza hori mantentzen dutelako suposizioan oinarritzen dute ikerketa. Horrela izanik, abiapuntu-hizkuntzako hitz bat xedeko hainbatekin erlazionatzeko itzulpen-matrize estokastiko bat erabiltzen dute.

Aurrerago ikusiko dugun gure NEen itzulpen-sistemaren ildotik (ikus III.3 atala), Tao *et al.* (2006) egileen lanean berriekin osatutako corpus konparagarrian oinarritutako sistema aurkezten da. Sistemak gainbegiratu ga-

beko estrategia erabiltzen du, ezaugarri fonetiko zein ezaugarri tenporalak konbinatuz, karaktereen ezabaketa, txertaketa eta ordezkapen konbinazioekin determinatzen duen bi hitzen arteko distantzia kalkulatzen du. Corpus konparagarrian lerrokatzea ezinezkoa denez, dokumentu antzekoak bilatzeko, publikazio-data erabiltzea proposatzen dute. Sistema hiru hizkuntzabikoteetarako ebaluatzen da ingeles-arabiera, ingeles-txinatar eta ingeles-hindi hizkuntzetarako. Hain zuzen ere, % 78,8, % 87,5 eta % 76,1 dira, hurrenez hurren, lortzen dituzten asmatze-tasarik onenak.

Sproat *et al.* (2006) egileek ingelesaren eta txinatarraren arteko NEen itzulpena burutzeko, Tao *et al.* (2006) egileen antzera, ezaugarri fonetiko eta tenporalak erabiltzen dituzte, antzeko testuen agerpen eta agerkidetza maiztasunekin konbinatuz. Maiztasunak eta antzekotasun neurriak kalkulatzeko 234 txinatarrerako testuz eta 322 ingeleseko testuz osatutako corpus konparagarria erabiltzen dute. Antzeko hurbilpen bat aurkezten du Klementiev eta Rothen (2006) lanak ere. Ingeles-errusiera bikoteko NEak itzultzeko hitzen karaktere azpi-sekuentziak lerrokatzeko estrategia proposatzen dute transliterazio hautagaiak erauziz. Lan horretarako errusieraz idatzitako 2.327 eta ingelesezko 978 berrietako testuak erabiltzen dituzte.

III.1.2.3 Web eta Wikipedia

Azken urteetan hizkuntzalaritzan Web-a corpus gisa erabiltzen dituzten ikerketek indarra handia hartu dute. Euskararako ideia horretan oinarritutako Leturia *et al.* (2007) egileen lana aipagarria da. Bertan Web-a euskarazko corpus erraldoi gisa erabiltzea ahalbidetzen duen CorpEus tresna aurkezten da.

Ildo beretik, NEen itzulpen-atazarako ere, Web-a corpus gisa erabiltzen dituzten lanak aurki daitezke. NEen itzulpenean, ez corpus paralelorik ez konparagarririk erabiltzen ez duen lehenbiziko lana 2002 urtean kokatzen da (Al-Onaizan eta Knight, 2002b). Bertan deskribatzen den sistemak, ingelesezko entitate-izenak arabierara itzultzeko, ezaugarri fonetiko eta fonologikoetan oinarritutako erregelak aplikatuz arabierako hautagai zerrendak sortzen ditu. Zerrendak ordenatu eta hautagai egokiena hautatzeko, sistemak agerkidetzak eta testuingurua kontuan hartzen dituzten Web kontaketak burutzen ditu. Sistema entrenatzeko zein ebaluatzeko erabilitako corpusak oso txikiak dira, 21 eta 20 testu hurrenez hurren.

Ildo berean baina Web-a soilik itzulpen egokiaren identifikaziorako beharrean, itzulpen sorkuntza fasean ere erabiltzea proposatzen duten lanak ere

aipa daitezke (Chen eta Chen, 2006; Lam *et al.*, 2007; Jiang *et al.*, 2007), non Web-aren informazioa itzulpenean lagungarri zenbateraino gerta daitekeen erakusten den.

Web-a orokorrean erabili beharrean, Web-eko baliabide zehatzak esplotatzen dituzten lanen artean, Wikipedia entziklopedia da gehien erreferentziatzen den baliabidea. Baliabide hori NEen inguruko atazetan ere erabili izan da azken urteetan. Itzulpen-ataza gisa aurkeztu ez arren, gure lanarekin zerikusia duen ikerketa aurkezten da Sorg eta Cimianoren (2008b) lanean, non ingelesa-alemana hizkuntza-bikoteko Wikipediako sarreren artean existitzen ez diren estekak definitzeko saiakera aurkezten den. Esteka berrien identifikazio prozesuan, hizkuntza desberdinetako Wikipedia sarreren tituluetako hitzen arteko edizio-distantzia erabiltzen da, besteak beste.

Wentland *et al.* (2008) egileen lanean, berriz, NEen baliabide eleanitz bat eraikitzeko saiakeran, prozesuaren urrats batean Wikipediako hizkuntzen arteko estekak erabiltzen dituzte NEak itzultzeko estrategia bezala.

Aipagarria da ere, Pouliquen *et al.* (2006) egileen NEen itzulpen-tresna, NewsExplorer[2] izeneko berriak aztertzeko sisteman integratuta dagoena. Sistema horrek berriez osatutako testu multzo eleanitz batetik hasita, alde batetik pertsona bera erreferentziatzen dituzten aldaerak identifikatzea eta, bestetik, NE desberdinen agerkidetzetan oinarrituz, pertsonen arteko erlazioak inferitzea du helburu.

III.2 Ingurune esperimentala

Aurreko atalean ikusi dugun bezala, NEen itzulpenerako (NET) sistemak eraikitzeko bide eta aukera desberdinak daude. Ondorengo puntuetan euskarazko NET tresna garatzeko aplikatu diren estrategiak eta horietan erabilitako baliabideak aurkezten dira. Konkretuki hiru dira erabili ditugun estrategiak, eta hirurak metodo ez-gainbegiratu eta corpus konparagarrietan oinarritutakoak izan arren, teknika eta baliabide desberdinak erabiltzen dituzte.

Gure lan honetan nagusiki bi sistema mota bereiz daitezke: hizkuntzaren menpeko sistema eta hizkuntzarekiko erdi-independentea. Lehenengoak hizkuntza-bikoteen ezaugarri fonetiko/fonologiko esplizituak erabiltzen ditu itzulpenak egiteko, eta, ondorioz, sistema ez da beste bikote batzuentzat

[2] `http://press.jrc.it/NewsExplorer/entities/en/1.html`

zuzenean hedagarria. Hizkuntzarekiko erdi-independentea den sistemak, berriz, itzulpenak metodo generiko batekin ebazten ditu. Azken hori, egunkarietako berriak zein Wikipedia erabiliz ebaluatu dugu, *hizkuntzekiko erdi-independentea* eta *Wikipedian oinarritutako NET tresna* izenekin aurkeztuko direnak, hurrenez hurren. Jarraian, horien garapenerako beharrezkoak izan diren baliabideen deskribapen zehatza egiten da.

III.2.1 Baliabideak

Euskararako NET tresnek, hirurek, abiapuntu-hizkuntza nagusi gisa euskara izango dute eta xedekoak, berriz, ingelesa eta gaztelera. Sarrera, hortaz, euskarazko testuak izango dira, eta bertan agertzen diren NEak itzultzea ataza honen helburua. Euskal testuetako NEak erauzteko, II. kapituluan azaldutako *Eihera+* tresna erabiliko da, IAko metodoak eta ezaugarri linguistikoetan oinarritutako estrategiak konbinatuz, euskaraz idatzitako sarrerako testuan NEak mugatuta eta sailkatuta itzultzen dituen tresna, alegia.

Esan bezala abiapuntu-hizkuntza nagusia euskara bada ere, hurrengo atalean deskribatzen diren baliabideen eskuragarritasuna dela eta, hizkuntzekiko erdi-independentea den sistema gaztelera abiapuntu-hizkuntza eta ingelesa xede-hizkuntza gisa duen sistema ere diseinatu da, estrategiaren hedagarritasuna ebaluatzeko asmoz, euskara abiapuntu-hizkuntza ez den bikote berrien aurrean estrategiaren portaera aztertu ahal izateko, hain zuzen ere. Hizkuntza-bikote berri horrekin lan egiteak ere zubi-lanak egiten dituen hizkuntza batekin (gaztelera gure kasuan) NET ataza ebazteko sistemen portaera aukera eskaintzen digu. Beraz, euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa izango dira, NET atazan landu diren hizkuntza-bikoteak. Jarraian, hizkuntza-bikote horien NET tresnen garapen zein ebaluaziorako erabilitako baliabideak aurkezten dira.

III.2.1.1 Corpus konparagarriak

Nagusiki bi corpus konparagarri desberdin erabili dira tesi-lan honetan NET tresnen garapenerako: hizkuntza desberdinetako egunkarietako berrietan oinarritutako corpus konparagarria, Hermes corpus eleanitza, hain zuzen ere; eta hizkuntza desberdinetako Wikipediako sarreretan oinarritutako corpusa. Corpus konparagarri horien eraketa prozesua deskribatzen dute ondoko bi puntuek.

III.2.1.1.1 Hermes corpusa

Hermes corpusa bilaketa eleanitz eta erauzketa semantikorako hemeroteka elektronikoak sortzea helburu duen Hermes[3] proiektuaren testuinguruan osatutako corpusa da.

Corpusak 2000 urteko hizkuntza desberdinetako egunkarietako berri etiketatuak biltzen ditu. Urte berdinetako egunkariak erabili diren arren, ez dira bata bestearen itzulpena, hortaz, corpus paraleloa denik ezin da esan, bai, ordea, konparagarri kontsideratu. Izan ere, urte bereko gai antzekoak jorratzen dituzte egunkarietako berrietan NE antzekoak agertzea espero da, bereziki kirola, politika eta nazioarteko gaietako artikuluetan. Eleaniztasuna ustiatuz, bi hizkuntza-bikoterentzat corpusak osatu dira: euskara-gaztelera eta gaztelera-ingelesa bikoteetarako corpusak, hain zuzen ere.

Euskarazko berriak *Euskaldunon Egunkaria*koak dira eta ingeles eta gaztelerako testuak berriz, *EFE*[4] agentziako berrietakoak. Testu horien etiketatze-prozesuan, besteak beste NEak automatikoki etiketatu eta sailkatu dira. Etiketa horietan oinarrituz, hain zuzen ere, eraiki dugu berrietan oinarritutako corpus konparagarria. Corpusen tamainak III.1 taulan laburbildu dira.

	Artikulu	**Hitz**	**NE**
Euskara	40.648	9.655.559	130.505
Gaztelera	16.914	5.192.567	106.473
Ingelesa	16.942	3.631.335	49.768

III.1 Taula: Hermes corpuseko hizkuntza bakoitzeko tamainak

Hizkuntza desberdinen kopuruei erreparatuz, ikus daiteke artikulu kopuruari dagokionez euskarazko corpusa handiena dela. Gaztelerazko eta ingelesezko corpusei dagokienez, artikulu kopuru aldetik oso antzekoak diren arren, gaztelerazkoan NEen kopurua bikoizten da. Kopuru alde hori egon arren, agentzia berdinetik etorrita antzeko NEak agertzea espero da, NEen itzulpen-atazarako baliabide eleanitz oso erabilgarria izanik.

Hizkuntzen araberako banaketa eginez, beraz, bi corpus lortzen dira: euskara-gaztelera NEen corpus konparagarria eta gaztelera-ingelesa corpusa.

[3] http://nlp.uned.es/hermes/

[4] `http://www.efe.com/`

Bi corpus horiek, hainbat sistemen urrats anitzetan helburu desberdinekin erabiliko dira, diseinu atalean (ikus III.3 atala) ikusiko den bezala.

III.2.1.1.2 Wikipedian oinarritutako corpusa

Wikipedia entziklopedia eleanitza web-oinarriduna eta eduki askekoa da, eta bertako sarrerak mundu osoko boluntarioek idazten dituzte. Sarrera bakoitzaren identifikadore unibokoa titulua da. Titulua sarrerak deskribatzen duen entitatearen forma usuena izan ohi da, eta usuenak ez diren entitatearen aldaerak edo formak berbideratze- eta desanbiguazio-orrien bitartez lotzen dira sarrera nagusiarekin.

Wikipedia baliabide eleanitza denez, sarrera bera deskribatzen duten hizkuntza desberdinetako artikuluak topa daitezke. Esate baterako, euskarazko *Euskal Herria* eta ingelesezko *Basque Country* sarrerek NE bera deskribatzen dute hizkuntza desberdinetako Wikipedietan. Sarrera horien arteko estekak ezartzeko asmoz, Wikipediak hizkuntza-arteko estekak (*Wikipedia Interlanguage Link*, WIL) definitzen ditu.

WILak ustiatuz, beraz, NEen itzulpenak lor daitezke inolako itzulpen-prozesu konplexuagorik aplikatu gabe. Zoritxarrez NE guztiek ez dute Wikipedian sarrera definituta eta are gutxiago WIL estekarik. Hainbat sarreren kasuan, hizkuntza desberdinetan sarrera baliokidea existitu arren, WIL esteka definitu gabe dute. Hortaz, metodo egokia dirudien arren, WILen bitartez itzulpenak lortzea ez da nahikoa NET tresna sendo bat eraikitzeko. Bai, ordea, itzulpen-sistema baten urrats bezala erabiltzeko, eta hori da, hain zuzen ere, tesi-lan honetako Wikipedian[5] oinarritutako tresna garatzean egiten dena (ikus III.3.3 atala).

WIL estekez gain, Wikipedia corpus konparagarri gisa ere erabil daiteke, hizkuntza desberdinetako Wikipediako sarrerez osatutako corpusa osatuz. Azken batean, antzeko sarrerak eta sarrera berak hizkuntza desberdinetan deskribatzen dituzten testuak konparagarri kontsidera daitezke.

Wikipedian oinarritutako NEen corpus konparagarria eraikitzeko, *Yahoo!*k semantikoki etiketatutako ingelesezko Wikipedia bertsiotik[6] hasi gara. Wikipediako bertsio hau izan da eta ez beste bat, ataza honetan lanean hastean, beste hainbat anotazioren artean, titulu zein deskribapen testuetako NEak (automatikoki) etiketatuta zituen bertsio bakarra baitzen.

[5] Wikipedian eta bere esteken bitartez bilaketak egiteko http://www.mediawiki.org/wiki/API APIa erabili da.

[6] http://barcelona.research.yahoo.net/dokuwiki

Tesi-lan honetako NET estrategia guztietan abiapuntu-hizkuntza nagusia euskara denez, ingeleseko NEen euskarazko NE baliokideak lortzea da helburua. Lehen urratsa beraz, ingelesez etiketatutako Wikipedia horretako 162.720 NEak erauztea izan da. Ondoren, NE sorta horretatik euskarazko Wikipedia bertsioarekin lotzen dituzten WIL estekak jarraituz euskarazko NE baliokideak lortu dira, euskara-ingeles bikoterako 6.193 NE zehazki. Esan bezala abiapuntuko NE askok ez dute euskarazko Wikipediarekin lotzen duten WIL estekarik definituta, ingelesezko NE adina bikote ez lortzea eragin duena.

NE bikote horietan NE berari erreferentzia egin arren bi hizkuntzetan forma desberdina erabiltzen dituzten NE bikoteak topa daitezke. Esate baterako, euskarazko *Tourmalet* sarrera ingelesezko *Col du Tourmalet* sarrerakin erlazionatzen du WIL estekak, euskarazko bertsioan *Tourmaleteko lepoa* sarrera existitzen den arren.

WIL esteken bidez erauzitako bikote horiek, NE bera erreferentziatu arren, ezin dira NEen itzulpen baliokide kontsideratu eta, hortaz, benetan baliokide diren beste bikoteetatik banatu behar dira. Banaketa-prozesua automatizatzeko asmoz, abiapuntu- zein xede-hizkuntzetako NEen osagai kopuruan oinarritutako irizpidea aplikatu da, hau da, luzera bereko NE bikoteak multzo batean eta gainerakoak beste zerrenda batean gorde dira. Banaketaren ondoren, batetik osagai kopuru bereko 4.859 NE bikote eta bestetik, osagai kopuru desberdineko 1.334 bikote lortu dira, bi multzo osatuz.

Wikipedian oinarrituta, beraz, euskara-ingelesa hizkuntzetarako baliabidea osatu da, orokor daitekeen metodo bat erabiliz. Prozesu bera jarraituz, alegia, *Yahoo!*k semantikoki etiketatutako ingelesezko Wikipediatik hasita, edozein beste hizkuntzetara zuzendutako WILak esplotatuz, ingelesa eta beste hizkuntza horren bikoterako sortu liteke guk eratutako corpus konparagarriaren pareko baliabidea.

III.2.1.2 Hiztegi elebidunak

Edozein itzulpen-sistema automatikotan bezala, kapitulu honetan landutako NET itzulpen-sistemetan ere hiztegi elebidunak erabili dira, transformazio fonetiko/fonologikoekin itzuli ezin daitezkeen osagaien itzulpenak eskuratu ahal izateko, behinik behin. Hiru izanik landutako hizkuntza-bikoteak, hiru dira erabilitako hiztegi elebidunak: euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa hiztegi elebidunak, hain zuzen ere. Hiru hiztegietatik, hitz-bikoteen itzulpen-zerrenda bana eskuratu da. Izan ere, NET tresnen diseinu

atalean ikusiko den bezala, hiztegien atzipena azkartzeko asmoz, hiztegiak automata bihurtu ditugu, eta bihurketa horretarako 1:1 korrespondentzia beharrezkoa da. Hortaz, hiru zerrenda elebidun, bikoteko bana, erabiltzen dira NET tresnetan hiztegi elebidun gisa.

Existitzen diren euskara-gaztelera hiztegi elebidunen artean, lan honetarako *Elhuyar Fundazioa*ren *Elhuyar 2000*[7] hiztegia erabili da, proiektuaren hasieratik *Ixa Taldeak* formatu digitalean eskuragarri izan duen hiztegia. Hiztegi elebidun horretan euskarazko sarrera bakoitzeko gaztelerazko n itzulpen/definizio eskuragarri daude, zehazki 99.516 euskarazko sarrera daude definituta. Hizkuntzekin menpekoa den NET sisteman hiztegi elebiduneko itzulpen guztiak erabiltzen badira ere, hizkuntzekin erdi-independentea den sisteman hiztegia automata bezala erabiltzen da. Eta hiztegia automata bihurtzeko bide horretan, hiztegitik euskara-gaztelera hitz-bikoteak sortu dira, euskarako sarrera bakoitzean definitutako gaztelerako itzulpen-forma adina bikote sortuz. Bikote-sorkuntza prozesu horretan, euskarazko osagai sinpleak, alegia hitz bakarreko sarrerak hautatu dira eta, itzulpenei dagokienez, luzera berdina edo oso antzekoa duten itzulpenak besterik ez dira erabili. Bi irizpide horiek jarraituz, 74.330 euskara-gaztelera hitz-bikoteko baliabidea automatikoki sortu dugu.

Euskara-gaztelera itzulpenean ordea, *Elhuyar Fundazioa*ren hiztegi elebiduna ez da izan erabili den hiztegi bakarra. Izan ere, NEetan eta oro har hitz-elkarketetan badira hitz berezi batzuk, euskaraz izen kategoriakoak izan arren, gazteleraz adjektibo kategoria betetzen dutenak. Esaterako, euskarazko *Itsas Armada* NEan *Itsas* hitzak izen kategoria du, eta gaztelerazko *Ejército Marítimo* itzulpenean berriz, *Marítimo* hitzak adjektibo funtzioa du. Horrelako kasuak ustiatzen dituen hiztegirik erabili ezean, aurrerago III.3.1.2 atalean azalduko dugun bezala, hizkuntzen menpeko NET sistemak ezingo du horrelako NEen itzulpen egokia osatu, eraketa hori hitzen kategorian oinarrituta baitago.

Kategoria aldaketa jasaten duten hitz berezi horien zerrenda osatzeko, *Elhuyar Fundazioa*ren gaztelera-euskara hiztegi elebidunetik gaztelerazko adjektiboen zerrenda bat eskuratu dugu. Zerrenda horretatik eskuz bereizi ditugu euskaraz hitz elkartu gisa itzultzen diren adjektiboak eta horien euskarazko izenak eskuratu ditugu 349 bikoteko hiztegi bat osatuz. Hiztegi hori da, hain zuzen ere, aurrerantzean **hitz berezien hiztegi elebidun** bezala aipatuko den hiztegia.

[7] `http://www.elhuyar.org/hizkuntza-zerbitzuak/eu/Elhuyar-Hiztegia`

Euskara-ingelesa hiztegi elebidunari dagokionez, OpenTrad[8] proiektuan erabilitako ingelesa-euskara hiztegi elebidunetik hasi gara. Hiztegi horretan, 24.047 ingeleseko sarreretarako euskarazko hainbat itzulpen posible definituta daude. Aurreko hiztegian bezala, hitz-itzulpen bikoteak eratu ditugu, antzeko luzerako ingeles-euskara bikoteak sortuz, soilik osagai bakarreko euskarazko itzulpenak erabiliz. Horrela, bikoteei buelta emanez, 41.854 euskara-ingeles hitz-bikoteko zerrenda lortu dugu.

Azkenik, gaztelera-ingeles bikoterako erabilitako hiztegi elebiduna MCR (*Multilingual Central Repository*)[9] izan da. Aurreko bietan erabilitako irizpide berdinak erabiliz, 73.784 gaztelera-ingeles hitz-bikote erauzi ditugu.

Lau hiztegietatik erauzitako zerrenda elebidunen tamainak III.2 taulan laburbiltzen dira, eta bertan ikus daitekeen bezala, euskara-gaztelera eta gaztelera-ingeleserako ohiko hiztegi elebidunetatik lortutako baliabideak tamainan antzekoak badira ere, euskara-ingelesaren kasuan, bikote kopurua nahiko txikiagoa da. Guztien artean txikiena, hitz berezien hiztegia da, izan ere oso portaera berezia duten hitzen zerrenda besterik ez baitu biltzen.

	Bikote-kopurua
euskara-gaztelera	74.330
euskara-gaztelera (hitz bereziak)	349
euskara-ingelesa	41.854
gaztelera-ingelesa	73.784

III.2 Taula: Hiztegi elebidunen tamainak

III.2.1.3 Ebaluazio-corpusak

Hiztegien kasuan bezala, NET tresnak hizkuntza-bikote desberdinekin ebaluatu ahal izateko, bikote bakoitzeko gutxienez ebaluazio-corpus bat osatu behar da. Gutxienez diogu, tresnen diseinuaren atalean (ikus III.3 atala) ikusiko den bezala, tresnak baliabide eta domeinu desberdinetan oinarrituz garatu baitira. Hortaz, estrategiaren portaera ondo neurtzeko asmoz, tresna bakoitza ebaluatzeko hizkuntzak kontuan izateaz gain, domeinua ere kontsideratu behar da. Bi irizpide horien konbinaketatik, ondoko ebaluazio-corpusak definitu ditugu:

[8] `http://www.opentrad.org/`

[9] `http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=53`

- *Elhuyar*en onomastikan oinarritutako euskara-gaztelera ebaluazio-corpusa.
- Hermes corpusean oinarritutako euskara-gaztelera eta gaztelera-ingelesa ebaluazio-corpusak.
- Wikipediako WIL esteketan oinarritutako euskara-ingelesa ebaluazio-corpusa.
- 2009 urteko CLEF ResPubliQA[10] galderen itzulpen sortan oinarritutako euskara-ingelesa ebaluazio-corpusa.

III.2.1.3.1 Onomastika (euskara-gaztelera)

*Elhuyar*en onomastika, *Elhuyar Fundazioa*k luzatutako izen berezien zerrenda bat da. Bertan, pertsona-, toki- eta erakunde-izenen euskara, gaztelera eta frantses aldaerak biltzen dira. Zerrendak 1.570 sarrera dituen arren, itzulpenak proposatzeko garaian formatu erregularrik jarraitzen ez duenez, ezin izan da prozesu automatiko bat definitu sarrera bakoitzetik euskara-gaztelera NE bikote bana eskuratzeko. Automatizazioa ezinekoa izanik, eskuzko berrikuspen baten beharraren aurrean zerrenda horretatik zoriz 600 sarrera hautatu dira eta, berrikuspena eta gero, 557 NE bikote eskuratu dira. Euskara-gaztelerazko 557 bikote horiek dira, hain zuzen ere, *Elhuyar*en onomastikan oinarritutako lehen corpusa osatzen duten NEak. Ebaluazio-corpus hau, emaitzen atalean ikusiko dugun bezala (ikus III.4 atala), hizkuntzen menpeko NET tresna ebaluatzeko erabili da.

III.2.1.3.2 Hermes (euskara-gaztelera eta gaztelera-ingelesa)

Hermes corpusean oinarritutako euskara-gaztelera eta gaztelera-ingelesa ebaluazio-corpusak eratzeko ere, eskuzko lana egin behar izan dugu. Tresnak trebatzeko corpus konparagarrien atalean (ikus III.2.1.1 atala) aztertutako 2000 urteko berriak erabili direnez, ebaluaziorako corpusa osatzeko gai antzekoak baina urte desberdinekoak diren 2002 urteko berriak erabili ditugu. Atal horretan azaldu den bezala, Hermes copuseko berrietan, NEak mugatuta daude. Eta abiapuntu- zein xede-hizkuntzetako testuetan NEak mugatuta egon arren, beraien arteko korrespondentziarik definituta ez dagoenez, ezin dugu

[10] *Cross Language Evaluation Forum* konferentziako galdera-erantzun sistema eleanitzen ebaluaziorako ataza partekatua

jakin zein NE den zeinen itzulpen. Horregatik izan da beharrezkoa eskuzko lana. Zehazki, corpusaren osaketaren prozesua ondokoa izan da: hizkuntza-bikoteko abiapuntu-hizkuntzen berrietako NEen agerpen-kopuruak kontatu ditugu eta bakoitzeko maiztasun handieneko 200 NEak hautatu ditugu; jarraian, NE bakoitzerako eskuzko itzulpena egin da, hala, 200 NE bikoteko euskara-gaztelera eta gaztelera-ingelesa corpusak osatuz.

III.2.1.3.3 Hermes + WIL (euskara-ingelesa)

Eskuzko lana ahal den heinean ekiditeko eta Wikipediako domeinuko ebaluazio-corpusaren eraketa automatizatzeko bidean, Hermeseko *Euskaldunon Egunkaria*ko 2002 urteko berrietan etiketatutako NEak eta Wikipediako WILak konbinatu ditugu. Abiapuntua euskarazko berrietan agertzen diren 142.464 NEak izan dira. Horietatik guztietatik maiztasun handieneko 1.000 NEak hautatu ditugu, aurreko corpus eraketan bezala. Oraingoan, eskuzko itzulpena Wikipediako WILen interpretazio automatikoarekin ordezkatu dugu, ordea. Horrela, ingelesa xede-hizkuntza izanik, guztiz automatikoki lortzen ditugu sarrerako euskarazko NEen ingeleseko formak. Prozesu automatiko horretatik euskara-ingeleseko 680 NE bikote eskuratu ditugu eta, garapenerako corpusean egin den bezala, NE multzo hori bitan banatu da NE bikoteen osagai kopurua banaketa irizpide izanik: osagai kopuru bereko 575 NE bikoteko eta osagai kopuru desberdineko 105 NEko zerrendak lortuz. Lan honetan, sistemaren diseinu atalean ikusiko dugun bezala (ikus III.3 atala), itzulpena ebazteko definitu dugun estrategian abiapuntu- eta xede-hizkuntzan luzera desberdinak dituzten NEei ez diegu arreta berezirik eskaini. Hori dela eta, azken ebaluazio-corpus hori, etorkizunean sisteman bikote mota horien tratamendua gehitzean erabiliko da, baina ez kapitulu honetako emaitzetan.

III.2.1.3.4 CLEF ResPubliQA (euskara-ingelesa)

Wikipedian oinarritutako sistema ebaluatzeko ordea, ez da azken ebaluazio-corpus hori izan erabili den bakarra. Bigarren ebaluazio-corpus bat garatzeko beharrean egon gara, aurrerago diseinu atalean (ikus III.3.3 atala) ikusiko dugun bezala, tresna horren parte baitira WIL estekak. Hortaz, WIL esteketan oinarritutako ebaluazio-corpusa, sistema bere osotasunean ebaluatzeko aproposa denik ezin da esan. Hori dela eta, WIL estekak erabili gabe sistemaren portaera aurreko ebaluazio-corpusarekin ebaluatu dugu, eta sistema osoaren portaera berriz, jarraian azaltzen dugun corpusarekin. Gainera, ebaluazio-

corpus honen eraketan Wikipedia erabili ez izanak, corpuseko NE bikoteen artean xede-hizkuntzako formak xedeko Wikipedian sarrerarik ez duten NEak ere agertu ahal izatea dakar, horrelako egoeretan sistemaren sendotasuna eta portaera neurtu eta aztertzeko aukera eskainiz.

Wikipediatik independentea den ebaluazio-corpus hau osatzeko, 2009ko CLEF ResPubliQA atazarako prestatutako hizkuntza desberdinetako 500 galderen itzulpenak erabili ditugu. Euskara eta ingeleseko galderetatik eskuzko berrikuspenaren ostean 72 NE bikote identifikatu dira. 72 bikote horietako 9 NEen ingelesezko itzulpenek ingelesezko Wikipedian ez dute sarrerarik.

Laburbilduz, beraz, euskara-gaztelera eta euskara-ingeleserako bi ebaluazio-corpus sortu ditugu eta, gaztelera-ingelesa bikoterako bakarra. Ebaluazio-corpus horien jatorri eta tamainak III.3 taulan laburbiltzen dira.

Jatorria	**Hizkuntza-bikotea**	**NE kopurua**
Onomastika (III.2.1.3.1)	euskara-gaztelera	557
Hermes (III.2.1.3.2)	euskara-gaztelera	200
Hermes (III.2.1.3.2)	gaztelera-ingelesa	200
Hermes (III.2.1.3.3) + WIL	euskara-ingelesa	575
CLEF ResPubliQA (III.2.1.3.4)	euskara-ingelesa	72

III.3 Taula: Ebaluazio-corpusen jatorria eta tamainak

III.2.2 Ebaluazio-neurriak

NET sistema bat ebaluatzeko garaian hainbat ikuspegi kontuan hartu behar dira. Alde batetik sistemak itzultzen dituen NEen zuzentasuna neurtzeko doitasun metrika ezaguna erabiliko da, NET sistemak itzultzen dituen NEen artean itzulpen egokia zenbatetan lortzen duen neurtzen duena.

$$doitasuna = \frac{zuzen\ itzulitako\ NE\ kopurua}{itzulitako\ NE\ guztien\ kopurua}$$

Sistemaren itzulpen estaldura beste faktore garrantzitsu bat da, non edozein NEen aurrean sistemak itzulpen bat eskuratzeko ahalmena neurtzen den:

$$estaldura = \frac{zuzen\ itzulitako\ NE\ kopurua}{itzuli\ beharreko\ NEen\ kopurua}$$

Doitasun eta estaldura, biak konbinatzeko F_1 neurri ezaguna erabiliko da.

$$F_1 = \frac{2*Doitasuna*Estaldura}{Doitasuna+Estaldura}$$

Sistemak ebaluatzeko garaian, beraz, bakoitzaren doitasuna, estaldura eta F_1 neurriak erabiliz aurkeztuko dira emaitzak.

III.3 Euskarazko NET sistemaren diseinua

Sarreran euskarazko NE bat jasoz, xede-hizkuntzako NEaren forma edo itzulpena eskuratzea helburu duen NET tresna garatzeko, aurretik aipatu den bezala, bi bide desberdin jorratu ditugu. Lehenengoak itzulpen-prozesua garatzeko hizkuntza-bikote jakin bateko ezaugarri fonetiko/fonologiko eta sintaktiko zehatzak erabiltzen ditu. Estrategia hori jarraituz diseinatu dugu lehen euskara-gaztelera NET tresna. Bigarren estrategiak, itzulpen-irizpideak corpus konparagarrietatik informazioa erauzten du itzulpen-prozesua edozein hizkuntza-bikoterako orokortuz. Azken hurbilpen horrekin, desberdintasun handiena erabiltzen duten corpusa izanik, bi sistema desberdin diseinatu ditugu: egunkarietako berriekin osatutako corpus konparagarrian oinarritzen den sistema eta Wikipedia entziklopedia corpus konparagarri gisa ustiatuz eraikitako sistema.

Hiru sistemetan sarrera abiapuntu-hizkuntzako NE lematizatua/etiketatua izango da. NEa NET sistemara bidali aurretik, beraz, lematizatu egin beharko da. Izan ere, lematizazio prozesua, itzulpen-prozesu baten aurreprozesu kontsidera daiteke, urrats hori NET sistematik at utziz. Aurreko kapituluan bezala, euskarazko NEak lematizatzeko, *Eustagger* tresna erabili dugu, lemaz gain, osagai bakoitzeko kategoria/azpikategoria, kasua eta pluraltasun informazioa ere eskuratuz, besteak beste.

Erabilitako estrategiak desberdinak izan arren, sistemaren helburua bera izanik, hiruetan sekuentzialki aplikatzen diren modulu nagusiak, funtzionaltasunaren ikuspuntutik behinik behin, komunak dira:

- NEen osagaiz osagaiko itzulpenerako modulua
- NE osoaren itzulpen-eraketarako modulua
- NEen itzulpen-hautagaien aukeraketarako modulua

Lehenbizikoa, NEen osagaiz osagaiko itzulpen-modulua, NEaren osagaien banakako itzulpenaz arduratzen da, abiapuntu-hizkuntzako NEaren osagai bakoitzeko xede-hizkuntzan itzulpenak sortuz. Izan ere, NE bat osagai bat

baino gehiagoko espresioa izan daiteke, eta beraz, NE osoaren itzulpena eskuratzeko, osagai bakoitza xede-hizkuntzan nola adierazten den jakitea beharrezkoa da NEaren itzulpen osoa eskuratu ahal izateko.

Bigarren moduluan, behin osagai bakoitzeko itzulpen-hautagaiak eskuratu direla, sistemak NE osoaren itzulpen-hautagaiak eratzen ditu osagaietako itzulpen-hautagai horiek konbinatuz eta ordenatuz. Ordenazioa beharrezkoa gertatzen da, batez ere egitura sintaktiko desberdineko bikoteak erabiltzean. Horrelakoetan, oso ohikoa da NEaren osagaiak bi hizkuntzetako espresioetan toki berean ez agertzea, *Ipar Poloa -> Polo Norte* kasu. Osagaien ordenazio tratamendurik gabe, askotan NEen itzulpen egokia lortzea ezinezkoa izango litzateke.

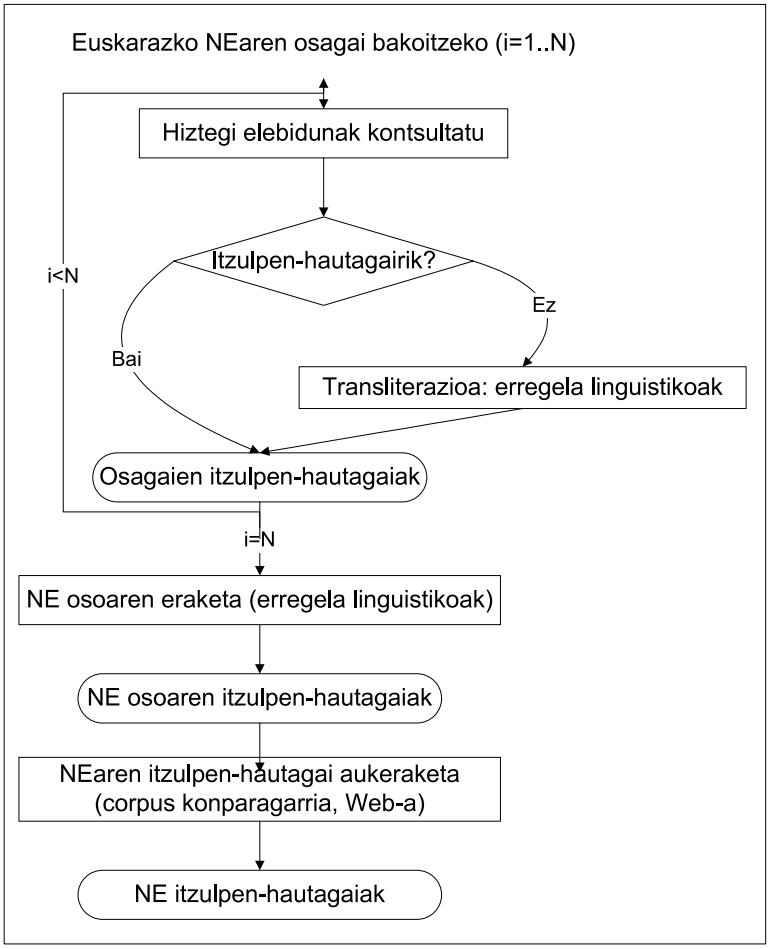
Azkenik, NEen itzulpen-hautagaien aukeraketa moduluan, aurreko moduluan sortutako hautagai guztien artean egokiena zein den ebazten da.

Sistemako modulu nagusien funtzionaltasunak komunak badira ere, modulu horien diseinu eta inplementazioa hiru tresnetan teknika desberdinekin gauzatu dira. Desberdintasun horiek ondo azaltzeko asmoz, tresna bakoitzaren diseinua banan-banan aztertuko da. Lehenbizi hizkuntzen menpekoa den tresna aurkeztuko da. Ondoren hizkuntzekiko erdi-independentea den tresna azalduko da eta, bukatzeko, azken horren estrategia oso antzekoa jarraitzen duen Wikipedian oinarritutako NET tresnaren diseinua azalduko da.

III.3.1 Hizkuntzen menpeko NET tresna

Euskara-gaztelera hizkuntza-bikoterako sistema hau, euskarazko NE bat jasoz gaztelerako itzulpen-hautagai zerrenda bat bueltatzeko diseinatu da. Kontuan hartuz sarrerako NEa osagai bakarrekoa zein anitzekoa izan daitekeela, sistemak lehenik eta behin osagaiz osagaiko itzulpena burutuko du, eta behin osagai guztien bakarkako itzulpenak eginda, horiek konbinatuz NE osoaren itzulpena eraikiko du. Azkenik, hautagai guztien artean itzulpen egokiena aukeratzeko asmoz Web bilaketak egingo ditu besteak beste. Urrats hauek guztiak III.3 irudiko eskeman laburbilduta ikus daitezke.

Lehenengo urratsetan osagaiz osagaiko itzulpena egiteko hiztegi elebidunak eta transformazio fonetiko/fonologikoak simulatzeko erregela linguistikoak erabiltzen ditu. Hiztegi elebidunetik itzulpenik lortzen ez duenean soilik, euskara-gaztelera bikoteko hitzen arteko aldaera fonetiko/fonologikoen eskuzko azterketa batetik ondorioztatutako erregela linguistikoak aplikatuko ditu.



III.3 Irudia: Hizkuntzen menpeko NET tresnaren diseinua

Behin NEeko osagai guztien itzulpen-hautagaiak izanik, hautagaiak konbinatzeari ekingo dio NE osoaren itzulpen-proposamenak eratzeko. NE osoaren eraketa horretan, osagaien ordenazioa burutzen da, euskara-gaztelera bikotearen izaera sintaktikoen desberdintasunak kontsideratuz, osagai bakoitzeko itzulpenak, xede-hizkuntzako espresioan zein posiziotan kokatu behar diren ebatziz.

Azkenik, aurreko urratsean sortutako NE osoaren itzulpen guztietatik sarrerako NEarentzako itzulpen zuzenena zein den hautatzeko urratsera jotzen du sistemak. Bertan, III.3.1.3 puntuan sakonki azalduko den bezala, baliabide eta estrategia desberdinak erabiltzen dira: aurreko urratsetan sortutako

osagaien tarteko pisuak eta Google edota Wikipedia bezalako kanpoko baliabideak. Biak konbinatuz, pisu altuena lortzen duen hautagaia izango da sistemak itzulpen gisa emango duena.

Ondorengo puntuetan fase bakoitza zehazki nola burutzen den azaltzen da.

III.3.1.1 NEen osagaiz osagaiko itzulpen-modulua

Modulu honen helburua NEa osatzen duten osagai guztietarako itzulpen-hautagaiak aurkitzea da. Osagaiz osagaiko itzulpena burutzeko, bi urrats sekuentzial aplikatzen dira, hiztegi elebidunaren kontsulta eta transliterazioa, hain zuzen ere. Bietan, sarrerak euskarazko NEaren osagaien lemak izango dira, aurretik esan den bezala. Ondorengo puntuetan, beraz, osagaia esatean osagaiaren lema ulertu behar da.

III.3.1.1.1 Hiztegi-kontsulta

Hizkuntzen menpekoa den NET tresna honetan kontsultarako erabiltzen diren euskara-gaztelera hiztegi elebidunak bi dira: euskaraz izen eta gazteleraz adjektibo diren hitz berezien hiztegi elebiduna batetik; eta ohiko hiztegi elebiduna bestetik. Bi hiztegiak baliabideen atalean azaldu dira (ikus III.2.1.2 atala).

Sistemak jasotzen duen osagaia izen arrunt motakoa bada hitz bereziak kudeatzeko hiztegira joko du lehenbizi. Bertatik itzulpenen bat eskuratzen badu, itzulpena gordetzeaz gain sistemak hitz horren kategoria izenetik adjektibo kategoriara bihurtzen du. Hitz berezien hiztegitik hautagairik lortzen ez duenean edota sarrerako osagaia izena ez den beste kategoria batekoa denean, sistemak hiztegi elebidun arruntera jotzen du.

Oro har kontsulta horietatik itzulpen-hautagaiak (bat baino gehiago izan daitezke) lortzea gerta daiteke, baina kontrakoa ere, arrazoi nagusiak bi izan daitezkeelarik:

1. Hiztegietan hitz horri dagokion sarrerarik ez egotea.

2. Hitz horri dagokion itzulpen/definizioa baliagarria ez izatea. Esaterako, *aholkulari* sarrerarentzat *persona que aconseja* bezalako itzulpen baino definizio itxurako proposamenak izatea.

Lehenengo kasua identifikatzea erraza da, hiztegi atzipenen emaitza hutsa

izango baita bi hiztegietan, alegia hiztegien kontsultatik ez da hautagairik lortuko.

Bigarren kasua, berriz, identifikatzen zailagoa bada ere, soilik ohiko hiztegi elebidunaren atzipenean eman daiteke. Izan ere, izen berezien kudeaketarako zerrenda, aldez aurretik eskuz landu da eta itzulpen okerrak baztertu eta baliozko hitz bakarreko itzulpenak besterik ez baitira gorde. Beraz, zerrenda horretan itzulpen-hautagairik definituta badago, baliozko hautagaia izango da beti. Ohiko hiztegi elebidunaren gainean eskuzko berrikuspen hori egin ez denez, aldiz, bertatik lortzen diren osagaien itzulpenen zuzentasuna ezin da bermatu.

Zuzentasunean ziurtasun falta horrek eragin ditzakeen erroreak minimizatzeko asmoz, baina eskuzko berrikuspena ekidinez, ohiko hiztegi elebidunetik eskuratutako hautagaien baliagarritasuna kontrolatzeko heuristiko bat definitu da. Heuristikoan erabiltzen diren irizpideak ondorengoak dira:

- 2 hitzez osatutako itzulpena bada eta bi osagaietako bat artikulua ez bada, itzulpen-hautagaia baztertu egiten da.
- 2 hitz baino itzulpen luzeagoa bada, itzulpen-hautagaia baztertzen da (definizio motako itzulpenak baztertuz).

Heuristikoak beraz, irizpide horiek betetzen dituzten hautagaiak, zerrendatik ezabatzen ditu. Ezabaketa prozesua, ez da aurreprozesu gisa egin, baizik eta itzulpen-prozesuko urrats bat bezala gehitu da banakako osagaien itzulpen-moduluan. Hori dela eta, hiztegi elebidunetik hitz bati dagozkion itzulpenak lortu ondoren aplikatzen den garbiketa urratsa denez, osagai baterako hiztegi-kontsultatik eskuratutako hautagaien kopurua jaistea eragin dezake. Are gehiago, hiztegiko kontsultatik itzulpenak eskuratu arren, garbiketa prozesuaren ondoren, euskarazko NEaren osagaia gaztelerako hautagairik ez izatera pasa daiteke, hautagai guztiek ezabaketa irizpideren bat betetzearen ondorioz.

Itzulpenen bat geratzen denean, sistemak NE osoaren eraketa ondo burutu ahal izateko osagai mailako genero eta pluraltasunari buruzko azterketa egitea beharrezkoa da. Izan ere, hiztegira lemarekin jotzean xedehizkuntzako hautagaiaren pluraltasun marka galdu egiten da. Eta bestetik, euskaraz orokorrean genero-markarik ez dagoenez, gaztelerako itzulpena proposatzeko garaian hizkuntza horretan generoa dagoela kontuan hartzea ezinbestekoa da, maskulino zein femenino formak aukeren artean emateko. Demagun euskarazko *Nazio Batuak* NEa gaztelerara itzuli nahi dela

eta hiztegitik *nación* eta *unida* itzulpen-hautagaiak eskuratzen direla hurrenez hurren. *Batuak* osagaia, euskarazko NEaren forman plural marka duela kontsideratzen ez bada gaztelerazko itzulpen zuzenik ez da lortuko. *Eustaggeren* analisitik, beraz, lemaz gain pluraltasun markari buruzko informazioa ere erabiltzen du sistemak. Horrela, sistemak euskarazko osagaia pluralean dagoenean xede-hizkuntzako hautagaian plural marka gehituko du. Pluraltasunaren tratamendu berezi hori, soilik izen arrunt eta adjektiboei aplikatzen zaie, bestelako osagaietan pluraltasun marka hori lexikotik etorriko dela suposatzen baita.

Generoari dagokionez, euskarak genero morfologikoaren markarik ez duenez euskarazko hainbat hitz gaztelerazko bi hitzi egokitu dakieke. Adibidez, gaztelerazko *secretario* eta *secretaria* idazteko euskaraz *idazkari* hitza besterik ez da erabiltzen. Hiztegiek *secretario* forma besterik ez dutenez proposatzen itzulpen bezala, sistemak *secretaria* forma sortzeko beharra du, hainbat NEek osagaien forma femeninoa eta ez maskulinoa eskatzen baitute. Esaterako, euskarazko *Janet Reno Idazkaria* itzultzean, gaztelerazko forma egokia *Secretaria Janet Reno* da eta ez *Secretario Janet Reno*; genero tratamendua egin ezean sistemak inoiz lortuko ez duen itzulpena. Sisteman genero sorkuntza hori, itzulpen-hautagai bat *oz* bukatzen denean hautagai bera *a* bukaerarekin sortu eta hautagai zerrendan gehituz ebazten da.

Osagaiaren hautagai zerrendan bi sorkuntzak/moldaketak aplikatuta, beraz, pluraltasun eta genero markak ebatzita geratzen dira hurrengo urratsetarako.

III.3.1.1.2 Transliterazio-gramatika

Edozein dela arrazoia, hiztegi-kontsulta eta garbiketaren ostean, sistemak hiztegi-kontsulta fasetik itzulpen-hautagairik lortzen ez duenean, Al-Onaizan eta Knighten (2002a) sistemaren antzera, transliterazio teknikaren bitartez itzulpenak sortzeko saiakeran ezagutza linguistikoa aplikatuko du.

Euskara-gaztelera itzulpenean, asko dira fonetikoki berdinak izan arren, desberdin, nahiz eta oso antzera, idazten diren hitzak, hala nola, Kuba/Cuba, Ibarretxe/Ibarreche, etab. NE bikoteetan ematen diren aldaera fonetiko/fonologiko horiek bihurtzen dituen teknika da, hain zuzen ere, transliterazioa. Tesi-lan honen testuinguruan, euskara eta gazteleraren artean berdin ahoskatu eta desberdin idazten diren hitzak eskuz aztertu dira, NEetan ematen diren transliterazio-erregelak identifikatzeko asmoz.

Eskuzko azterketa horretan, Ixa Taldeko bi hizkuntzalarik idatzizko tes-

tuen azterketa sakon bat egin dute Hermes corpuseko (ikus III.2.1.1.1 atala) gaztelera eta euskarazko testu sorta bat erabiliz. Hizkuntzalarietako bat izen berezien aldaketak lantzen jardun du eta bestea aldiz, edozein izen eta adjektibok jasaten dituen aldaketak, ezin baita ahaztu NEak pertsona-izenez gain erakunde- eta toki-izenak ere badirela, eta hauek izen arrunt zein adjektiboz osatuak egon daitezkeela, *Udako Euskal Unibertsitatea*ren kasuan bezala. *Elhuyar*eko onomastika zerrenda ere oso lagungarria gertatu da lan horretan. Zenbait kasutan zerrenda horrek biltzen dituen pertsona-, toki-zein erakunde-izenen euskara eta gaztelera aldaerak erabili dira eskuzko azterketan zehar.

Hizkuntzalarien azterketa horretan identifikatutako aldaketa guztiak erregela-zerrenda batean bildu dira, 24 transformazio erregela fonetiko/fonologiko, hain zuzen ere. Zerrenda hori sisteman ezaugarri linguistikoetan oinarritutako NERC tresna (ikus III.3.1 atala) garatzeko erabili den *XFST* tresna bera erabili dugu. Horrela, esaterako *ze,zi -> ce,ci* transformazio erregela inplementatzeko, transliterazio-gramatikan ondoko erregela definitu dugu:

```
define Rulez [z (->) c "*" R 5, Z (->) C "*" R 5 _ [e | i | E | I]];
```

non, NEaren osagaian *z* bat eta jarraian *e* edo *i* bokalak topatzean *z* kontsonantea *c* bihurtzen duen, eta *c* kontsonantearen alboan **R5* erregelaren identifikadore marka txertatzen du. Sarrerako osagaia *Eskozia* balitz erregela horrek, *Eskoc*R5ia* itzulpen-hautagaia sortuko luke. Identifikadore marka hori txertatzen da, gramatikan 24 erregela definituta badaude ere, erregela horien kasuistika eta konbinatorioa oso zabala delako, eta hautagaiak osatzeko prozesuan parte hartu duten erregelak kontrolatzea beharrezkoa baita, bukaeran soilik zentzuzko itzulpenak aukeratu ahal izateko. Izan ere, aurrerago azaltzen den bezala, hainbat erregelen konbinazioek zentzugabeko hautagaiak sortzen baitituzte, eta horiek hurrengo urratsetan ez kontsideratzeko hautagaia lortzeko aplikatutako erregelak ezagutzea beharrezkoa da. Euskaraz eta gazteleraz forma antzekoak dituzten hitzak itzultzeko hizkuntzen menpeko NET sistema honen oinarri izango den transliterazio-gramatika osoa C eranskinean deskribatzen da.

NEaren osagai bati transliterazio-gramatika aplikatzean, irizpideak betetzen dituzten erregelak aplikatzen zaizkio irteeran itzulpen-hautagaien zerrenda bat lortuz. Zerrenda bat eta ez hautagai bakarra, askotan osagai bat transliteratzeko definitutako erregeletatik bat baino gehiago aplika baitaitezke eta, aplikazio ordenaren arabera hautagai desberdinak lortzen baitira.

Eskozia adibidearekin jarraituz, esaterako, *z -> c* transformazioa aplikatzeaz gain, gramatikan definituta dagoen *k -> c* bihurketa ere aplikagarria da. Hortaz, gramatika *Eskozia* sarrerari aplikatzean 3 hautagai sortuko dira, erregelak bakarka aplikatuta hautagai bana, eta biak konbinatuta beste hautagai bat: *Eskocia*, *Escozia* eta *Escocia*, hurrenez hurren.

Hasierako zerrendaren osagaien ordenak inolako zentzu berezirik ez badu ere, kontuan hartzen badugu erregela batzuk beste batzuk baino probableagoak direla, eta are gehiago, erregela-konbinazio batzuek ez dutela zentzurik, erregelen aplikazioan horien pisuak/probabilitateak markatzeak azken zerrendako osagaien ordenaketa eta aukeraketa erraztu dezake (Alegria *et al.*, 2010). Jarraian, erregelen aplikazioa eta haztaketa sistemaren zehaztasunak azaltzen dira.

Hasteko, erregela bera hitz bakar batean behin edo behin baino gehiagotan aplikagarria izatea gerta daiteke, *Aiastui -> Ayastuy* itzulpenean gertatzen den bezala. *i -> y* erregela bitan aplikatzen da. Oso maiz gertatzen da ere, hitz bakar batean erregela bat baino gehiago aplikatu behar izatea itzulpen egokia lortu ahal izateko. Esaterako,

- *Etxeberria -> Echeverria* transliteratzeko , *tx -> ch* eta *b -> v* erregelak aplikatu behar dira.
- *Kanpandegi -> Campandegui* transliteraziorako berriz, *c -> k*, *np -> mp*, *gi -> gui* erregelak.

Gramatika inplementatzeko erabilitako *Xerox*eko *XFST* tresnak erregelen aplikazio sekuentzial eta paraleloa simulatzeko aukera ematen du. Aurreko bi adibideetan erregelak sekuentzialki edo paraleloan aplika daitezke, izan ere, aplikatu beharreko erregelek elkarren artean ez baitute eragiten. Ez da gauza bera gertatzen jarraian azaltzen den *Kolonbia* adibidearekin.

Suposatu *Kolonbia* sarrera transliteratu nahi dela, eta transliterazio-erregelen artean ondoko hiruak aurki daitezkeela[11]:

- *k -> c*
- *nb -> mb*
- *b -> v*

[11] C eranskinean ikus daitekeen bezala erregelak hautazkoak dira eta hori adierazteko "(->)" karaktere-katea erabiltzen da. Adibideak sinplifikatzeko, ordea, formalismo hori ezabatu egin dugu. Hori bai, erregelak hautazkoak balira ulertu behar dira.

Erregelak biltzen dituen gramatika soilik paraleloan aplikatzeko aukera izanez gero, erregelak aplikatutako ordenaren arabera bi irteera posible egongo lirateke (k -> c erregelak ez du besteekin elkarrekintzarik, beraz, ez da adibidean aipatuko):

- Colonvia (b -> v)
- Colomvia (nb -> mb, b -> v)

Bi irteeretatik bat bera ez da zuzena. Gramatika sekuentzialki aplikatuko balitz aldiz, irteera posibleak ondokoak lirateke:

- Colonvia (b -> v)
- Colombia (nb -> mb)
- Colomvia (nb -> mb, b -> v)

Estrategia sekuentzial horrekin itzulpen zuzena ez diren bi hautagai sortu arren, itzulpen egokia eratzea ere lortzen da. Adibide horretatik, beraz, ondoko bi ondorioak atera daitezke:

1. Erregelen aplikazio-ordena garrantzitsua da.
2. Erregelen aplikazio paraleloaz gain, sekuentziala izatea kasu askotan ezinbesteko ezaugarria da itzulpen-hautagai egokia lortu ahal izateko.

Erregelen konbinatoria hain zabalarekin lan egitean, hautagaien gainsorkuntza handiaren zein hainbat erregela zentzugabeko konbinazioetatik lortutako hautagaien arazoekin topa gaitezke. Horrelako erregela aplikazio zabalaren ostean, beraz, hautagai aukeraketa egitea ezinbestekoa da, hautagaien zerrenda murriztu eta gaizki sortutako hautagaiak ezabatzeko.

Aukeraketa prozesuak hautagaien haztaketan oinarritutako bi irizpide nagusi jarraitzen ditu:

- hautagaia transliteratzeko aplikatutako erregela kopurua erabiltzen duen irizpidea;
- aplikatutako erregela-konbinatoriak kontrolatzen eta ponderatzen dituen irizpidea.

Lehenbiziko irizpideak, aplikatutako transliterazio-erregela kopurua aintzat hartuz, zenbat eta erregela gehiago erabili hautagaiaren pisua gutxitzen du, erregela bakoitzaren aplikazioa eta gero eratutako hautagaiaren pisua dekrementatuz.

Bigarren irizpidean aldiz, erregela kopurua erabili beharrean, erregela eta erregela-bikoteen ponderazioa erabiltzen da, non erregela bakoitzak pisu bat esleituta duen, eta erregela bat transliterazio prozesuan erabiltzean, pisu hori bere aplikaziotik sortzen den hautagai berria haztatzeko erabiltzen den. Pisu hauek estimatzeko eta zentzugabeko erregela-konbinazioak identifikatzeko *Elhuyar*reko hiztegi elebiduna erabili da.

Konkretuki, hiztegi elebiduneko euskara-gaztelera hitz-bikote bakoitzeko abiapuntu-hizkuntzako hitzei, euskarazko hitzei, alegia, gramatikan definitutako transliterazio-erregelak aplikatu zaizkie gaztelerazko hautagai multzo bat eskuratuz. Bikoteen jatorria hiztegi elebiduna izanik, xede-hizkuntzako itzulpen-forma egokia ezaguna da. Hortaz, automatikoki eskuratzen da transliteraziotik lortutako hautagaien artean zein den zuzena eta zeintzuk okerrak, eta ondorioz, zeintzuk izan diren erregela-konbinazio egokiak eta zeintzuk konbinazio okerrak.

Erregela-bikoteen zuzentasun probabilitateak automatikoki kalkulatzeko bidean, bikote bakoitzean transliterazio-erregelen konbinazio egoki eta okerren kontaketa eta normalizazio prozesu bat aplikatu dugu, hautagai egokia lortzen duten erregela-bikoteei hautagai okerrak itzuli dituzten erregela-konbinazioei baino probabilitate altuagoak esleituz.

Erregela-bikoteen probabilitate horiek izango dira, hain zuzen ere, hautagaien aukeraketa-estrategiak bigarren irizpiderako erabiliko dituen pisuak.

Transliterazio-gramatika aplikatzean beraz, sistemak pisuak esleituta dituzten itzulpenak eskuratuko ditu. Hautagai horien pisuetan oinarrituz, horrela, handienetik txikienera ordenatutako zerrenda bat sortuko da, goiko postuetan hautagai probableenak agertuko direlarik. Haztaketaren arrazoiak hautaketa prozesu batean jatorria duela kontsideratuz, eta, beraz, helburua zerrenda laburtzea dela, sistemak euskarazko NEaren osagai bakoitzeko soilik lehenbiziko n itzulpen-hautagaiak erabiliko ditu NE osoaren itzulpen-prozesuan jarraitzen duten urratsetan. n parametroaren balioa esperimentazio fasean definituko da.

Laburbilduz, lehen urrats honetan hiztegietatik zein transliterazio-erregelak aplikatuz, euskarazko NE baten osagai bakoitzeko gehienez ere baliozko n gaztelerazko itzulpen-hautagai dituen zerrenda bat eskuratzen da.

III.3.1.2 NE osoaren eraketa

NEaren hitz edo osagai bakoitzeko gehienez ere n itzulpen izanik, horien konbinazioetatik sor daitezkeen NE osoen hautagaiak eratzea da modulu honen funtsa. Euskarak eta gaztelerak egitura sintaktiko desberdina jarraitzen dutela kontuan izanik, osagaien itzulpen-hautagaien konbinazio egokiak sortzeko, NE osoarekiko horiek duten agerpen posizioa kontsideratu beharrean dago sistema.

NE osoa eraikitzerakoan euskaraz zein gazteleraz ordena bera mantentzen duten NEak dira errazenak osatzen. Ezaugarri hori bete ohi duten NEak pertsona-izenak dira. Esate baterako, euskarazko *Juan Jose Ibarretxe*ren gaztelerazko ordaina *Juan Jose Ibarreche* da, non osagaiek euskarazko zein gaztelerazko formetan ordena bera betetzen duten. Ordena mantentzen duten bestelako izen bereziak ere badaude, izen + adjektibo patroi sintaktikoa betetzen dituzten NEak adibidez. Esaterako *Fronte Iraultzaile -> Frente Revolucionario* erakunde-izen bikoteak ezaugarri hori betetzen du.

Hala ere, askotan NEen osagaien ordena euskarazko eta gaztelerazko formetan ez dira berdinak izaten. Esaterako, demagun *Lomeko Bake Akordio* NEa itzuli nahi dela. Osagai bakoitzeko itzulpen-hautagai bana ailegatzen dela suposatuz[12], gaztelerazko itzulpena euskarazko NEaren osagaien ordena bera mantenduta eratuz gero, *Lome Paz Acuerdo* forma okerra lortuko litzateke, NEaren osagaiak ordena okerrean utziz.

Osagaien ordenaren tratamendua beraz, beharrezkoa da NEen itzulpenak ondo sortzeko. Osagaien ordena identifikatzeko asmoz, *IXA talde*ko hizkuntzalari batek euskararen eta gazteleraren NEen patroi sintaktikoen arteko eskuzko azterketa burutu du. Azterketan, gehien errepikatzen diren aldaketa sintaktikoen portaerak, hitzen kategoria/azpikategoria eta deklinabidekasuetan oinarritutako erregela multzo batean taldekatu ditu.

Guztira 10 dira informazio morfosintaktikoan oinarritzen diren erregelak, eta D eranskinean aurki daitezke. Erregelak III.3.1 adibidearen antzekoak dira.

III.3.1 Adibidea Informazio morfosintaktikoan oinarritutako berrordenaketa erregela.

Bi izen arrunt jarraian eta lehenengoa deklinatu gabe dagoenean osagaien itzulpen-hautagaiei buelta eman eta *de* partikula tartekatu.
Adibidea: Bake Akordio

[12]Itzulpen-prozesu errealean horrela izan ez arren, adibidea sinplifikatzeko asmoz osagai bakoitzeko hautagai bakarra eskuratzen dela suposatuko da.

- *Bake*: *IZE_ARR* (izen arrunt), deklinatu gabe, *PLU-* (singularra)
- *Akordioetatik*: *IZE_ARR* (izen arrunt), *ABL* (ablatibo kasuan deklinatua), *PLU+* (plurala)

Sistemak, beraz, *Eustagger*en analisitik osagai bakoitzeko lema, pluraltasun eta kasuari buruzko informazioa eskuratzeaz gain, kategoria/azpikategoria informazioa ere eskuratu beharko ditu ordenazio-erregelak aplikatu ahal izateko. *Lomeko Bake Akordio* NEaren itzulpen-prozesuan esaterako, euskarazko NEari buruzko ondoko informazioa beharrezkoa izango du sistemak:

- *Lomeko*: *IZE_LIB* (leku-izen berezia), *GEN* (genitibo kasuan deklinatua), *PLU-* (singularra)
- *Bake*: *IZE_ARR* (izen arrunta), deklinatu gabe, *PLU-* (singularra)
- *Akordio*: *IZE_ARR* (izen arrunta), deklinatu gabe, *PLU-* (singularra)

Ordenazio-erregelak aplikatzeko garaian, osagaiak binaka eta eskuinetik ezkerrera tratatuko dira, azken osagaiaren deklinabide-kasuan erreparatu gabe, horrek ez baitu NEaren osagaien ordenari buruzko informaziorik eskaintzen. Beraz, eskuinean dauden lehenengo bi osagaiak izango dira lehenak aztertzen.

Aurreko *Lomeko Bake Akordio* NEaren itzulpen-adibidearekin jarraituz, ordenazio modulu honetan sistemak *Bake Akordio* osagaiek osatzen duten bikotea izango da aztertuko duen lehen bikotea. Bi izen arruntez osatutako bikotea izanik, III.3.1 adibideko erregela (D eranskineko *2.a.ii*) aplikatuko du sistemak, bi osagaien ordena trukatu eta *de* preposizioa tartekatzeko irizpidea definitzen duen erregela, hain zuzen ere, *Acuerdo de Paz* gaztelerazko itzulpen-hautagai partziala sortuz. Aurrerantzean itzulpen partzial hori osagai bakarra kontsideratuko da, ondorengo osagaiek eragindako aldaketak baleude, horiek itzulpen partzial osoari blokean aplikatuko zaizkiolarik.

Hurrengo urratsean, *Lomeko Bake* bikotearekin jarraituko du sistemak itzulpen partziala osatzen. Horretarako, sistemak erregelen artean genitibo kasuan dagoen izen berezi bat (*Lomeko*) eta jarraian izen arrunt (*Bake*) bat duen patroirik existitzen den aztertuko du. Patroi hori betetzen duen erregela, D eranskineko *1.b* erregela, hain zuzen ere, topatu eta horrek agintzen duen bezala, *Lomeko* osagaiaren hautagaia bukaerara pasako du, genitibo

kasua ordezkatzeko *de* preposizioa ere tartekatuz. Lortuko duen NE osoaren itzulpena, beraz, *Acuerdo de Paz de Lome* izango da.

Sistemak oro har, ordenazio-erregeletan bilaketak egiteko entitateko osagai-bikoteen egitura sintaktikoa kontuan hartzen du. Uneko egituraren pareko patroirik topatzen ez badu, abiapuntu-osagaien ordena xede-hizkuntzan mantendu egiten dela suposatzen du. Definitutako patroietan, azken finean, ordena-aldaketa behar duten kasuak besterik ez baitira biltzen.

Laburbilduz, beraz, modulu honetan NEaren banakako osagaien itzulpen-hautagaietatik hasiz NEaren hautagai osoak eratzen dira. Horretarako osagai bakoitzeko itzulpenak konbinatzen dira eta, NEaren osagaien itzulpenak xede-hizkuntzako espresioko zein posiziotan kokatu ebazteko, erregela sintaktikoetan oinarritutako ordenaketa erregelak aplikatzen dira. Irteeran gaztelerazko NEaren itzulpen-hautagai osoak itzultzen dira.

III.3.1.3 NEen itzulpen-hautagai egokien aukeraketa

NE osoaren hautagaiak lortzeko, aurretik aipatu den bezala, osagai bakoitzeko gehienez n hautagaien arteko konbinazioak egiten dira. Beraz, aurreko bi urratsetatik NEaren gehienez $n \times n$ itzulpen sortuko ditu sistemak. Hautagai horietan, hizkuntza-bikotearen ezaugarri morfosintaktikoak kontuan izan badira ere, ez da NEaren zentzua konprobatu bere osotasunean. Modulu honek, hain zuzen ere, hori du helburu: sortutako itzulpen guztien artean egokiena(k) aukeratzea, zentzugabekoak ezabatuz.

Hautagai guztien artean itzulpen egokiena(k) aukeratzeko, sistemak xede-hizkuntzan idatzitako testu sorta batera joko du itzulpen-hautagai horien erabilera kontrastatzeko asmoz. Emaitzen atalean ikusiko den bezala (ikus III.4.1 atala), hiru baliabide desberdinekin ebaluatu dugu azken modulu hau, bilaketak Google, Wikipedia eta III.2.1.1.1 atalean aurkeztutako Hermes euskara-gaztelera corpus konparagarrian burutuz. Azkenengoan Webera jotzea beharrezkoa ez badu ere, sistema lehenbiziko biekin ebaluatzean kontsulta online egiten du. Horrek, sistemaren erantzun-denbora, Web-aren kontsultaren erantzun-denboren menpe egotea eragingo du.

$n \times n$ gehienezko hautagaien kopuru maximoa oso handia izan daitekeenez, hautagai guztiekin erabilera kontrastatzeko baliabidera jotzeak sistemaren erantzun-denbora asko igo dezake, bereziki, Google eta Wikipediara jotzean. Erantzun-denbora hori jaisteko asmoz, sistemak baliabide horietako edozeinetara jo aurretik itzulpen-hautagaien aukeraketa bat burutuko du. Aukeraketa horretarako, NEen osagaiz osagaiko itzulpen-moduluan era-

bilitako pisuetan oinarritutako hautagaien aukeraketa-estrategia bera (ikus III.3.1.1.2 atala) erabiliko du.

Beraz, batetik, transliterazio-gramatikatik lortutako pisua erabiliko du, eta bestetik, hiztegi elebidunetik jasotako hautagaiei, guztiz fidagarritzat jotzen direnez, pisurik altuena emango zaie, bateko pisua, alegia. Osagai guztien hautagaien pisuekin, sistemak pisuen batezbestekoa kalkulatzen du, NE osoari pisu global hori esleituz. Pisu horretan oinarrituz, sistemak NEaren itzulpen-hautagai guztien artean n hoberenak edo, zehatzago, batezbesteko altueneko NEaren n itzulpenak aukeratzen ditu. Horrela, $n \times n$ hautagaietatik $n$ra laburtuko du azken zerrenda.

Ondoren esperimentuaren arabera sistemak Google, Wikipedia edo corpus konparagarrira, hiru baliabideetako batera joko du, itzulpen egokiena(k) aukeratzeko. Hiruetan, hautagaien agerpen-kopuruak erabiliz ebatziko da hautagai aukeraketa. Alegia, agerpen-kopuru altueneko hautagaia(k) itzultzen d(it)u sistemak sarrerako euskarazko NEaren itzulpen gisa[13].

Sistemak osatutako itzulpen guztien artean, batek berak ere ez duenean agerpenik erabilera kontrasterako dagokion baliabidean, sistemak ez du itzulpenik proposatuko.

III.3.1.4 NEen itzulpen-prozesuaren adibidea

Sistemaren moduluak hobeto ulertu eta sistemaren portaera osotasunean azaltzeko asmoz *Akabako Golkoa* euskarazko NEaren itzulpen-prozesu osoa azaltzen dugu jarraian.

NET tresnak, ondoko informazioa jasoko du NEa gaztelerara itzultzeko prozesua martxan jartzeko:

- *Akabako*: *Akaba* (lema), *IZE_LIB* (leku-izen berezia), *GEN* (genitiboan deklinatua), *PLU-* (singularra)
- *Golkoa*: *golko* (lema), *IZE_ARR* (izen arrunta), ABS (absolutiboan deklinatua), *PLU-* (singularra)

Lehenbizi, osagai bakoitzerako, itzulpen-hautagaiak eskuratzen saiatuko da sistema, aurrena *Akaba* osagairako eta ondoren *golko* osagairako. *Akaba*k hiztegi elebidunetan sarrerarik definituta ez duenez, sistemak transliterazio-gramatika aplikatuko du ondoko hautagai zerrenda lortuz (hautagai bakoitzerako aplikatutako erregelen araberako pisuak ere eskuratuko ditu):

[13] Google-en kasurako, GoogleApi zerbitzua erabili da. Eta Wikipedia kontsultatzeko berriz, `http://www.mediawiki.org/wiki/API` zerbitzua.

- Akaba 1
- Akava 0,125
- Acaba 0,81
- Acava 0,055

golko osagaiaren kasuan, horrek hiztegi elebidunean sarrera duenez, lehenengo urratsean ondoko hautagai zerrenda eskuratuko du. Gainera, itzulpenen jatorria hiztegia denez, sistemak hautagai guztiei 1eko pisua emango die.

- golfo 1
- conciencia 1
- pecho 1
- regazo 1
- seno 1

Izen arrunt horren itzulpenak ez direnez izen/adjektibo zerrenda berezitik eskuratu, *golko* osagaiaren kategoria, hurrengo urratsetan izen arrunta izaten jarraituko du.

Genero eta pluraltasun markei erreparatuz, lehenengo osagaiaren kasuan, hiztegitik ez direnez eskuratu, sistemak ez du itzulpen-hautagai zerrenda moldatuko. *golko* hautagaiaren kasuan, aldiz, horietako batzuk *oz* bukatzen direnez sistemak generoaren erregela aplikatu behar die[14], irteeran ondoko zerrenda osatua sortuz:

- golfo 1
- conciencia 1
- pecho 1
- regazo 1
- seno 1
- golfa 1
- pecha 1

[14] genero sorkuntzan sortzen den hautagaia existitzen denik ez da egiaztatzen

- regaza 1
- sena 1

Osagai bakoitzerako itzulpenak izanik, sistema NEerako itzulpen-proposamen osoak eratzen hasiko da. Puntu honetan sistema ordenazio-erregelak aplikatzen hasiko da osagaiak eskuinetik ezkerrera eta bikoteka aztertuz. Adibide honetan bikote bakarra egonik, hortik hasiko da. Leku izen berezi bat eta jarraian izen arrunt bat patroia erregelen artean bilatuz, sistemak D eranskineko *1.b* erregela topatuko du, non osagaiak alderantzizko ordenan eta *de* partikula tartekatu behar dituela adierazten zaion. Erregela hori aplikatuz, sistemak ondoko itzulpenak lortuko ditu:

- golfo de Acaba 0,9
- conciencia de Acaba 0,9
- ...
- golfo de Acava 0,525
- conciencia de Acava 0,525
- ...
- golfo de Akava 0,5625
- conciencia de Akava 0,5625
- ...
- golfo de Akaba 1
- conciencia de Akaba 1
- ...

Adibidean ikus daitekeen bezala, zerrendako NEaren itzulpen oso bakoitzak pisu bat du esleituta, NEaren osagai bakoitzaren hautagaien pisuetan oinarrituta kalkulatutako batezbesteko pisua, hain zuzen ere. Pisu horietan oinarrituz, balio altuenak dituzten n hautagaiak aukeratzen ditu sistemak, egiaztapenerako baliabidera jotzeko. Suposa dezagun, n altu bat definituta duela sistemak, esaterako 50. Hala izanda, aurreko zerrendako itzulpen-hautagai guztiak pasatuko dira egiaztapenerako baliabidean bilatzeko urratsera. Erabilera kontsultaren ostean, sistemak *Golfo de Akaba* eta *Golfo*

de Acaba itzuliko ditu *Akabako Golkoa* NEaren itzulpen-proposamen gisa, ordena horretan Google-en *Golfo de Akaba* espresioak *Golfo de Acaba*k baino agerpen gehiago baititu, eta gainerako hautagaiak baztertu egingo ditu, Google-en ez baitute agerpenik.

III.3.2 Hizkuntzekiko erdi-independentea den NET tresna

Sistema honen helburua, hizkuntzen menpekoarena bezala, abiapuntu-hizkuntzako NE bat emanik, xede-hizkuntzako baliokidea lortzea da. Aurrekoak bezala, sistemak hiru urrats nagusitan banatzen du lana. Lehenbizi, osagaiz osagaiko itzulpena burutuko du, transliterazio-gramatika eta hiztegi elebidunak konbinatuz, berriz ere, baina aurrekoarekin desberdintasun nagusi batekin: transliterazio-gramatikako erregelak hizkuntza-bikoteko ezaugarri linguistikoetan oinarrituta egon ordez, hizkuntzekiko independenteak diren erregelekin definituta daude.

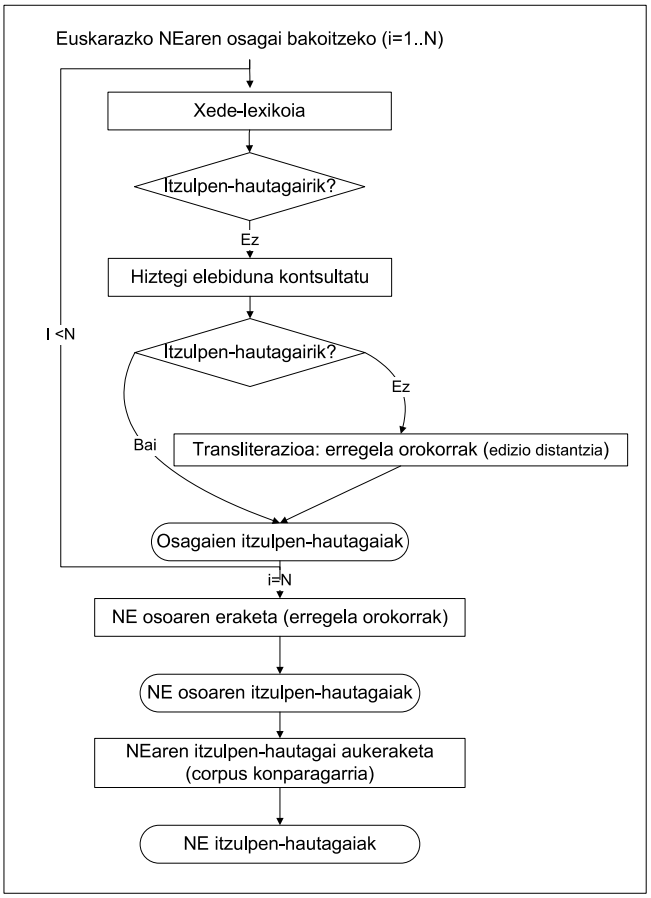
Ondoren osagai bakoitzaren itzulpen-hautagaiekin, NE osoa eraikitzen saiatuko da, horretarako hizkuntzekiko independenteak diren elementuen ordenazio-patroiak aplikatuz.

Azkenik sistemak, sortutako NEaren hautagaietatik egokienak aukeratzeko, corpus konparagarri batean egingo du kontsulta, formarik usuenak itzulpen-proposamen bezala bueltatzeko. Urrats horiek guztiak III.4 irudian bilduta daude.

Sistemaren moduluak ikusita, hiztegi elebidunen erabilera dela eta, ezin da esan hizkuntzekiko erabat independentea denik tresna hau. Bai, ordea, erdi-independentea, izan ere, hizkuntzen ezaugarri linguistikorik ez baita erabiltzen ez transliterazio ezta ordenazio patroietan ere.

Hizkuntzekiko erdi-independentea izanik sistemaren izaera, bi hizkuntza-bikoterekin ebaluatu da: euskara-gaztelera eta gaztelera-ingelesa. Lehenbiziko bikoteak, hizkuntzen menpeko NET tresnaren emaitzekin konparatzeko aukera ematen digu eta bigarrenak, aldiz, bestelako hizkuntza-bikote batera eramatean sistemak agertzen duen portaera ebaluatzeko.

Sistemak ahalik eta ezaugarri linguistiko gutxien erabiltzea du helburu. Dena den, hiztegi elebiduna eta osagaien lemak ezagutzea beharrezkoa du hainbat urratsetan, esaterako lehenengo urratseko hiztegi elebidunen kontsultan. Euskara eta gaztelera izanik definitutako bi bikoteen abiapuntu-hizkuntzak, lematizatzaile bana behar izan da NET sistemari beharrezko informazioa luzatzeko. Euskararako *Eustagger* erabili da eta gaztelerarako,



III.4 Irudia: Hizkuntzekiko NET tresna erdi-independentearen diseinua

berriz, *Freeling*[15] tresna (Atserias *et al.*, 2006), *TALP* ikerkuntza-taldeak garatutako hainbat hizkuntzen analisirako tresna, besteak beste gaztelera, ingelesa eta katalana analizatzeko gaitasuna du.

Ondorengo ataletan, urrats bakoitzaren azalpen sakonagoa egiten da.

III.3.2.1 NEen osagaiz osagaiko itzulpena

NEaren osagai bakoitzeko itzulpen-hautagai zerrenda bat lortzea helburu duen modulu honek, bi osagai nagusi ditu: transliterazio-gramatika eta hiztegi elebiduna. Lehenbizikoa bi hizkuntzetan antzera idazten diren hitzen

[15] `http://nlp.lsi.upc.edu/freeling/`

itzulpenak eskuratzeko erabiliko da eta bigarrena, aldiz, bestelako osagaien itzulpenak eskuratzeko.

III.3.2.1.1 Transliterazio-gramatika

Sistema honetako transliterazio-gramatika, edizio-distantzian (Kukich, 1992) oinarritutako erregela multzoa eta xede-hizkuntzako lexikoia konbinatuz lortzen da. Sistema azkarra izan dadin automata batean konpilatzen da. Edizio-distantzian oinarritutako erregelek karaktereen oinarrizko hiru edizio eragiketak hartzen dituzte kontuan: karaktereen ezabaketa, ordezkapena eta txertaketa. Jarraian dauden bi karaktereen trukaketarako erregelarik ez da definitu, eragiketa hori ezabaketa eta gehikuntza erregelak konbinatuz simulatu baitaiteke eta, hortaz, ez da beharrezkotzat jo. Edizio-distantzian oinarritutako hiru eragiketa horiek *XFST* tresna erabiliz definitu dira.

III.3.2 adibidean oinarrizko hiru eragiketa horien *XFST* erregelak deskribatzen dira.

III.3.2 Adibidea Edizio-distantzia kalkulatzeko oinarrizko eragiketen *XFST* erregelak.

```
define Alfabeto [ a | b | c | d | e | f | g | h | i | j | k | l | m |
       n | ñ | o | p | q | r | s | t | u | v | w | y | z | x | %- | %.];
define Ezabatu [Alfabeto (→) "1"];
define Txertatu1 [Alfabeto (→) ... "2" Alfabeto];
define Txertatu2 [Alfabeto (→) "2" Alfabeto ...];
define Txertatu [Txertatu1 .o.  Txertatu2];
define Ordezkatu [Alfabeto (→) "3" Alfabeto];
```

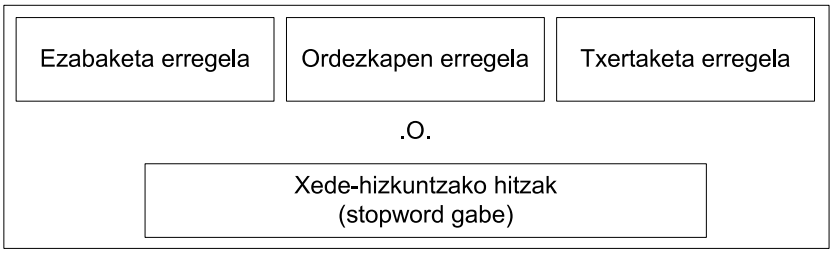
Erregela guztietan agertzen diren zenbakiak, osagaia eratzean aplikatu diren erregelei buruzko informazioa eskuratzen laguntzen du, erregela bakoitza zenbaki batekin identifikatuz. Txertaketa erregelaren kasuan, edozein karakteretik hasita, txertaketa aurretik zein atzetik eman daitekeenez, bi erregela definitzen dira, eta ondoren, biak konbinatuz *Txertatu* erregela orokorra definitzen da.

Edizio-distantziako hiru eragiketa horiek, edozein izanik hizkuntza, hitz bakoitzeko nahi adina aldiz (m) aplika daitezke: bakoitza bere aldetik zein beste batzuekin konbinatuz. Zenbat eta m altuagoa ezarri, orduan eta irteerako osagai kopuru handiagoa lortuko da. Hipotesi kopuru handia ekiditeko eta prozesua azkartzeko, tesi-lan honetan bi erregela aplikatzeko murriztapena ezarri da, m=*2* alegia. Horrela, transliterazio-automata eraikitzeko erregelak xede-hizkuntzako lexikoi batekin konbinatu dira (III.5 irudian .O.

espresioarekin adierazita), prozesu horretan lexikoi sarrera bakoitzeko gehienez ere 2 aldaketa erregela aplikatuz. Modu horretan, transliteratutako osagai kopuru bideragarria eskuratuko da eta sistemak hurrengo urratsak modu eraginkorragoan aplikatu ahal izango ditu.

Adierazitako lexikoia eskuratzeko Hermes corpusa (ikus III.2.1.1.1 atala) erabili da. Konkretuki, euskara-gaztelera bikoterako lexikoia berrien sorta horretan gaztelerako hizkuntzako testuetan identifikatutako 106.473 NEen **55.656** osagai desberdinekin osatu da. Gaztelera-ingelesa bikoterako berriz, Hermes corpuseko 49.768 NEen **37.500** lema desberdinekin osatu da ingelesezko lexikoia.

Lexikoi horietatik *stopword* zerrendak[16] erabiliz artikuluak, preposizioak, eta oro har hitz gramatikal guztiak ezabatu ondoren, lexikoiak edizio-distantzia erregelekin konbinatu dira III.5 irudian ikus daitekeen bezala.



III.5 Irudia: Transliterazio-automata

Guztira hiru transliterazio-automata (TA) sortu dira. Funtsean aplika daitekeen transformazio kopurua zehazten duen m parametroa besterik ez da aldatzen, irteerako osagaia sortzeko aplika daitekeen erregela kopuru maximoa, alegia:

- Transformaziorik aplikatzen ez duen automata, sarrerako hitza irteeran berdin itzultzen duen automata
- Gehienez ere transformazio bat aplikatzen duen automata
- Hitzak eratzeko gehienez bi transformazio aplikatzen dituen automata

Edizio-distantziako erregelak eta lexikoiak konbinatuz sortutako transliterazio-automata beraz, *Kuba* euskarazko osagairako *Cuba* gaztelerazko forma transliteratua sortzeko gai izango da K karakterea C

[16] http://www.lc.leidenuniv.nl/awcourse/oracle/text.920/ a96518/astopsup.htm

karakterearekin ordezkatuz. Eta gaztelerazko *Constitución* osagairako, aldiz, *Constitution* ingelesezko itzulpena lortuko du *c -> t* eta *ó -> o* karaktereen ordezkapenak burutuz.

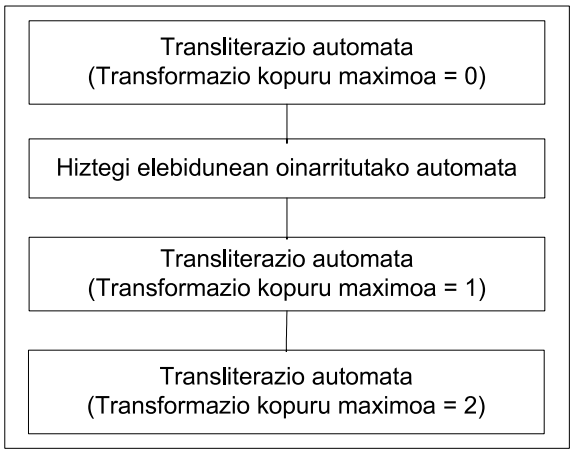
III.3.2.1.2 Hiztegi elebidunak

Aurretik aipatu den bezala, hainbat hitzen itzulpenak lortzeko transliterazioa ez da nahikoa, horien formak oso desberdinak izan baitaitezke hitzen abiapuntu- eta xede-hizkuntzetan. Hitz-mota horien itzulpenak eskuratzeko, sistemak hiztegi elebidunak erabiliko ditu, gutxienez bikote bakoitzerako hiztegi bat erabiliz. Euskara-gaztelera bikoterako baliabideen atalean deskribatutako (ikus III.2.1.2 atala) *Elhuyar Hiztegia* erabili da eta gaztelera-ingelesa bikoterako, berriz, MCR (*Multilingual Central Repository*) hiztegia.

Hiztegi horien erabilera transliterazioaren automaten metodologia bera jarrai dezan, hiztegi bakoitzeko, bertako hizkuntza-bikote bakoitzerako itzulpena burutzen duen automata bat sortu da *XFST* tresnarekin. Automata horiek, abiapuntu-hizkuntzako hitz bat emanda, sarrera horretarako hiztegian definituta dauden itzulpen-hautagaiak itzultzen dituzte. Horrela, hitz baten itzulpena eskuratzeko edozein dela aplikatu beharreko baliabidea, hiztegia edo transliterazioa, sistemak automata bat edo beste aplikatu besterik ez du egin beharko.

Automata horien guztien aplikazio ordena III.6 irudian aurkezten da. Lehenbizi, sarrerako osagairako sistemak itzulpen-hautagaiak transformaziorik aplikatu gabe eskura daitezkeen aztertzen du. Alegia, *m=0* parametroarekin sortzeko automata exekutatzen du, sarrerako osagaia lexikoian dagoen ikusteko; hortik proposamenik lortzen ez denean, hiztegi elebidunean oinarritutako automatatik hautagaiak eskuratzen saiatuko da, eta baten bat eskuratu ezean, alegia, osagaia hiztegi elebidunean definituta ez dagoenean, transformazio-kopuru maximoa handitzen automatak aplikatzen joango da. Irteeran, sarrerako osagaietarako n hautagai proposatuko dira. n hautagai baino gehiago sortzen diren kasuan, aplikatutako erregela kopuruan oinarritutako aukeraketa aplikatuko da: zenbat eta erregela gehiago orduan eta pisu txikiagoa izango du hautagaiak. Hortaz, pisu handieneko lehenengo n hautagaiak itzuliko ditu itzulpen gisa ondorengo moduluetan NE osoaren itzulpen-prozesuan erabili ahal izateko.

Oro har, NEen osagaiz osagaiko itzulpen-moduluak, beraz, corpus konparagarrietan zehar agertzen diren hitz transliteratuetarako proposamenak sortzeaz gain, transformazioak nahikoak ez direnean hiztegi elebidunen infor-



III.6 Irudia: Osagaiz osagaiko itzulpenerako automata

mazioa beharrezko duten *erakunde - organización* bezalako hitzak itzultzeko ere gai da.

III.3.2.2 NE osoaren eraketa

Hizkuntzen menpeko NET tresnan bezala, NEaren hitz edo osagai bakoitzeko itzulpen-hautagai zerrenda definituta, horien konbinazioetatik sor daitezkeen NE osoen hautagaiak sortzea da modulu honen funtsa. Euskara-gaztelera eta gaztelera-ingelesa hizkuntza-bikoteetan, egitura sintaktiko desberdineko hizkuntzak izanik, xede-hizkuntzako NEak eratzean osagaien ordena aldaketa tratatu beharreko berezitasuna da. Hala izanik, hizkuntzekiko erdi-independentea den sistema honetan, bi hizkuntza espezifikoren patroi sintaktikoak kontuan harturik osagaien ordenaketa egin beharrean, hizkuntzekiko independentea den estrategia jarraitu da.

Ordenazio estrategia hori, abiapuntuko edozein osagai xedeko NEaren edozein posiziotan ager daitekeeneko hipotesian oinarritzen da. Hortaz, modulu honek, osagai bakoitza xedeko espresioan posizio guztietan ager daitekeela kontsideratuz, osagai bakoitzeko itzulpen-hautagaiak beste osagaien hautagai guztiekin konbinatuko ditu posizio posible guztietan, NErako itzulpen osoen zerrenda luze bat sortuz. Esate baterako, *Nazioarteko Federazioa* euskarazko NErako, *internacional* eta *federación* osagai itzulpenak jasoz, modulu honek *federación internacional* eta *internacional federación* hautagaiak emango ditu, osagai bakoitza bi posizioetan ager daitekeeneko hipotesia apli-

katuz.

Modulu honek euskara-gaztelera bezalako bikoteetan, non euskarazko deklinabide-kasuek itzulpen-atazan preposizioen sorkuntza eragin dezaketen (*Bake Akordio -> Acuerdo de Paz* bezalako) kasuetan, preposizio horiek NEen osaketan beharrezkoak izango dira eta, ondorioz, ez du hautagai egokirik sortuko, bertan horrelako sorkuntzarik ez baita tratatu. Gabezia hori hurrengo moduluan tratatuko da, ordea.

III.3.2.3 NEen itzulpen-hautagai egokien aukeraketa

Sistemak NEen itzulpen-hautagaiak sortu dituenean, hautagaien artean egokienak aukeratzen hasiko da. Aukeraketarako, hautagaiaren zuzentasuna egiaztatzeko xede-hizkuntzako corpus konparagarria erabiliko du. Web-era ez jotzearen arrazoia emaitzen atalean hizkuntzen menpeko tresnaren ebaluazioan ikusiko dugun bezala (ikus III.4.1 atala), corpus konparagarriekin lortzen diren emaitzen parekoak lortzen direla da, eta erantzun-denbora aldetik, sistema hobea da Web-era jotzen duenean baino.

Zehazki, sistemak hautagai bakoitzerako corpus konparagarriko xede-hizkuntzako berrietan duen agerpen-kopurua eskuratuko du. Agerpen-kopuru horietan oinarrituz, zerrenda handienetik txikienera ordenatuko du eta lehenengo postuetan NE osoaren itzulpen usuenak geratuko dira. Lehenbiziko n postuetako hautagaiak izango dira, hain zuzen ere, sistemak itzulpen gisa itzuliko dituen NEak.

Corpus konparagarrian itzulpen-hautagaien bilaketa egitean, xedeko corpusean NEaren osagaien artean tartekatuta preposizioak ager daitezkeela aurreikusten da. Are gehiago, xedeko corpuseko bilaketan NEa partikula gramatikalen bat tartekatuta duela topatzen bada, forma berri hori izango da azken zerrendan agertuko dena, eta ez bilaketarako erabilitako forma. Horrela, aurreko atalean aipatu dugun gabezia ebatzita geratzen da. Adibidez, sistemak *Frantziako Gobernua* NErako *Gobierno Francia* hautagaia lortzen duenean, eta corpus konparagarrian *Gobierno de Francia* forma topatzean, *Gobierno de Francia* azken forma hori sartuko du bukaerako itzulpenen artean, eta ez bilaketa aurretik sortutako *Gobierno Francia* forma. Bilaketa flexiblea egiten du sistemak, beraz, sortu gabeko partikulak bilaketan eragin negatiborik izan ez dezaten. Gainera, partikulen tratamendua horrela egiteak, sistemaren hizkuntzekiko independentzia mantendu egiten du, ez baita inolako kasuen ez egitura sintaktikoen azterketarik egin behar partikulen sorkuntza egiteko.

Osatutako NEaren itzulpen guztien artean xede-hizkuntzako corpusean gutxienez agerpen bat duen hautagairik ez badago, alegia, azken zerrenda hutsa denean, sistemak ez du itzulpenik proposatuko.

Pauso guztiak hobeto ulertzeko, ondorengo puntuan adibide baten bitartez deskribatzen da prozesu osoa.

III.3.2.4 NEen itzulpen-prozesuaren adibidea

Kapituluan zehar erabilitako adibideak euskara-gaztelera hizkuntza-bikotekoak izan direnez, ondorengo azalpenerako gaztelera-ingelesa NE bikote bat erabiliko da. *Comisión Europea - European Commission* bikotea hain zuzen ere.

Lehenbizi, sistema osagaiz osagaiko itzulpena burutzen hasiko da. *Comisión* hautagaiaren kasuan zero transformazioko automata aplikatzean ingelesezko lexikoian forma bera duen hitzik topatuko ez duenez, hiztegi elebidunean oinarritutako automata aplikatuko da. Automata horretatik ondoko itzulpenak eskuratuko ditu:

- Commission
- Commissioning
- Committee
- Delegacy
- Delegation
- Deputation
- Mission
- Perpetration

Bigarren osagairako aldiz (hiztegian *Europeo* sarrera existitzen den arren, *Europea* sarrera ez denez existitzen) transformazio bat gehienez onartzen duen automata aplikatuta ondoko itzulpen-hautagaiak eskuratuko ditu:

- Europea
- Europa
- Europe
- European

Osagai bakoitzerako itzulpenak izanik, NE osoa osatzera pasako da. Horretarako bi osagaien itzulpen guztiak konbinatuko ditu permutazio posible guztiak sortuz. Zerrenda oso luzea denez, pausu hau ikusteko osagai bakoitzerako bi hautagai besterik ez direla eskuratzen suposatuko dugu: *Commission* eta *Committee*, eta *Europe* eta *European*, hurrenez hurren. Bina hautagai horiekin sistemak ondoko zerrenda hau sortuko luke:

- Commission Europe
- Committee Europe
- Commission European
- Committee European
- Europe Commission
- Europe Committee
- European Commission
- European Committee

Zerrenda horretako hautagaiak corpus konparagarrian bilatuz eta horien agerpen-kopuruan oinarrituz, sistemak itzulpen egoki bezala *European Commission* itzuliko du, hori baita guztietatik agerpena duen ordain bakarra.

III.3.3 Wikipedian oinarritutako NET tresna

Euskara abiapuntu-hizkuntza eta ingelesa xede-hizkuntza duen Wikipedian oinarritutako NET tresnak, NEen itzulpenak automatikoki burutzeko aurreko NET sistemetako hiru modulu nagusiak erabiltzen ditu. Are gehiago, NEen osagaiz osagaiko itzulpen-modulua eta NE osoaren eraketa moduluak III.3.2.1 eta III.3.2.2 ataletan deskribatutako estrategiekin eraiki dira, hurrenez hurren, oraingoan euskara-ingelesa bikoteko lexikoiak Wikipediatik eskuratuz. Eta NEen itzulpen-hautagaien aukeraketa moduluaren kasuan, hautagaien erabilera maila Wikipedian egiaztatzen da.

Sistema hau, beraz, Hermes corpus konparagarrian oinarritu beharrean, hizkuntza desberdinetako Wikipedietan oinarritzen da (ikus III.2.1.1.2 atala). Konkretuki, gure esperimentuetarako euskara-ingelesa bikotea aukeratu dugu. Euskara tesi-lan honetako interesa izanik, euskara sistemako abiapuntu-hizkuntza izateak inolako zalantzarik ez du sortzen. Eta xede-hizkuntza ingelesa aukeratu dugu, Wikipedia guztien artean landuen dagoena izanik, corpus

konparagarri osoena eta handiena osatzeko aukera eskaintzen digun hizkuntza baita batetik. Eta bestetik, lan honekin hasi ginean testuan guztian zehar NEak (automatikoki) etiketatuta zituen Wikipedia bertsio bakarra ingeleseko bertsioa baitzen.

Wikipedian oinarritutako NET tresnak, beste sistemetan erabilitako 3 modulu nagusiez gain, Wikipediaren egituraren ezaugarriak baliatuz itzulpen-hautagaiak eskuratzeko beste modulu berri bat du bere arkitekturan. Modulu berri hori, beste moduluak baino lehenago aplikatzen da. Eta sistemak bertatik itzulpenik eskuratzen ez duenean bakarrik NEaren osagaiz osagaiko modulura joko. Modulu berri horretan itzulpena eskuratzen duenean, beraz, zuzenean aukeraketa modulura salto egingo du sistemak. Modulu berri horrek, funtsean bi urrats ematen ditu.

Lehenbizikoan, Wikipediak eskaintzen dituen hizkuntzen arteko estekak (WILak) baliatzen ditu. WIL estekak jarraituz itzulpen-proposamenak bilatzen saiatzen da. Alegia, sistemak abiapuntu-hizkuntzaren Wikipedian sarrera bilatzen du eta hau existitzen bada eta horri lotuta xede-hizkuntzara zuzendutako WIL estekarik badago, esteka hori jarraituz Wikipediatik eskuratzen den NEa sistemaren aukeraketa modulura pasako du zuzenean, elementuz elementuko itzulpena burutu gabe, III.7 irudian ikus daitekeen bezala.

Bigarren urratsean, WIL esteken bidez sarrerako NEerako itzulpenik lortzen ez denean, abiapuntu-hizkuntzako NEaren espresioa xede-hizkuntzako Wikipedian bilaketa burutzen du. Bilaketa horretan xede-hizkuntzako Wikipedian abiapuntu-hizkuntzako NEarekin bat datorren sarrera topatzen denean, sarrerako forma izango da aukeraketa modulura bidaltzen den itzulpen-proposamen bakarra.

Sisteman gehitutako Wikipedian oinarritutako bi faseko urrats berritik itzulpen-hautagairik topatzen ez denean soilik, beraz, aplikatuko dira osagaiz osagaiko itzulpena eta NE osoaren eraketa moduluak. NEen aukeraketa modulua, aldiz, beti aplikatuko da. Sistemaren egitura osoa, III.7 irudian ikus daiteke[17].

Emaitzen atalean ikusiko dugun bezala (ikus III.4.3 atala), modulu berria egindako bi esperimentuetatik bakar batean ebaluatu dugu. Arrazoi nagusia, esperimentuetako baterako ebaluazio-corpusa Wikipediaren WIL estekak esplotatuz eraiki dela da. Horrela izanik, sistema ebaluatzeko Wikipediaren WILak esplotatzen dituen urratsa erabiliz gero, sistemak beti asmatuko lu-

[17] Tamaina dela eta, irudi hau kapituluaren bukaeran dago eskuragarri

ke, sistemaren portaera erreala hori izan ez arren. Hori dela eta, Wikipediatik at eraikitako ebaluazio-corpuseko esperimentuetan soilik ebaluatu dugu sistema bere osotasunean. Beste corpusarekin, aurreko sistemekin konpartitzen dituen hiru moduluak soilik aplikatuz ebaluatu da. Izan ere, beste hiru modulu horiek Wikipediaren informazioa erabiliz eraiki badira ere, soilik xede-hizkuntzako informazioa erabili da, sarreren artean WILik existitzen den aztertu gabe. Hortaz, hiru modulu horien ebaluaziorako agertoki erreala dela esan daiteke.

III.3.3.1 NEen osagaiz osagaiko itzulpena

III.3.2.1 puntuan azaldutako urrats berberak aplikatu dira modulu hau eraikitzeko garaian: edizio-distantzian oinarritutako erregelak xede-hizkuntzako lexikoi batekin konbinatzen dira hiru automata desberdin sortuz, funtsean horiek onartzen dituzten transformazio kopuruan desberdintzen direlarik.

Erabilitako xedeko lexikoia, ordea, ez da berdina izan. Aurrekoan berrietan oinarritutakoa bazen, oraingoan tresnaren izenak agertzen duen bezala, xede-hizkuntzako Wikipedian oinarrituta dago. Euskara-ingelesa izanik tresnak lantzen duen hizkuntza bikotea, transliterazio-automatak eraikitzeko ingelesezko Wikipediako NEetan oinarritutako lexikoia erabili da.

Hiztegi bidezko itzulpenerako automata eraikitzeko berriz, OpenTrad[18] proiektuan erabilitako euskara-ingelesa hiztegi elebiduna erabili da.

Edizio-distantzian oinarritutako hiru automatak eta hiztegian oinarritutako automata sekuentzialki aplikatzen dira, hurrengo urratsa soilik uneko urratsean itzulpen-proposamenik lortzen ez denean aplikatzen delarik, III.6 irudian agertzen den bezala.

III.3.3.2 NE osoaren eraketa

Modulu honetan III.3.2.2 atalean deskribatutako metodologia bera aplikatzen da. Alegia, aurreko modulutik eskuratutako osagai bakoitzaren itzulpen-hautagai guztiak beste osagaien itzulpen-proposamenekin konbinatzen ditu, abiapuntu-espresioko edozein osagai xede-hizkuntzako espresioan edozein posiziotan ager daitekeela kontsideratuz, permutazio guztiak sortuz.

Esaterako, *Nazioarteko Federazioa* euskarazko NEarentzat, *International* eta *Federation* osagai itzulpenak jasoz, modulu honek *Federation Internatio-*

[18] `http://www.opentrad.org/`

nal eta *International Federation* hautagaiak itzuliko ditu, osagai bakoitza bi posizioetan ager daitekeeneko hipotesia aplikatuz.

III.3.3.3 NEen itzulpen-hautagai egokien aukeraketa

Aurreko moduluetatik eskuratutako NE osoaren itzulpen-hautagai guztien artean egokiena(k) aukeratzeko, Wikipedian oinarritutako NET sistema honek bi baliabide erabiltzen ditu bilaketa egiteko:

- Wikipediaren berbideratze-estekak
- Wikipedian oinarritutako corpus konparagarria

Sistemak, osagaiz osagaiko itzulpena eta NE osoaren eraketarako urratsak eman ez dituenean, Wikipediaren ezaugarrietan oinarritutako urratsetik zuzenean lortutako itzulpen-proposamenak eskuratu dituenean, alegia, azken zerrenda osatzeko sistemak soilik Wikipediako berbideratze-estekak erabiliko ditu. Esteka horien bitartez, Wikipedian NE hori adierazteko beste formarik definituta dagoen aztertzen da. Mota horretako estekaren bat existitzen bada, NEaren itzulpena eta bertara berbideratutako orri guztien formak itzuliko ditu itzulpen-proposamen gisa.

Osagaiz osagaiko itzulpena eta NE osoaren eraketaren bidetik lortutako hautagaiekin gauza bera egiten da. Alegia, berbideratze-estekak erabiltzen dira azken zerrenda osatzeko. Baina osaketa horren aurretik, sistemak osatutako itzulpen guztien artean, Wikipedian oinarritutako corpus konparagarria erabiliz xedeko Wikipedian sarrera duten hautagaiak soilik filtratutako ditu. Filtraketa hori gainditzen duten hautagaiekin eta horietara berbideratutako orrien formekin osatuko da azken zerrenda, hain zuzen ere.

III.3.3.4 NEen itzulpen-prozesuaren adibidea

Aurreko tresnetan aipatutako adibideak osagaiz osagaiko itzulpena eta ondoren eraketa eginez itzulitako NEen kasuak izan direnez, oraingo honetan, Wikipedia bera bakarrik erabilita itzultzen den kasu bat azaltzen da, beste bidea aurreko tresnetan ikusitakoaren berdina baita.

Hortaz, demagun *Donibane Lohizune* euskarazko NEa itzuli behar duela sistemak. Lehenbiziko pausuan, euskarazko Wikipedian NE horretarako sarrerarik definituta dagoen begiratuko du sistemak, eta orri hori existitzen

dela ikusiko du. Hurrengo urratsa, beraz, orri horrekin WILen bitartez lotutako orrien artean, ingelesezko Wikipediako orririk dagoen aztertuko du, eta ingelesezko esteka jarraituz *Saint-Jean-de-Luz* forma aurkituko du.

Saint-Jean-de-Luz itzulpena izango da, beraz, NEaren aukeraketa fasera zuzenean bidaliko den proposamena. Itzulpen-hautagaia Wikipedian oinarritutako urratsetik datorrenez, ez dago egiaztatu beharrik xede-hizkuntzako Wikipedian sarrera existitzen denik, hori ziurtatuta baitago. Beraz, zuzenean berbideratze-estekarik definituta dagoen aztertzera joko du. Kasu honetan berbideratze-esteka bat topatuko du, eta hori jarraituz, hain zuzen ere, sistemak ondoko itzulpen-zerrenda hau osatuko du:

- *Saint-Jean-de-Luz*
- *Donibane Lohizune*

Gainerako kasuetan III.3.2.4 ataleko adibidean azaldutako pausoak jarraituko lirateke, baina moduluak eraikitzeko baliabide desberdinak erabili direnez, emaitza desberdinak sor daitezke.

III.4 Emaitzak

Atal honetan, aurreko atalean deskribatutako tresnen ebaluazioa aurkeztuko dugu. Sistema bakoitza hizkuntza-bikote desberdinekin eta ebaluazio-corpus desberdinekin ebaluatu denez, hiru sistemen portaerak banan-banan aztertuko ditugu. III.4 taulan egindako esperimentu guztien laburpena ikus daiteke.

Tresna	**Hizkuntza-bikotea**	**Ebaluazio-corpusa**
Hizkuntzen menpekoa	Euskara-gaztelera	Onomastika
Hizkuntzen menpekoa	Euskara-gaztelera	Hermes
Hizkuntzekiko erdi-independentea	Euskara-gaztelera	Hermes
Hizkuntzekiko erdi-independentea	Gaztelera-ingelesa	Hermes
Wikipedian oinarritutakoa	Euskara-ingelesa	Hermes + WIL
Wikipedian oinarritutakoa	Euskara-ingelesa	CLEF ResPubliQA

III.4 Taula: NET tresnen esperimentuen laburpena

III.4.1 Hizkuntzen menpeko NET tresna

Hizkuntzen menpeko NET tresna, euskara-gaztelera bikotean oinarrituta diseinatu denez, ebaluazio-corpusa ere hizkuntza horietarako prestatu da. Aurretik III.2.1.3 atalean aipatu den bezala, sistema hau bi ebaluazio-corpusekin ebaluatu dugu. Lehen ebaluaziorako *Elhuyar fundazioa*k luzatutako izen berezien zerrendatik (onomastikatik) erauzitako 557 NE bikoteek osatzen duten corpusa erabili dugu. Eta ondoren, ebaluazio errealagoa burutu nahian, Hermes euskara-gaztelera corpuseko euskarazko 200 NE maizenen eskuzko itzulpenarekin osatutako corpusa erabili dugu.

Esan beharra dago, bi corpusekin egindako ebaluazioetan osagaien sorkuntza eta ordenazio faseetan ez dela murriztapenik jarri. Alegia, NEaren hitz bakoitzeko osagaiz osagaiko itzulpen-moduluan sortzen diren hautagai guztiak erabili dira NE osoaren eraketa moduluan. Eta horiekin guztiekin ordenazio erregelak aplikatu ondoren sortutako NEaren itzulpen oso guztiekin *Web*era jo da hautagai erabiliena zein den jakiteko. Hala izateko, bi esperimentuetarako III.3.1 atalean deskribatutako sistemaren n parametroa balio altu batekin definitu da.

III.5 taulan onomastika taldeko zerrendan oinarritutako copursarekin lehendabiziko ebaluazioan burututako bi esperimentuen emaitzak ikus daitezke: lehen esperimentuan hautagaiak azken zerrendan sartzeko Googleen agerpen bat izatea nahikoa da; eta bigarren esperimentuan, berriz, gutxienez 100 agerpen izateko murriztapena gehitu da hautagai bat itzulpen-hautagaien zerrendan sartzeko. Agerpen-minimo horiek *fr-min* parametroarekin deskribatzen dira III.5 taulan. Taulako x parametroak, berriz, ebaluaziorako erabiliko den hautagai kopurua adierazten du. Alegia, sistemak proposatzen dituen hautagai ordenatuen artean lehen $x$en artean itzulpen egokia badago asmatu egin duela kontsideratuko da, eta bestela huts egin duela.

III.5 taulako emaitzak aztertuz, NEen aukeraketan agerpen atalase-balio minimo bat eskatzean sistemaren estaldura okertzen bada ere, bataz bestean (F_1), sistemak portaera hobea duela esan daiteke. Izan ere, NE gutxiago itzultzen baditu ere, sistemak itzulpenen fidagarritasun maila igo egiten du doitasunean irabaziz.

Onomastika, zerrenda itxia izanik, agertoki oso erreala izango ote ez den zalantza kentzeko, ebaluazio errealagoa egitea erabaki da. Horretarako maiz erabiltzen diren NEak biltzen dituen ebaluazio-corpus bat erabili da. 200 euskara-gaztelera NE bikotez osatutako Hermesen oinarritutako ebaluazio-

fr-min	x	Doitasuna	Estaldura	F_1
1	**1**	% 70,00	% 62,00	% 65,75
100	**1**	% 80,40	% 58,16	% 67,49
1	**3**	% 37,00	% 65,00	% 47,15
100	**3**	% 61,00	% 60,00	% 60,50
1	**10**	% 33,00	% 65,52	% 43,89
100	**10**	% 53,60	% 61,00	% 57,10

III.5 Taula: Hizkuntzen menpeko NET, NEen aukeraketa Google-en oinarrituz, ebaluazioa euskara-gaztelera onomastika ebaluazio-corpusarekin

corpusa, hain zuzen ere.

Aurreko ebaluazioan ez bezala, bigarren ebaluazio honetan, Web-era jotzean bi baliabide desberdin ustiatu ditugu: Google eta Wikipedia. Bi baliabideekin egindako esperimentuetan, sistema x balio berdinekin ebaluatu badugu ere, *fr-min* parametroaren balioak III.6 eta III.7 tauletan ikus daitekeen bezala baliabidearen menpekoak izan dira. Alegia, agerpen kopuru minimoa definitzean, kontsulta-baliabidearen izaera eta ezaugarriak kontuan izan ditugu.

fr-min	x	Doitasuna	Estaldura	F_1
10	**1**	% 73,96	% 62,50	% 67,74
100	**1**	% 75,75	% 62,50	% 68,50
250	**1**	**% 78,71**	**% 61,00**	**% 68,73**
500	**1**	% 79,13	% 55,50	% 65,24
10	**3**	% 60,36	% 67,00	% 63,50
100	**3**	% 64,87	% 66,50	% 65,67
250	**3**	**% 70,65**	**% 65,00**	**% 67,70**
500	**3**	% 71,60	% 58,00	% 61,93
10	**10**	% 54,43	% 67,50	% 60,26
100	**10**	% 61,18	% 67,00	% 63,96
250	**10**	**% 67,52**	**% 65,50**	**% 66,50**
500	**10**	% 68,82	% 58,50	% 63,24

III.6 Taula: Hizkuntzen menpeko NET, NEen aukeraketa Google-en oinarrituz, ebaluazioa euskara-gaztelera Hermes ebaluazio-corpusarekin

Oro har, Al-Onaizan eta Knighten (2002b) emaitzekin konparazio zuzena egin ezin bada ere, lan horretako emaitzak NEen kategoriaren arabera aurkezten baitira, antzeko emaitzak lortzen direla esan daiteke, toki-izenetan % 85eko doitasuna lortzen badute ere, pertsona-izenetan % 64 baita lortzen

duten doitasun tasarik altuena. Erakundeekin, berriz, problematika zailagoa dela eta ez dute ebaluaziorik aurkezten. Bataz bestean, beraz, hizkuntzen menpeko NET sistema honek portaera parekoa eta hainbat kasutan hobea erakusten duela esan dezakegu.

x parametroari erreparatuz gero, antzeman daiteke zenbat eta hautagai gehiago erabili emaitza egokia bilatzeko, sistemak doitasunean galtzen badu ere estalduran irabazi egiten duela. Hala, x handitzean sistemaren F_1ean antzeko emaitzak lortzen dira, baina esan daiteke x handiagoekin estaldura eta doitasun balioen arteko oreka hobea dela. Esperimentatutako x balio desberdinekin *x=1*etik *x=3*ra pasatzean hobekuntza nabarmenena lortzen dela esan dezakegu. Konkretuki x balio aldaketa horretan sistema estalduran 3 eta 4 puntu artean hobetzen da, eta *x=3*tik *x=10*era pasatzean aldiz % 0,5ekoa da.

Agerpen-kopuru minimoa definitzen duen *fr-min* atalase balioak bestetik, doitasunean du eragina. Izan ere, atalase-balio hori handitzean baldintza betetzen duten hautagai-kopurua gutxitu egiten da. Beraz, *fr-min* optimoa aukeratzeko, balio hori doitasuna handitzen laguntzen duen bitartean neurria handitzen jarraitu behar da, eta doitasuna eta estalduraren arteko oreka apurtzen dela ikustean geratu.

Gure esperimentuetan atalase-balio hori, Google-en bilatuz gero, 250 inguruko balioa onena dela ematen du. Are gehiago, balio hau bikoiztean, sistemaren asmatze-tasak ez dira mantentzen, okertu egiten dira. Beraz, 250eko atalase-balioa erabiliko da aurreratzean konparaketak egiteko. Esperimentuak azkartzeko asmoz $x$ren atalasea, berriz, 1ean definituko dugu konparaketetarako. Izan ere, x handitzeak sistemaren eraginkortasunean eragin zuzena du x balio handietarako sistemaren itzulpen-denbora nabarmenki hazten baita.

Aurretik aipatu bezala, NEen aukeraketarako Google ez den beste baliabide batzuk erabiliz sistemaren portaera aztertzeko asmoz, beste bi esperimentu burutu dira, hautagai aukeraketarako iturria aldatuz: Wikipedia entziklopedia eta Hermes corpusetik erauzitako euskara-gaztelera corpus konparagarria erabiliz, hain zuzen ere.

Aukeraketa Google-en egin ordez Wikipedian oinarrituta egitean, sarrera kopuruari dagokionez Wikipedia Google baino baliabide askoz txikiagoa izanik, agerpen-kopuru minimoa (*fr-min*) zehazten duen atalase-balioa 1en ezarri dugu. Esperimentu horien emaitzak III.7 taulan agertzen dira. Bertan ikus daitekeen bezala, esperimentu guztietan Google-eko sistemaren emaitzekin konparatuz gero estaldura jaitsi egiten da. Doitasunean aldiz, ez da hori

gertatzen. Izan ere esperimentu guztietan Google-eko emaitzak doitasunean 1 eta 3 puntu artean gainditzen dira. Hala, Wikipediako sarreren informazioa Google-ekoa bezain esanguratsua edo esanguratsuagoa dela esan daiteke.

fr-min	x	Doitasuna	Estaldura	F_1
1	**1**	% 81,83	% 60,00	% 69,10
1	**3**	% 72,78	% 61,50	% 67,75
1	**10**	**% 68,50**	**% 62,00**	**% 65,10**

III.7 Taula: Hizkuntzen menpeko NET, NEen aukeraketa Wikipedian oinarrituz, ebaluazioa Hermes ebaluazio-corpusarekin

Hautagaien azken aukeraketarako Web-eko baliabideak atzitu beharrean, sortutako hautagaiak corpus konparagarrian kontsultatzean, sistemaren doitasuna % 10an hobetzea lortzen da (ikus III.8 taula). Estalduran aldiz, ez da ia hobekuntzarik ematen.

x=1	fr-min	Doitasuna	Estaldura	F_1
Google	**250**	% 78,71	% 61,00	% 68,73
Wikipedia	**1**	% 81,83	% 60,00	% 69,10
Corpus Konparagarria	1	% 91,85	% 62,00	% 74,00

III.8 Taula: Hizkuntzen menpeko NET, NEen aukeraketa baliabidearen araberako emaitzak *x=1*erako, ebaluazioa euskara-gaztelera Hermes ebaluazio-corpusarekin

Batezbesteko portaera egonkorragoa lortzeko asmoz, azken urratsean, corpus konparagarria eta Web-a konbinatzen dituzten esperimentuak egin ditugu. Konkretuki, bi esperimentu egin dira, non bietan lehenengo urratsa corpus konparagarriko kontsultan oinarritzen den. Kontsulta horretatik agerpenik aurkitu ezean, bigarren urratsean Web-era jotzen da, Google eta Wikipedia iturriak atzituz, hurrenez hurren. Esperimentu horien emaitzak III.9 taulan ikus daitezke.

Bi esperimentuetan antzeman daiteke, corpus konparagarria soilik erabiltzean baino doitasuna okerragoa lortzen duela sistemak bi urratsetako estrategia jarraitzean. Ez da gauza bera gertatzen ordea estaldura neurrian, non % 3 inguruko hobekuntza lortzen den aukeraketarako bi iturriak konbinatzean. Erantzun gehiago lortzen dira, baina doitasunari erreparatuz, ez dira erantzun zuzenak proportzionalki igotzen.

Web iturria	**Doitasuna**	**Estaldura**	**F_1**
Google, 250	% 82,80	% 65,00	% 72,83
Wikipedia, 1	% 83,22	% 64,50	% 72,70

III.9 Taula: Hizkuntzen menpeko NET, NEen aukeraketa Corpus Konparagarria + Web-a, ebaluazioa euskara-gaztelera Hermes ebaluazio-corpusarekin

Hala ere, *fr-min* eta x atalase-balio berdinekin egindako esperimentuetan, baliabide bakarra eta, corpus konparagarri eta Web-a erabiltzen duten bi estrategiekin ebaluatutako sistemen portaera hobetzea lortzen da. Konkretuki, F_1 balioa bi kasuetan ia 4 puntu hobetzen da. Are gehiago, sistema motelagoa den *fr-min* bera baina *x=3* edota *x=10* balioekin ebaluatutako Web-a soilik atzitzen duten sistemen emaitzekin konparatuz gero, F_1en 5 eta 4 puntu arteko hobekuntzak lortzen dira.

Balio baxuko x sistemak hobetzeko beraz, sistema moteltzea ekartzen duen x handitzea baino, aukeraketa urratsean corpus paraleloa eta Web-a konbinatzea bide egokia dela esan daiteke. Izan ere, estrategia horrekin aukeraketa fasean soilik baliabide bat (bereziki Web-a) atzitzen duten x altuko sistema motelagoen emaitzak hobetzea lortzen dira.

Erroreen azterketa egitean, mota desberdinetako erroreak aurki daitezke. Sistemak erabiltzen dituen tresna automatikoetatik hedatutako erroreek lehenengo errore-mota osatzen dute. Akats horiek, sistemaren % 100eko asmatze-tasa lortzea ezinezko egiten dute. Besteak beste, *Eustaggeren* etiketatze morfosintaktikoan nahasten denean eta analisi desegokiak itzultzen dituenean; esaterako *Ukraina Subkarpatika*, *Ukraina Subkarpa* bezala lematizatzen duenean.

Beste errore mota bat, hiztegi elebidunetatik datorrena da. Abiapuntuhizkuntzako hitzak hiztegian sarrera duenean, baina NE osoa eratzeko xedeko hizkuntzako hitz baliokidea ez dagoenean bertan. Hiztegiko sarrera hitzaren adiera posible guztiekin ez dagoenean osatuta, alegia. Errore mota horrekin *Ipar Polo* NEa itzultzen saiatzean topo egiten du sistemak, hain zuzen ere. *Ipar* hitzarekin hiztegira jotzean, *septentrional* itzulpena soilik eskuratuko du eta, beraz, sistemaren proposamen bakarra *Polo Septentrional* izango da, *Polo Norte* NE egokia sortzea ezinezkoa izanik.

Ordenazioari dagokionez, oso errore gutxi sortzen dira urrats honetan. Eta ematen direnean, gehienetan salbuespenen bat duten NEen itzulpenak dira. Adibide bat ikustearren, *Kristal Hautsien Gaua* izenburuaren itzulpe-

na azalduko dugu. *Eustagger*ek *Hautsien* osagaiaren kategoria ADI dela dio, eta horrela da, baina entitatearen barruan adjektibo funtzioa du. Beraz, sistemak jasotzen duen informazioa ez denez esperotakoa, itzuliko duen ordena ez da zuzena izango. Hau konpontzeko, patroietan beste egitura morfosintaktiko bat aurreikusi beharko litzateke: aditza partizipio modura agertzea. Baina salbuespen guztiak ezin dira bildu gramatika batean, beti egongo baita salbuespen berriren bat.

III.4.2 Hizkuntzekiko erdi-independentea den NET tresna

Baliabide zein diseinu ataletan azaldu den bezala, tresna hau bi hizkuntza-bikoteren corpus konparagarrietan oinarrituz garatu da: euskara-gaztelera eta gaztelera-ingelesa, hain zuzen ere. Bikote bakoitzaren portaera aztertzen da jarraian, 200 entitateko ebaluazio-corpus bana erabiliz (ikus III.2.1.3 atala). Euskara-gazteleraren kasuan, hizkuntzen menpeko tresna ebaluatzeko erabili den ebaluazio-corpus berdina erabili dugu. Horri esker, bi estrategiak konparatu ahal izan ditugu, alegia hizkuntzen menpekoa eta erdi-independentea. Bestetik estrategia bera hizkuntza bikote desberdinekin ebaluatzeak estrategiaren hedagarritasuna aztertzea ahalbidetu digu.

Edozein dela hizkuntza-bikotea, sistemak azken urratsean, NEen itzulpen-hautagai aukeraketa fasean, alegia, corpus konparagarrian agerpen gehien dituzten hautagaiak hobetsiko ditu, maiztasun handienekoa itzuliz. Beraz, sistemaren asmatzea kontsideratuko da, agerpen-kopuru handiena duen itzulpena, ebaluazio-corpusean proposatutakoarekin bat datorrenean.

Hizkuntzen menpeko NET tresna euskara-gaztelera ebaluazio-corpus berarekin eta *x=1* eta hautagai aukeraketarako hautagaia corpus konparagarrian eta Web-a kontsultatzen duen estrategiarekin, hain zuzen ere, % 72,8 inguruko F_1 lortzen du (ikus III.9 taula). III.10 taulan agertzen diren euskara-gaztelera neurketak aztertuz ikus daiteke, hizkuntzen erdi-independentea den sistema erabiltzean hizkuntzen menpeko tresnaren emaitza horiek 4 puntu inguru hobetzea lortzen dela, lehenbizikoa informazio azterketa sakonarekin diseinatuta egon arren.

III.8 taulan, hautagai aukeraketan baliabide bakarra erabiltzen duten hizkuntzekiko menpeko NET tresnaren emaitzarik onenekin konparatuz gero, corpus konparagarrian bilaketa egiten denean alegia, emaitzak F_1en 3 puntu inguru hobetzen dira estrategia erdi-independentea erabiltzean. Doitasunean galera nabarmena eman arren (% 91,85tik % 82,02ra jeisten da), estalduran emaitzak 11 puntu gainetik lortzen dira eta.

hizkuntza-bikotea	Dointasuna	Estaldura	F_1
Euskara-gaztelera	% 82,02	% 73,00	% 77,50
Gaztelera-ingelesa	% 73,80	% 62,00	% 67,38

III.10 Taula: Hizkuntzekiko erdi-independentea den NET tresna, ebaluazioa Hermes ebaluazio-corpusarekin

Estrategia bera, gaztelera-ingelesa bikotearekin ebaluatzean aldiz sistemak portaera okerragoa azaltzen du. Galera nabarmen horren arrazoiak bilatzeko asmoz, ebaluazioan oker itzulitako NE guztiak eskuz errebisatu dira. Berrikuspen horretan 26 NE identifikatu dira, non sistemak itzulpen egokia sortu arren, aukeraketa fasean, forma egokiak xede-hizkuntzako corpusean abiapuntu-hizkuntzako NEak baino agerpen-kopuru txikiagoa duten. Hortaz, aukeraketan bi hautagai horien artean sistemak abiapuntuko NEa bueltatzea erabakiko du, hutsegitea eraginez. Esaterako, *Italia* gaztelerazko NEa itzultzean sistemak *Italia* eta *Italy* proposamenak sortuko ditu, eta ingelesezko corpusean lehenbizikoak azkenak baino agerpen-kopuru handiagoa duenez, *Italia* itzuliko du *Italy* beharrean. Forma desegokiek forma egokiak baino agerpen kopuru handiagoa izatearen jatorria, *EFE* agentziako kazetariek egindako errore tipografikoetan egon daiteke. Edozein kasutan, argi dago corpuseko zaratak ebaluazioan eragin zuzena duela, ebaluazio-corpuseko % 13a osatzen baitute 26 NE horiek.

Beraz, corpus konparagarriaren zuzentasunak sistemaren portaeran eragin handia du. Hainbeste, ezen egindako ebaluazioan, nabarmenki hobetuko bailirateke emaitzak, euskara-gaztelera bikoterako lortutako emaitzak gainditzea posible bihurtuz.

III.4.3 Wikipedian oinarritutako NET tresna

Wikipedian oinarritutako NET tresnaren portaera aztertzeko bi ebaluazio-corpus erabili dira:

- *Euskaldunon Egunkari*atik erauzitako 566 NEko corpusa, euskarazko Wikipedian sarrera izateaz gain, ingeleserako WILa ere definituta dutenak biltzen dituena.

- CLEF ResPubliQA atazan definitutako 500 euskara-ingeles galdera elebidunetatik erauzitako 72 NE bikotetako corpusa.

Bien deskribapena ingurune esperimentalean aurki daiteke (ikus III.2.1.3 atala). Hermes ebaluazio-corpusarekin, sistemaren lehenbiziko urratsa, Wikipediako WIL estekak erabiliz itzulpena lortzen saiatzen den urratsa alegia, ez da aplikatzen % 100eko asmatzea lortuko bailuke sistemak, baina ez litzateke ebaluazio erreala izango, esteka horietan oinarrituz eraiki baita Hermes corpus hori. III.11 taulan sistemak Wikipedian oinarritutako corpus horretako NEak itzultzean, fase bakoitzean prozesatutakoaren banaketa agertzen da, guztira fase bakoitzean itzulpena zenbat NEentzat lortu den eta zenbatetan proposamen egokia lortu den adieraziz.

Urratsa	**NE totala**	**Ondo itzulita**
Hiztegitik	10	10
Wikipediatik[1]	366	362
Itzulpen-Prozesutik[2]	49	46
Itzulpenik gabe	75	0

[1] *Sarrerako euskarazko forma ingelesezko Wikipedian bilatuz lortutako NEen itzulpenak*
[2] *Osagaiz osagaiko itzulpena, NEen eraketa eta azkenik aukeraketa fasea aplikatuz itzulitako NEak*

III.11 Taula: Itzulpenen banaketa

Banaketak aztertuz, euskarazko NEaren forma asko ingelesez forma egokitzat jotzen dela antzeman daiteke, asko baitira abiapuntu-forma ingeleseko Wikipedian sarrera dutenak. Erdia baino gehiago direnez urrats horretan itzultzen diren NEak, euskara-ingelesa forma berdina duten NEak, horiek sistemaren oinarrizko neurria mugatzeko erabili dira. Sistemaren oinarrizko neurria eta ebaluazioa III.12 taulan aurkezten dira.

	Doitasuna	**Estaldura**	**F_1**
Oinarrizko neurria	% 59,82	% 59,82	% 59,82
Sistema	% 93,36	% 75,82	% 83,68

III.12 Taula: Wikipedian oinarritutako NET, ebaluazioa euskara-ingelesa Hermes+WIL ebaluazio-corpusarekin

Argi geratzen da III.12 taulan NET tresnak oinarrizko neurria nabarmenki gainditzen duela. Eta beste sistemekin zuzenean konparatu ezin bada ere, emaitza onak direla esan daiteke (F_1 balio altuena lortzen duen tresna da).

Ebaluazioan gertatutako erroreak analizatzean, batzuetan Wikipediako hizkuntzen arteko WIL estekek, NE bera lotzen duten arren, forma bera erlazionatzen ez dutela identifikatu da. Esaterako sistemak *Dorre Bikiak* itzultzean, *Twin Towers* proposamena bueltatzen du. Forma hau ingelesezko Wikipedian sarrera du, baina ebaluazio-corpusa sortzeko erabilitako WIL estekak, *Dorre Bikiak* sarrera *Twin Towers* sarrerarekin erlazionatu beharrean, *Word Trade Center* orriarekin lotzen du, sistemaren asmatzea erroretzat joz.

Portaera hau ikusita, sistema itzulpenerako erabilgarria izateaz gain, Wikipediako WIL esteka berriak gehitzeko lagungarri izan daitekeela esan daiteke.

Beraz, WIL estekak esplotatzen dituen modulua aplikatu gabe, sistemak portaera ona azaltzen duela esan daiteke. Domeinuko NEak itzultzeko garaian eta NEaren itzulpen egokia xede hizkuntzan sarrera duten NEekin ebaluatzean, behintzat.

Oro har, Wikipedian agerpenik ez duten NEen aurrean itzulpen-sistema honen portaera neurtzeko bidean, CLEF ResPubliQA ebaluazio-corpusa erabili da, non corpusa osatzen duten 72 NE bikoteetatik 9k ez duten sarrerarik ingelesezko Wikipedian. Corpus horretako NEak itzultzean, lehenbiziko corpuserako ez bezala, WIL estekak esplotatzen dituen modulua ere erabili da. Izan ere, CLEF ResPubliQA ebaluazio-corpusaren osaketan WIL loturak ez direnez erabili, WIL estekak ere baliatzen dituen sistema osoa ebaluatzeko agertoki egokia dela esan daiteke.

Eratutako NEen itzulpen-hautagai guztien artean, xede-hizkuntzako Wikipedian sarrerarik duen hautagairik topatzen ez denean, sistema bi portaera desberdinetan ebaluatu dugu: isila eta hiztuna. Hortaz, sistemaren erantzunean dago desberdintasuna. Portaera isila esatean, sistemak sortutako itzulpen-hautagai guztien artean ingelesezko Wikipedian agerpenik topatzen ez duenean, ezer ez duela itzultzen erreferentzia egiten zaio, alegia, sistema isilik geratzen denean. Modu hiztuna, aldiz, ingelesezko Wikipedian agerpenik ez duten itzulpenak besterik sortu ez dituenean, eta beraz, irteerako hautagai zerrenda hutsik dagoenean, sistemak euskarazko NEa itzultzeko portaera hartzen duela esan nahi du. Ebaluazio horien emaitzak III.13 taulan daude.

Oinarrizko neurriaren zenbakiak aztertuz, ebaluazio-corpus hau sistemarentzat aurrekoa baino zailagoa dirudiela esan daiteke. Eta zailtasun horren aurrean sistemaren portaera argi azaltzen da, emaitzak nabarmenki okertuz. Hala ere, sistemak oinarrizko neurriarekiko aurkezten duen hobekuntza, aurreko ebaluazio-corpusarekin ematen dena baino handiagoa da, modu isilean

	Doitasuna	Estaldura	F_1
Oinarrizko neurria	% 23,69	% 23,69	% 23,69
Isila	% 92,68	% 52,77	% 67,25
Hiztuna	% 55,50	% 55,50	% 55,50

III.13 Taula: Wikipedian oinarritutako NET, ebaluazioa euskara-ingelesa CLEF ResPubliQA corpusarekin

behintzat.

Galera nabarmenaren jatorria identifikatzeko erroreak analizatzean, hutsegite asko osagaiz osagaiko itzulpen-fasean sistemak hiztegitik eskuratzen dituen itzulpen-proposamenen artean, NEerako itzulpen egokia ez dagoela identifikatu da. Esaterako *Nazio Batuak* ingelesera itzultzeko prozesuan, sistemak hiztegitik *Union* eta *Nation* itzulpenak eskuratzen ditu, itzulpen onak bai, baina ez egokiak NEaren itzulpenerako, alegia. Hortaz, horrek ez du esan nahi hiztegi elebiduna ona ez denik, bai, ordea, NEen itzulpenerako oso egokia ez dela. Errore mota hori hizkuntzen menpeko NET tresnaren errore analisia egitean ere identifikatu da (ikus III.4.1 atala).

Hiztegi elebidunaren moldaketak sistemaren portaera hobetuko duelakoan, hiztegia automatikoki aberasteko saiakerak burutu ditugu. Aberasketa horretan, Wikipedian oinarritutako corpuseko ebaluazioan gaizki itzulitako 84 NEetatik hasiz, NE bikote bakoitzeko euskarazko osagaiei dagokien ingelesezko hitzekin lotzea izan da lehenbiziko urratsa. Korrespondentzia automatikoki lortzeko, lehenbizi euskarazko hitza ingelesezko NEean forma berarekin agertzen den kontsultatu da eta esteka lortu ez denean, hiztegi elebidunera jo da.

Hiztegitik eskuratutako itzulpen-hautagaiak ingelesezko NEean kontrastatu dira eta xedeko NEean hautagaia aurkitzean euskara-ingelesa hitz-bikotearen korrespondentzia esteka sortu da. Korrespondentziarik lortzen ez denean hurrengo elementua tratatzera pasatzen da. Bukaeran, abiapuntu zein xedeko NE bikotean osagai bakarra geratzen bada lotu gabe, orduan bi hitz horiek bata bestearen itzulpena kontsideratuko dira eta hiztegi elebiduna hitz-bikote horrekin aberastuko da. Prozesua behin eta berriz aplikatuko da, iterazio bakoitzean hiztegi aberastua erabiliz, hitz-bikote berririk topatzen ez den arte.

Aberasketa prozesua *Europako Parlamentua - European Parliament* NE bikotearekin errepasatzen da jarraian. Lehenbizi, euskarazko hitzak ingelesezko NEan bilatuz esteka egiterik lortzen ez denez, hiztegi elebidunera jo

behar da. Lematizaziorik aplikatzen ez denez prozesaketa honetan, hiztegi elebidunean *Europako* hitzarentzat ez da itzulpenik aurkituko eta beraz, ez da estekarik egingo. *Parlamentua* hitza bilatzean aldiz, *Parliament* itzulpena eskuratuko da, ingelesezko NEean dagoen osagaietako batekin bat datorrena. Korrespondentzia aurkitzean *Parlamentua* - *Parliament* hitz-bikotearen esteka ezartzen da. Osagai guztiak korritu ondoren, estekarik gabeko *Europako* eta *European* hitzak besterik ez dira geratzen. Beraz, bi hitzak baliokideak direla kontsideratu eta hiztegi elebiduna hitz-bikote berri horrekin aberastuko da.

	Doitasuna	**Estaldura**	**F_1**
Isila	% 92,68	% 52,77	% 67,25
Hiztuna	% 55,50	% 55,50	% 55,50
Isila-aberasketarekin	**% 93,00**	**% 55,50**	**% 69,51**
Hiztuna-aberasketarekin	**% 58,33**	**% 58,33**	**% 58,33**

III.14 Taula: Wikipedian oinarritutako NET tresna hiztegi aberasketarekin, ebaluazioa euskara-ingelesa CLEF ResPubliQA corpusarekin

Hiztegi aberasketa saiakera txiki horrekin, III.14 taulako emaitzetan ikus daitekeen bezala, modu isilean sistemaren % 2ko hobekuntza lortzen da, eta modu hiztunean, berriz, % 3koa. Hortaz, sistemaren portaerak hobetzen duela ematen du, eta pentsa daiteke hiztegi aberasketa sakonagoarekin sistemaren portaera gehiago hobetzea dagoela. Edozein kasutan, ebaluazio-corpus hau oso txikia da, 72 NE soilik, eta aldaketa txikiek eragin handia dute, bai onerako bai txarrerako. Alegia, aberasketa sakonago horretan zarata sortzen bada, kontuan hartu behar da zarata hori kaltegarria izan daitekeela.

III.5 Ondorioak

Euskaraz idatzitako erakunde-, pertsona- eta toki-izenak automatikoki itzultzeko estrategia desberdinak jarraitzen dituzten hiru sistema aurkeztu dira kapitulu honetan, eta hainbatetan tresnak hizkuntza desberdinekin ebaluatu dira. Hiru estrategiak desberdinak izan arren, tresnen modulu nagusien helburuak hiruetan berdinak izan dira:

- NEaren osagaiz osagaiko itzulpen-modulua: transliterazio-erregelak eta hiztegi elebidunak konbinatuz osagaiz osagaiko itzulpenaz arduratzen den modulua.

- Osagaien itzulpen-proposamenekin, NE osoaren itzulpen-hautagaiak eratzeko osagaien ordenaketa modulua: euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa bikoteetako hizkuntzek jarraitzen duten egitura sintaktiko desberdina dela eta, xedeko osagaien itzulpenak xedeko forman posizio egokian kokatzeaz arduratzen den modulua.

- NEen itzulpen-hautagaien aukeraketa: behin hautagaiak sortuta, guztien artean abiapuntuko NEerako proposamen egokiena aukeratzeaz arduratzen den modulua.

Moduluen funtzionalitate nagusiak bat etorri arren, horien diseinua eta inplementazioa hiru tresnetan teknika eta baliabide desberdinak erabilita burutu dira. Hizkuntzen menpeko NET tresnarako, euskara-gaztelera hizkuntzen ezaugarri linguistiko eta sintaktikoetan oinarrituz, NEetako osagaien transliterazio transformazioak zein elementuen ordenazio gramatikak definitu dira, hurrenez hurren, bikotearekiko guztiz menpekoa den tresna bat eraikiz.

Beste bi tresnak ordea, hizkuntzekiko independenteagoa den sistema bat garatzeko helburuarekin, osagaien transliterazio zein elementu ordenaketa moduluak, teknika orokorrekin garatu dira: transliterazio-gramatika, edizio-distantzian oinarritutako karaktere mailako transformazioak burutzen dituen erregelekin osatuz; eta ordenazio patroiak, aldiz, abiapuntuko NEaren osagaiak, xede-hizkuntzako forman edozein posiziotan ager daitezkeela kontsideratzen duen gramatikan definitu dira. Oinarrian, beraz, hizkuntzekiko erdi-independentea den NET tresna eta Wikipedian oinarritutako NET tresna teknika berdinekin eraiki dira, corpus konparagarri desberdinetan oinarrituta, ordea. Lehenbizikoa, Hermes proiektuan etiketatutako hizkuntza desberdinetako egunkarien berrietako NEetan oinarritua dago, euskara-gaztelera zein gaztelera-ingelesa bikoteetarako garatu eta ebaluatu delarik. Hizkuntzen independentzia horri esker, beraz, bikote berri batentzat corpus konparagarria eta hiztegi elebiduna nahikoak izango dira itzulpen-sistema egonkor bat eraikitzeko. Esan ere, tresna bi hizkuntza-bikote horiekin ebaluatu dela, hasiera batean, euskara-ingelesa NEen itzulpenerako gaztelera zubi lanetarako pentsatu zelako.

Wikipedian oinarritutako tresna aldiz, izenak adierazten duen bezala, euskara eta ingelesezko Wikipedian oinarrituta dago, bi hizkuntza horiek sistemaren abiapuntu- eta xede-hizkuntzak direlarik, hurrenez hurren. Wikipedia transformazio eta bilaketa moduluetan corpus konparagarri gisa erabiltzeaz

gain, entziklopediak definituta dituen hizkuntzen arteko estekak (WILak) eta berbideratze-estekak ere erabiltzen ditu sistemak.

Euskara-gaztelera NEen itzulpenerako, beraz, bi estrategia oso desberdin jarraitzen duten bi sistema proposatu ditugu: hizkuntza-ezagutzan oinarritutakoa eta hizkuntzen ezaugarri linguistikoak esplizituki inplementatzen ez dituen hizkuntzen erdi-independentea. Biak corpus berdinarekin ebaluatzean emaitza antzekoak lortzen direla ikusi ahal izan dugu (% 77,5 F_1 inguru) (ikus III.8 eta III.10 taulak). Hizkuntzekin erdi-independente bezala aurkeztu dugun hurbilpena hizkuntzen ezagutza-linguistikoa izan gabe NEak modu fidagarrian itzultzeko sistemak garatzeko metodologia egokia dela frogatu dugu.

Azken hurbilpen horrek hizkuntzekin duen neurri bateko independentziari esker, lan handia egin gabe, sistema gaztelera-ingelesa bikoterako izaera antzekoa duen ebaluazio-corpus batekin ebaluatu ahal izan dugu. Eta baliabideetan agertzen diren hainbat errore zuzenduta, bi bikoteetarako antzeko emaitzak lortzea posible dela ikusi ahal izan dugu.

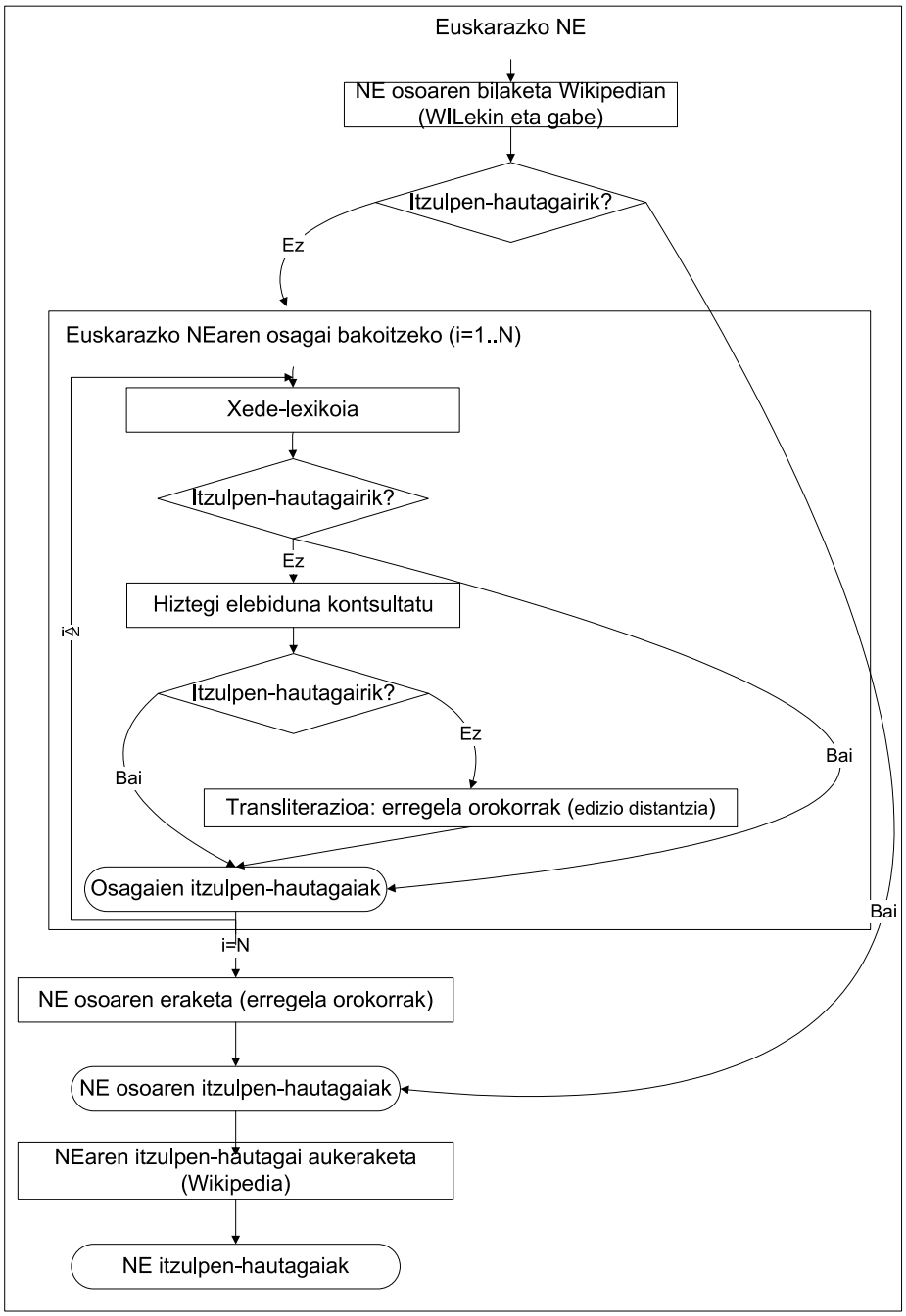
Euskara-ingelesa NEak zuzenean itzultzeko saiakeran, Wikipedian oinarritutako sistema ebaluatu da. Ebaluazioa bi corpus desberdinekin burutu dugu. Lehenbizikoan WILak ustiatuz osatutako corpus bat erabili denez, ebaluazio horretan sistemaren WILak erabiltzeko urratsa ez da erabili, beste sistemen izaera bera mantenduz. Eta bigarrenean aldiz WILak barne, sistema osoa ebaluatu ahal izan dugu. Azken corpus hori ordea, oso txikia denez, emaitzak ezin izan ditugu konfirmatu, bi ebaluazioen artean oso emaitza desberdinak lortzen baitira.

Edozein kasutan, hasiera batean euskara eta ingelesaren arteko itzulpena burutzeko gaztelera zubi lanak egiteko pentsatu bazen ere, itzulpena hizkuntza-zubi gabe egitea hobea dela ikusi dugu, % 78 inguruko asmatzetasako sistemekin lan egitean behintzat. Horiek sortzen dituzten erroreen metaketak azken sistemaren portaeran eragin negatibo nabarmena izango dute, noski.

Euskarazko Wikipedia hazten doan heinean, Wikipedian oinarritutako sistemen estaldura hobetzea espero da. Hazkunde horretan *IXA taldea*k, beste hainbat talderekin batera, ahalegin berezia egiten dabil azken urtean. Konkretuki, *Matxin* itzultzaile automatikoa aplikatzen ari dira Wikipediako hainbat sarreretan, eta ondoren boluntarioen eskuzko edizioarekin automatikoki sortutako erroreak zuzentzen dihardute, Wikipedia erdi-automatikoki handitu nahian.

Edozein kasutan, Wikipedian oinarritutako tresna itzulpen-sistema eraikitzeko baliagarria izateaz gain, sistemak berak, Wikipediako WIL estekak aberasteko bidea izan daitekeela ikusi dugu.

Azkenik, kapitulu honetan egindako Wikipediaren ustiapenari dagokionez, esan dezakegu baliabide horretan oinarrituta euskara-ingelesa hizkuntza bikoterako corpus konparagarria automatikoki osatzeko beste hizkuntza batzuetara hedagarria den estrategia definitu dugula. Alegia, *Yahoo!*k semantikoki etiketatutako ingelesezko Wikipediatik hasita, kapitulu honetan proposatutako metodologia jarraituz edozein beste hizkuntzetara zuzendutako WILak ustiatuz ingelesa eta beste hizkuntza bikoterako guk eratutako corpus konparagarriaren pareko baliabidea sortu daitekeela esan dezakegu.



III.7 Irudia: Wikipedian oinarritutako NET tresnaren diseinua

IV. KAPITULUA

Entitate-izenen Desanbiguazioa

Hitzen adiera-desanbiguazioa azken urteetan arreta handia jaso duen ikerkuntza arloa da. Hizkuntzaren prozesamenduaren munduan, hitzen adiera-desanbiguazioa (HAD) perpaus jakin batean hitz polisemiko batek duen adiera antzematean datzan prozesua da. Esate baterako, *partidu* hitzak adiera ezberdinak izan ditzake testuinguruaren arabera.

IV.0.1 Adibidea *Partidu* hitzaren bi adieren adibideak.

Partidu politiko guztiek bat egin zuten astelehenean *Eibarren* aurkeztutakoarekin.

Ipuruan jokatuko du gaur arratsaldean *Eibarrek* aurtengo azken *partidua*.

Hizkuntza menderatzen duen gizaki batentzat erraza da jakitea IV.0.1 adibideko perpaus bakoitzean *partidu* hitzak duen adiera. Alegia, lehenbiziko esaldian alderdi bati buruz ari dela eta bigarrenean berriz, futbol neurketa bati buruz.

Gauza bera gertatzen da *Eibar* NEaren agerpenarekin, IV.0.1 adibideko perpaus bakoitzean adiera desberdina du. Lehenbiziko esaldiko NEak, herriari egiten dio erreferentzia eta bigarrenekoak berriz futbol taldeari.

Ingelesezko hitzen adiera-desanbiguazioan lan eta ikerketa asko egin badira ere, euskararen kasuan lan gehien egin duen taldea *IXA taldea* izan da. Ingelesezko hitzen adiera-desanbiguazioarekin hasi eta jardun badira ere (Martinez, 2004), euskarazko semantika alorreko hainbat lan argitaratu dituzte. Besteak beste, hitzen adiera-desanbiguazioan (Lopez de Lacalle eta Agirre, 2010) eta Zapirainen (2011) rol semantikoen etiketatze automatikoaren inguruko tesi-lanak.

NEen desanbiguazio alorrean, aldiz, ataza berriagoa izanik, ikerketa gutxiago egin da *IXA taldea*n, eta egindakoek ingelesezko NEak lantzen dituzte (Agirre *et al.*, 2009; Chang *et al.*, 2010). Izan ere, ez *IXA taldea*n, ez hemendik kanpo, ez da euskarazko NEen desanbiguazioaren inguruko lanik ezagutzen.

Oro har, izenen eta NEen artean sinonimia eta polisemia fenomenoek ezinezkoa bihurtzen dute NEen inguruko arrazonamenduak egitea soilik NERC sistemetako irteeretan oinarrituta. Polisemiaren kasuan, izen batek NE desberdinak erreferentzia ditzakeela eragiten du, eta sinonimian, aldiz, izen desberdinek NE bera erreferentzia dezaketela.

Izen berarekin erreferentzia daitezkeen NEek ez dute zertan kategoria berdinekoak izan. Esate baterako, IV.0.1 adibideko bi esaldietan, *Eibar* izenak *Eibar Kirol Elkartea* erakundeari eta *Eibar* toki-izenari egiten die erreferentzia, hurrenez hurren, izen berak bi sailkapen desberdinetako NEak erreferentziatuz.

Sinonimiaren kasuan, *Maialen Lujanbio* NEa aipatzeko izen desberdinak erabiltzen duten adibide dira IV.0.2 adibideko bi esaldiak, lehenbiziko esaldian abizena eta bigarrenean izena erabiliz.

IV.0.2 Adibidea *Maialen Lujanbio* aipatzeko izen desberdinen adibideak.

Bertsolariei dagokienez, haien jarrera nabarmendu zuen *Lujanbiok* bere solasaldian, eta aipatu zuen bertsolarien artean badela joera, beraienak ez diren iritzi eta arrazoiak eurenak balira bezala plazaratzeko bertsotan.

Maialenek hartu zuen hitza eta deskribatu zuen bertsolari bakoitzak zeri egiten dion barre.

NEen alorrean, bi fenomeno horiek, sinonimia eta polisemia ebazteko bi funtzio desberdin bereizten dira: normalizazioa eta desanbiguazioa. Normalizazio funtzioarekin sinonimoa ebazten da eta desanbiguazioarekin, aldiz, polisemia. Normalizazioak, beraz, NE bat aipatzeko erabil daitezkeen forma guztiak identifikatzeaz arduratzen da, eta desanbiguazioa berriz, erreferentzia anbiguo baten adiera edo NE egokia bilatzean datza.

Normalizazioa, testu batean NEak identifikatu ostean, NE bera aipatzen duten forma guztiak lotzeko erabili ohi den funtzionalitatea da. Desanbiguazioak, bestetik, testuko NE anbiguo baten agerpenak, zein NEi egiten dion erreferentzia ebazten du. Azken funtzionalitate horrek, normalizazioak baino konplexutasun handiagoa du, eta tesi-lan honetan, funtzionalitate hori euskarazko NEetan nola ebatzi aztertu dugu. Konkretuki, euskarazko testuetako

NE anbiguoen agerpenak ezagutza-base bateko sarrerekin lotzeko problema nola ebatzi landu dugu, ezagutza-base gisa euskarazko Wikipedia erabiliz. NEen desanbiguazio ataza (NED), beraz, testu bateko edozein izen edota NEen agerpeni dagokion adiera (ezagutza-baseko sarrera) egokia antzematean datza.

Desanbiguazioa modu fidagarrian ebazten duten NED sistemek, hortaz, HPko prozesu-katean txertatu eta NEen inguruko arrazonamendu aplikazioak garatu ahal izatea ahalbidetuko dute. Gaur eguneko erabiltzaileen informazio eskakizun eta beharrak hobeto betetzeko bidea irekiz.

NEen adiera esleipen hau automatikoki egitea ordea ez da lan erraza. Ondorengo puntuan egindako ahalegin desberdinak aurkezten dira.

Kapitulu honetan lehenbizi, lana kokatu eta NEen adiera-desanbiguazio teknika eta sistemen errepaso bibliografiko bat egingo da, eta ondoren, euskarazko NED burutzeko erabilitako ingurune esperimentala aurkeztuko da. Jarraian diseinatutako esperimentuen deskribapena egingo da eta azkenik hauekin lortutako emaitzak eta ondorioak aurkeztuko dira.

IV.1 Lanaren kokapena

Azken urteetan SemEval-en[1] egindako *Web People Search task* (WePS)[2] ataza (Artiles *et al.*, 2007) edo TAC-en egindako *Knowledge Base Population* (TAC-KBP) (McNamee eta Dang, 2009; Ji *et al.*, 2010) bezalako atazei esker, NEen desanbiguazioak indarra hartu duen ataza izan da, ingelesa, gaztelera, frantsesa eta beste hainbat hizkuntzatarako NED sistemen diseinua bultzatuz. Ataza horietan, esfortzu handia egin da desanbiguaziorako baliabideak sortzen, bereziki ingeleserako.

NEen desanbiguazioa HPko hainbat aplikaziotan oso erabilgarria izan daiteke. Esaterako, Web-ean oinarritutako IR (WIR) sistemak hobekuntza hori balia dezaketen tresnen adibideak izan daitezke. Khalid *et al.* (2008) egileek beraien lanean deskribatzen duten bezala, Web galderen artean, galderen erregistroen (log-en) arabera, NEen inguruko galdera-motak portzentaje altua baitute. Hala, sistema horiek esaterako *Eibar* NEari buruzko galderak egitean, Web orri multzo bat itzuli beharrean, NEaren inguruko gertakizun edota ezaugarri konkretuak proposatu ahal izango dituzte beste erantzun posible bat bezala. Horrelako emaitza orriak sortzeko, aldiz, sistemak lehenik

[1] http://www.senseval.org/

[2] `http://nlp.uned.es/weps/`

eta behin galderetan agertzen diren NEen anbiguotasuna ebatzi beharra du. Are gehiago, eraginkortasuna eta erabilgarritasuna handitu nahi izanez gero, emaitzak agerpenen NE posible guztien gertakizun eta ezaugarriak nahastuta aurkeztu beharrean, adieraren arabera multzokatuta aurkeztea garrantzitsua da. WIR sistema mota horrekin, erabiltzaileak modu errazagoan eta intuitiboagoan aztertu ahal izango du informazioa.

NEen desanbiguazioak abantaila handia ekar diezazkiokeen HPko beste arlo bat Web-ean oinarritutako IE da. Kilgarriff eta Grefenstettek (2003) IE aplikazioak hobetzeko algoritmoen konplexutasun maila handitu beharrean hobekuntzak datu multzo handien erabilerarekin etor daitezkeela azaltzen dute. Lan horretan deskribatzen duten IE sistemaren helburua NEen arteko erlazioak erauztea da. Hala izanik, sistemaren hobekuntzarako NEen desanbiguazioa lagungarri suerta daitekeela argi ikus daiteke. Batetik adiera desberdinetako agerpenak identifikatu ahal izateko eta, bestetik, erlazio okerrak ekiditeko baliagarri izan daiteke.

Ondorioz, NED ebazpenak HPko alor desberdinetan ekar ditzakeen abantailak argi daudela esan daiteke.

Gaur egun euskarazko NEen desanbiguazioa ebazten duen sistemarik ez dago, ordea. Eta hutsune hori betetzeko asmoz, euskarazko NEak desanbiguatzeko egindako esperimentuei buruz hitz egingo da kapitulu honetan. Horretarako, batetik euskarazko Wikipedia erabili eta, bestetik, beste hizkuntzetan NED ataza ebazten portaera onena duten algoritmoekin egindako esperimentuak aurkeztuko dira.

Wikipedia txikiekin lan egitean NED ebazpenean arazo desberdinak aurkez daitezke, horretan ere gure ekarpena egin nahi izan dugu. Alegia, Wikipedia txikiekin lan egitean NED atazaren zailtasun gehigarriak nola ebatzi eta ingelesezko metodo onenen portaera nola aldatzen den aztertu nahi izan dugu. Alegia, euskara bezalako hizkuntza gutxituen Wikipedia txikiak erabili edo Wikipedia handi batekin lan egitean ondoren zerrendatzen diren ezaugarrien eragina aztertu nahi izan dugu:

- *Anbiguotasun txikiagoa*: artikulu kopurua txikiagoa izanik, alegia ezagutza-basean deskribatzen diren NEen kopurua txikiagoa izanik, NE anbiguo baten desanbiguazioan adiera posibleen kopurua txikiagoa izango da.

- *NE gehiago ezagatzu-basean sarrerarik gabe*: aurreko ezaugarriarekin lotuta, ezagutza-basean deskribatzen diren NEen kopurua txikia-

goa izango da. Alegia, ebaluazio-corpuseko NE gehiago egongo dira ezagutza-basean sarrerarik ez dutenak. Ezaugarri hau ezagutza-basearen zein ebaluazio-corpusaren menpekoa izango da, noski.

- *Artikulu laburragoak*: artikulu laburren portzentajea altuagoa izanik, desanbiguaziorako eskuragarri dagoen informazioaren aberastasuna txikiagoa izango da.

- *Artikuluen arteko esteka gutxiago*: aurreko ezaugarriari lotuta informazio gutxiago egonik ezagutza-baseko sarreren arteko lotura gutxiago egotea ere espero daiteke.

Ezaugarri horiek datuetan erakusteko asmoz, Bunescu eta Pascaren (2006) lanean (ikus IV.2.1 atala) erabiltzen duten 2006 urteko ingeleseko Wikipediako datuak eta garai bereko euskarazko datuak deskribatzen dira jarraian. Ingelesezko Wikipedia osoan 1.187.839 artikulutik abiatuz, 100 hitz baino gutxiago eta 5 sarrera zein irteera esteka baino gutxiago dituzten artikuluak ezabatuz 241.393 artikulu lortzen dituzte. Euskarazko Wikipedian ordea, inolako garbiketa egin gabe, 63.106 artikulu besterik ez daude erabilgarri.

Bestetik, Wikipedia grafo bezala irudikatu eta konparatuz gero, ingeleseko 2008ko bertsio garbiarekin (Yeh *et al.*, 2009), 1.002.411 nodo eta 30.939.288 arkuko grafoa lortzen bada ere, tesi-lan honetan erabilitako euskarazko 2006ko bertsioarekin, 63.106 nodo eta 458.026 arkuko grafoa besterik ez da lortzen. Baliabide tamaina hain desberdina izanik, eta etiketatutako datu sorta handia lortzeko zailtasunak direla eta, euskarazko NEak desanbiguatzeko bibliografian portaera onena aurkeztu duten algoritmo ez-gainbegiratuekin sistemak eratzea erabaki dugu. Agertoki hori baliabide murritzeko hizkuntza baterako errealagoa dela uste dugu.

IV.2 NEen desanbiguaziorako teknikak

NEen desanbiguazioa, NEen agerpenak kanpoko ezagutza-base bateko NEekin lotzen dituen ataza gisa defini daiteke. NEen definizioak biltzen dituen ezagutza-basea garatzea ordea, ez da lan erraza. Hitzen adiera-desanbiguazioan, adierak eta bere definizioak biltzen dituen WordNet (Fellbaum, 1998) baliabidearen parekoa litzateke. Dena dela, pertsona askoren urteetako lana beharrezkoa izan da baliabide horren garapenerako.

*WordNet*en NEen presentzia oso txikia dela medio, NEen desanbiguaziorako ez da baliabide egokia. Hori dela eta, gaur egun, NED alorrean baliabiderik erabiliena Wikipedia da.

2007 urtean pertsona-izenen desanbiguazioa ebazteko SemEval-en WePS (Artiles *et al.*, 2007) ataza burutu bazen ere, NE mota guztiak kontsideratuz NED alorreko lehen ataza partekatu formalak 2009 eta 2010 urteetan burutu dira. McNamee eta Dang (2009) eta Ji *et al.* (2010) egileen lanek TAC-KBP ataza partekatu horien laburpenak aurkezten dituzte. Bertara aurkeztutako sistemek, NEen agerpen anbiguo bat eta bere testuingurua emanik, ezagutza-base bateko zein entitate errepresentatzen duen ebazten dute. Bi ataza partekatuetako ezagutza-baseak, ingelesezko Wikipediako sarrera multzoekin osatuta daude, hain zuzen ere.

McNamee eta Dangen (2009) lanean deskribatzen den bezala, dokumentu multzo bat eta ezagutza-base bat emanik, dokumentuetako NEen agerpen bakoitza ezagutza-baseko zein NE erreferentziatzen duen (baten bat erreferentziazen badu) ebazten duen ataza da NEen desanbiguazioa. Horrela izanik, agerpenak eta ezagutza-basea lotzeko NED sistemetan bi urrats nagusi identifika daitezke: hautagaien identifikazioa eta NEen hautagaien desanbiguazioa. Hautagaien identifikazioan, NE zehatz baten agerpen konkretu bat ezagutza-baseko zein sarrerarekin erlaziona daitekeen identifikatzen da. Desanbiguazio urratsean, berriz, identifikatutako ezagutza-baseko hautagai guztien artean, agerpenerako egokia den sarrera hautatzean dago funtsa. Oro har, NED sistemen desberdintasun nagusia bigarren urratsa ebazteko metodoa edo estrategia izan ohi da.

Ebazpen posible bat desanbiguazioa sailkapen problema bat balitz bezala ebaztea da. HADan ohikoa izan den estrategia, hain zuzen ere. Horretarako, NE bakoitzeko sailkatzaile bat sortzen da, eremu honetan sailkapen posibleak NE bakoitzari ezagutza-basean definituta dauden eta egoki dakizkiokeen sarrerak izango dira. Ebazpen horretarako ezaugarri-bektoreak erabiliz gero, estrategia hau Wikipediako sarrera bakoitzerako pisu bektore bat ikastearen parekoa litzateke, non ikasi beharreko pisu kopurua Wikipedia osoaren hiztegiaren hitz kopuruarekin bat etorriko den.

Beste ebazpen posible bat, desanbiguazioa ordenazio edo *ranking* problema gisa ebaztea da, non NEen agerpenen ezagutza-baseko sarrera posibleen artean adiera egokia hautatzeko, hautagai eta agerpenaren artean antzekotasun funtzio bat erabiltzen den. Funtzioaren balioa maximizatzen duen hautagaia hartuko da NEaren agerpenaren adiera egokia gisa.

Edozein dela ebazpen-bidea, desanbiguatu nahi den NEaren adiera

ezagutza-basean ez egotea gerta daiteke. Agerpen horiek NIL kasuak bezala ezagutzen dira eta, aurrerago deskribatzen den bezala (ikus IV.2.3), tratamendu berezia jaso ohi dute NED sistemetan.

Oro har, NED sistemak, pisaketa funtzioaren definizioan berezituko dira, bi multzo nagusi bereizi daitezkeelarik: batetik, gainbegiratutako metodoekin modelatutako funtzioak dituzten sistemak, non NE eta ezagutza-baseko sarrera pare bakoitza sailkapenerako instantzia kontsideratzen den, eta ikasketarako etiketatutako datu-sorta beharrezkoa den; eta bestetik gainbegiratu gabeko edota erdi-gainbegiratutako estrategiak erabiliz definitutako sistemak, non etiketatutako informazio minimoa parametroak eta atalase balioak definitzeko erabiltzen diren, eta beraz, antzekotasun funtzioa etiketatu gabeko testuinguruan oinarritzen den.

Gure lanerako bigarren aukera kontsideratu dugu bietatik interesgarriena, baliabide urriko hizkuntzetarako ondorioak ateratzea baita gure helburuetariko bat, eta horrelakoetan baliabideak (datuak zein pertsonak) urriak izan ohi dira. Jarraian, estrategi gainbegiratu, ez-gainbegiratu eta erdi-gainbegiratuetan oinarritutako azken urteetako NED sistemen aurkezpena egiten da.

IV.2.1 Sistema gainbegiratuak

NEen desanbiguazio atazan metodo gainbegiratuetan oinarritutako sistemak garatzeko etiketatutako datu asko behar badira ere, metodologia hau erabiliz badira hainbat sistema, bereziki TAC-KBP 2009 eta 2010 urteetako ataza partekatuaren testuinguruan garatutakoak. Konkretuki, 2009 urtean aurkeztutako 13 sistemetatik 5 sistema onenen arteko 4 teknika gainbegiratuetan oinarritutakoak dira.

Pertsona-izenen desanbiguaziorako teknika gainbegiratuak eta Wikipedia erabiltzen duen lehen argitalpenetako bat Bunescu eta Pascaren (2006) lana da. Bertan, 2006ko ingelesezko Wikipedia bertsioko sarreren testuak, kategoriak, berbideratze-orriak eta sarreren arteko estekak erabiliz gainbegiratutako SVM eredu bat nola entrenatu azaltzen da, sistemaren helburua sarrerako testu bateko NE baten agerpen bat Wikipediako sarrera batekin lotzea izanik. Pertsona-izenekin aurkezten duten ebaluazioan % 55,4 eta % 88,4 arteko estaldura (ikus IV.3.3 atala) lortzen dute.

TAC-KBP 2009 edizioan, bigarren postua lortu zuen sistemak (Li *et al.*, 2009) NE anbiguo bat ezagutza-baseko sarrera batekin lotzeko, pisuen ikasketa eta Naïve Bayes Hedatua erabiltzen ditu. Bi urrats nagusi bereizten dira

desanbiguazio-fasean: lehenbizi NEaren desanbiguazio hautagaiak pisatzen dira eta 2. urratsean, pisu altuena lortu duen hautagaia adiera egokia den edo ez ebazteko sailkatzaile bitar bat aplikatzen da. Sistemak batez bestean % 80,33ko asmatze-tasa[3] lortzen du.

McNamee *et al.*en (2009) egileen sistemak berriz, SVM algoritmoaren ordenaziorako aldaera erabiltzen du NE baten agerpen bat eta ezagutza-baseko sarreren arteko antzekotasuna neurtzeko. SVM eredua ikasteko, 200 atributu atomiko erabiltzen dituzte, NE anbiguoaren agerpenaren testuinguruko zein ezagutza-baseko sarrerako ezaugarriak konbinatuz.

Agirre *et al.* (2009) egileen lanean, hiru desanbiguazio sistema proposatzen dira. Horietako bat, gainbegiratua den bakarra, SVM sailkatzaileak entrenatuz NE anbiguo eta ezagutza-baseko adiera posible bakoitzeko multi-sailkatzaile batean oinarritzen da. Bigarren sistemak, ezagutza-basea IR sistema bateko baliabidea bailitz kontsideratuz, ezagutza-baseko sarrerak zein NE anbiguoen dokumentuak bektore-espazioan errepresentatuz eta horien artean IRn ohikoa den bektoreen arteko kosinu konparaketarekin desanbiguazioa burutzen du. Eta hirugarrenak HADan oso ezaguna den maiztasun handieneko adiera itzultzeko heuristikoan oinarritzen da. Hiru estrategien konbinazioak sistema bakoitza bere aldetik aplikatuz baino desanbiguazio hobeak lortzen dituzte, eta konbinazio horri esker 2009ko TAC-KBP edizioan 4. sistema onena da, % 78,84ko asmatze-tasarekin.

Lan horretan, NEen desanbiguazio sistema desberdinen eta horien konbinaketaz gain, desanbiguaziorako ezagutza-baseko hautagaiak identifikatzeko karaktere-kate mailako hiztegi baten sorkuntza prozesuaren deskribapen zehatza egiten da. Azken batean, NED sistema batek lehenbizi agerpen baten ezagutza-baseko adiera hautagaiak identifikatzeko ahalmena izan behar du, ondoren, hautagai horien artean egokiena hautatu ahal izateko.

Sistemak, NE baten agerpenaren adiera hautagaiak bilatzeko, karaktere-kate mailako hiztegi hori erabiltzen du, non NE bat ezagutza-baseko sarrera posibleekin definitzen den, ezagutza-baseko adiera hautagaiekin, alegia. Hiztegia, beraz, sistemako funtsezko baliabidea da, adiera egokia ez bada agertzen NE anbiguoaren definizioan sistemak ez baitu sekula erantzun egokia itzuliko. Hiztegiaren eraketarako, besteak beste, Wikipediako sarreren tituluak eta berbideratze-estekak erabiltzen dituzte.

Hiztegiaren eraketa fasean, Wikipediaren markaketa-sistemako artikuluen

[3]TAC-KBP ataza partekatuko *micro-average accuracy* bezala ulertu behar da TAC-KBP atazan aurkeztutako sistemen asmatze-tasari buruz hitz egitean.

arteko erreferentzien loturak baliatuz, karaktere-kate bat (NEaren agerpen bat) artikulu jakin batera lotuta zenbat aldiz agertzen den kalkulatzen dute. Informazio hori hiztegian ezagutza-baseko sarrera eta NEaren agerpenaren sarrera bikotearekin batera gordetzen dute. Pisu horietan oinarritzen da, hain zuzen ere, lan horretan proposatzen den maiztasun handieneko adiera-desanbiguazio heuristikoa. Aurrerago, ingurune esperimentala deskribatzean (ikus IV.3.1.2 atala) azalduko dugun bezala, antzeko hurbilpena egin dugu euskarazko karaktere-kate mailako hiztegia sortu eta esplotatzeko.

2010 urteko TAC-KBP edizioan metodo ez-gainbegiratuek edo erdi-gainbegiratuek lehen postuak lortu bazituzten ere, gainbegiratutakoak, 4., 5., 6. eta 10. postuetan geratu ziren (McNamee, 2010; Chang *et al.*, 2010; Zhang *et al.*, 2010; de Pablo-Sánchez *et al.*, 2010). Guztiek azkenak ezik, desanbiguazioa erregresio logistikoekin ebazten duela. Gainerakoek SVM algoritmoa erabiltzen dute desanbiguazio ereduak ikasteko.

Oro har, TAC-KBP ataza partekatuetan lortutako emaitzak kontuan izanik, gainbegiratutako estrategiekin emaitza onak lor daitezkeela ondoriozta daiteke. Ahaztu gabe, noski, mota horretako sistemen garapenerako beharrezko baliabidea etiketatutako datuak direla, eta hori askotan lortzen oso zaila eta hainbatetan ezinezkoa dela.

IV.2.2 Sistema ez-gainbegiratuak eta erdi-gainbegiratuak

NEen desanbiguazio arloko lehen lanak, nagusiki ikasketarako datu faltak eraginda, metodo ez-gainbegiratuetan oinarritutakoak dira. Sistema horietan gaur egun hain erabilia den Wikipedia ez zen erabiltzen, oraindik garatu gabe baitzegoen, eta NE mota guztiak beharrean, pertsona-izenak soilik desanbiguatzea zuten helburu.

Gehienak, NEen agerpen anbiguo bat eta hautagai-zerrenda bateko adiera egokia hautatzeko, bektore-espazioan oinarritutako estrategiekin pisu edo antzekotasun-neurri altuena lortzen duen hautagaia itzuliz ebazten dute ataza. Pertsona-izenak desanbiguatzeko ezagutzen den lehen sistema (Bagga eta Baldwin, 1998) estrategia hori jarraituz garatu zen, hain zuzen ere. Sistema horrek, bi espresiok NE bera aipatzen duten ebazteko, bien agerpenen testuinguruetako hitzak erabiliz, agerpenak bektore-espazioan irudikatu eta bektoreen arteko kosinua kalkulatzen du. Antzekotasun-neurria sistemak definitutako atalase balioaren gainetik badago, bi agerpenak NE bera erreferentziatzen dutela esaten da. Sistema ebaluatzeko, *John Smith* NEaren 35 agerpen dituen corpus bat erabiltzen dute. Ebaluazioan sistemak % 85eko

F_1 lortzen du.

Mann eta Yarowsky (2003) eta Niu *et al.* (2004) egileen lanetan bektore-espazioaren errepresentazioa informazio biografikoarekin zabaltzen dute, lehenbizikoan estrategia ez-gainbegiratua aplikatuz eta bigarrenean berriz, metodo erdi-gainbegiratua erabiliz.

Lehen sistemak, biografian landutako atributu sorta zabal baten gaineko multzokatze (edo *clustering*) ez-gainbegiratua aplikatzen du, eta ondoren multzokatutako NEak zatitu eta datu biografikoetako dagokien espresioekin lotzen ditu.

Niu *et al.* (2004) egileen lanak, berriz, dokumentu arteko pertsona-izenak desanbiguatzeko hitzen agerkidetzekin batera, informazio erauzketako teknikak integratzen dituen algoritmo berri bat proposatzen du. Horretarako, lehenbizi gainbegiratze minimoarekin ikasketa-eskema bat garatzen dute, eta ondoren ezaugarri heterogeneoetan oinarrituta antzekotasunaren probabilitate banaketak irudikatzeko entropia maximoan oinarritutako eredua erabiltzen dute. Azkenik, suberaketa estatistikoa (*statistical annealing*) aplikatzen dute antzekotasunak konparatzeko eta NEen korreferentziak ebazteko. Bagga eta Baldwinen (1998) lanean deskribatutako sistema lan honetan aurkezten den ebaluazio-corpusarekin ebaluatzean, Niu *et al.* (2004) egileen sistemak baino portaera okerragoa duela erakusten du. Izan ere, Niu *et al.* (2004) egileek proposatzen duten sistemak, ebaluazio horretan Bagga eta Baldwinen (1998) sistema orokorrean F_1 25 puntutan gainditzen du.

Hassell *et al.* (2006) egileen sistemak, Mann eta Yarowskyren (2003) sistemaren antzera, erreferentzia biografikoak gordetzen ditu afiliazioa, interesak, posta-helbidea eta antzeko ezaugarriak erauzi eta dagozkion pertsona-izenetan atributu gisa esleitzeko. Horrela, NEaren agerpen baten adiera egokia hautatzeko, hautagai bakoitzaren agerpenaren testuinguruarekin konparatzen da antzekotasun-pisu bat esleituz. Pisu altueneko hautagaia NEaren agerpenaren adiera gisa itzultzen du. Sistemaren balorazioa *DBWorld*[4] webguneko hogei mezutako 758 ikerlari-izen desanbiguatzeko egindako saioarekin aurkezten dute, sistemak % 97,1 eta % 79,4 doitasuna eta estaldura lortuz, hurrenez hurren. Datu-base biografiko bateko informazioa eta lankidetzak erabili ordez, sozialki erlazionatutako pertsona-izenak desanbiguatzeko Web orrien esteketan oinarritutako lanak (Bekkerman eta McCallum, 2005) ere aurki daitezke.

Gooi eta Allanen (2004) lanean, dokumentuen arteko korreferentziak

[4] http://www.cs.wisc.edu/dbworld/

ebazteko multzokatze teknika eta metodo estatistiko desberdinak ebaluatzen dira. Baina lanaren ekarpen nagusiena, NED atazako sistemak ebaluatzeko corpusak automatikoki eratzeko proposatzen duten teknika da. Teknika hori, HAD atazako ebaluazio-corpusak automatikoki sortzeko erabilitako estrategiaren (Kulkarni, 2005) oso antzekoa da. Hitzen adierak artifizialki etiketatzen dituzte ebaluazio-corpusak sortzeko.

Oro har, Gooi eta Allanen (2004) laneko teknikak, *person-x* espresioa erabiltzen du corpus bateko pertsona-izenen agerpenak ordezkatzeko. Horrela, *person-x* espresioa NE baten agerpen anbiguoa kontsideratuz, agerpen anbiguoetarako desanbiguazio adierak (testuko NEa) ezagunak dituen corpusa lortzen dute. Estrategia hainbat corpus desberdinetan aplikatzen dute, ebaluazio-corpus desberdinak eskuratuz. Eta ebaluazio-corpus horietan, hain zuzen ere, NED ataza ebazteko informazio erauzketarako teknika desberdinak ebaluatzen dituzte.

Artiles *et al.* (2005) egileek pertsona-izenak desanbiguatzeko sistemak ebaluatzeko corpusa eratzeko beste saiakera bat aurkezten dute. Egileek ingelesezko 10 izen usuenak hautatu dituzte eta izen bakoitzeko 100 web-orri bildu eta NEaren adieraren arabera multzokatzen dituzte, ebaluazio-corpus bat eratuz. Corpus hori 2007 urtean burututako WePS (Artiles *et al.*, 2007) ataza partekaturako ebaluazio-corpusa izan zen. Ataza partekatu horretara aurkeztatuko 16 sistemetatik gehienak, ezaugarri mota desberdinak erabiliz metodo ez-gainbegiratuetan oinarritutakoak izan ziren, % 40 eta % 78 arteko asmatze-tasak lortuz.

Cucerzanen (2007) lanean edozein NEen agerpen Wikipediako sarrera batekin lotzeko desanbiguazio-paradigma formalizatzen da. Aurkezten den sistema Bunescu eta Pascaren (2006) Wikipedia informazioan oinarritutako ezaugarrien antzekoak erabiltzen baditu ere, oraingoan ezaugarriak Wikipediako NEak bektore-espazioan adierazteko erabiltzen dira eta ez SVM sailkatzaileak eraikitzeko. Horrela, edozein NE anbiguoren agerpen bere testuinguruko hitzak erabiliz, Wikipedia sarreren bektore-espazio berean irudikatu eta Wikipediako sarreren bektoreekin konparatuz desanbiguazioa burutzen da. Sistemak Wikipediako kategoria-sistema ere baliatzen du adiera hautaketa burutzeko, % 88 eta % 91 arteko asmatze-tasak lortuz.

Urte berean, ez NEen desanbiguazioaren alorrean baina bai Wikipedia oinarritzat harturik, dokumentu zein hitzen arteko erlazio semantikoak neurtzeko ESA algoritmoa aurkeztu zen (Gabrilovich eta Markovitch, 2007). Semantika alorrean algoritmo onenen artean kokatzen da gaur egun eta tesi-lan honetan duen garrantzia dela eta aurrerago deskribapen zehatza emango da

(ikus IV.3.2.3 atala). Oinarrian, Wikipedia aurreprozesatzen dute alderantzizko indizea kalkulatuz, Wikipedian agertzen den hitz bakoitza Wikipediako bektore-espazioan adieraziz. Horrela, aurrerantzean, bi hitzen arteko erlazio semantikoa kalkulatzeko nahikoa izango da dagozkion bi bektoreak konparatzea.

Antzeko metodoa deskribatzen dute Strube eta Ponzetok (2006) beraien lanean. Bertan, edozein bi hitzen arteko erlazio semantikoa neurtzeko Wikipedia erabiltzen dute. Sistemak, lehenbizi Wikipediako sarreren tituluetan bilatzen ditu hitzak. Ondoren hitzak barnean dituzten tituluetako Wikipedia sarreren deskribapenetan oinarrituz, artikuluen arteko antzekotasuna neurtzen du eta azken urratsean, artikuluen arteko distantzia konputatuz antzekotasun neurri finala kalkulatzen du. Azken kalkulu horretan sistemak Wikipediako artikuluen sailkapen zuhaitza erabiltzen du.

2009 eta 2010 edizioetako TAC-KBP ataza partekatuetan hainbat sistema ez-gainbegiratu aurkeztu dira NE anbiguoen agerpenak eta Wikipedia sarrerak lotzeko helburuarekin ere. Varma *et al.* (2009) egileen lanean esaterako, ataza Naïve Bayes algoritmoarekin desanbiguazio gainbegiratua eta IR sistema ez-gainbegiratua garatu eta konparatzen dira, azkenak % 75,39ko asmatze-tasarekin lehenengoak baino emaitza hobeak lortuz.

2009ko edizioko IR estrategiaren portaera ona baliatuz, 2010 urteko ataza partekaturako, IR estrategia mantendu eta beste urrats bat gehituz sistema aurkeztu dute Bysani *et al.* (2010) egileek, aurkeztutako 16 sistemen artean % 82,17ko asmatze-tasarekin 2. postua lortuz. Konkretuki, sistemak *tf-idf* haztaketan oinarritutako bektoreen kosinu konparaketa burutzen du hautagaien lehen ordenazioa egiteko. Eta ondoren, sarrerako NE anbiguoarekin kosinu bidez antzekotasun handiena lortzen duten hiru hautagaiekin, galdera-hedapena egiten du. Ondoren berrordenaketa burutzen du, azken adiera-hautaketarako lehen ordenazioko eta bigarren berrordenaketaren pisuen konbinazio lineala erabiliz.

TAC-KBP 2010 edizioko hirugarren sistema onenak (Radford *et al.*, 2010) Cucerzanen (2007) sistema abiapuntutzat hartzen du desanbiguazioko hautagaien lehen ordenazioa burutzeko eta ondoren, Wikipediako esteken egitura baliatuz, grafo bidezko pisaketan oinarritutako berrordenaketa aplikatzen du hautagaien arteko adiera hautatzeko. Desanbiguazio estrategia horrekin, sistemak % 82ko asmatze-tasa lortzen du.

NEen desanbiguaziorako grafoetan oinarritutako sistema da Gentile *et al.* (2010) egileen lanean deskribatzen dena ere. Lan horretan, NEen anbiguotasuna ebazteko Wikipedia grafo gisa erabiliz erlazio semantikoa kal-

kulatzen da, % 91,46 eta % 89,83 asmatze-tasak lortuz ebaluatutako bi corpusetan. Yeh *et al.* (2009) egileen lanean Wikipedia grafoaren gainean erlazio semantikoa ebazteko *PageRank Pertsonalizatua* (Haveliwala, 2002) nola aplikatu azaltzen da. Fernández *et al.* (2010) egileen sistemak berriz, Wikipediaren egitura NEen agerkidetza balioztatzeko erabiltzen du, horretan ere *PageRank* algoritmoa erabiliz.

NED ataza ebazteko, beraz, *bag-of-words* ereduan oinarritutako sistema gainbegiratuak zein ez-gainbegiratuak, edota grafoetan oinarritutako algoritmo ugari erabili izan dira, gehienak ingelesezko Wikipedia moduko ezagutza-base handietan oinarrituta edota ebaluatuta.

IV.2.3 NIL kasuak

Aurreko ataletan ikusi dugun bezala, landu diren NED sistema gehienek, NEen agerpen anbiguo bat ezagutza-base bateko sarrera batekin lotzea dute helburu. Horrelako agertokian, ezinbestekoa da kontuan hartzea NEen agerpen anbiguo baten adierak, ezagutza-basean sarrerarik ez izatea gerta daitekeela. Alegia, NED sistemetan NE baten agerpen anbiguo bati ezagutza-baseko sarrerarik ezin esleitzeko egoera aurreikusi behar da. Egoera horren aurrean TAC-KBP ataza partekatuetan sistemak NIL erantzuna eman beharko duela definitu zen. Gertaera horri aurre egiteko, TAC-KBPra aurkeztutako sistemek estrategia desberdinak erabiltzen dituzte.

Gainbegiratutako sistemen artean proposatutako estrategien artean bi nagusi identifika daitezke: NIL erantzuna adiera posible gisa trebatutako sistemak, eta NIL adiera kontsideratu gabe trebatutakoak.

NED ataza ebaztean NIL erantzuna adiera posible kontsideratuz trebatutako sistemaren adibidea, McNamee *et al.* (2009) egileena da. Sistemak SVMn oinarritutako sailkatzailea trebatzeko, egileek ikasketa-fasean NIL erantzunen laginak gehitzen dituzte. Estrategia horrekin, NIL kasuen % 86 inguruko asmatze-tasak lortzen dituzte esperimentu desberdinetan, NIL ez direnetan baino 15 puntu inguruko portaera hobea lortzen dute.

Bestetik, TAC-KBP 2009 edizioan NIL erantzuna kontsideratu gabe trebatutako sistema gainbegiratu onenetarikoa, bigarren postua lortu zuen Li *et al.* (2009) taldearen sistema da. Gogoratzeko, NE anbiguo bat ezagutza-baseko sarrera batekin lotzeko, problema bi urratsetan ebazten du sistema horrek: lehenbizi NEaren desanbiguazio hautagaiak pisatzen ditu Naïve Bayes Hedatuan oinarrituz trebatutako sailkatzailearekin; eta 2. urratsean, pisu altuena lortu duen hautagaia adiera egokia den edo ez ebazteko sailkatzaile

bitar bat aplikatzen du, ikasketa fasean ere trebatutakoa baina lehenengo urratseko sailkatzailetik at. Bigarren urrats horretan sailkatzailearen erantzuna ezezkoa denean, sistemak NIL itzultzen du. Estrategia horrekin, sistemak NIL erantzunen aurrean (% 82,84) ezagutza-basean adiera duten NE anbiguoen aurrean (% 77,25) baino portaera hobea lortzen du.

Sistema ez-gainbegiratuen artean, NIL prozesu osoan erantzun posible gisa tratatzeko hurbilpen bat Varma *et al.* (2009) egileen lanean aurkezten da. Proposatzen duten sisteman NIL kasuak identifikatzeko sistemak NEaren agerpen baten desanbiguazio-hautagaiak bilatzean soilik ataza partekatuan luzatutako ezagutza-basean[5] bilatu beharrean, Wikipedia osoarekin zabaldutako ezagutza-base batean burutzen da bilaketa. Horrela desanbiguazioa, ezagutza-basean existitzen diren zein ezagutza-basean sarrera ez duten hautagaiekin burutzen da. Prozesua aplikatu ostean, hautatutako desanbiguazio-adierak ataza partekatuko ezagutza-basean sarrerarik ez badu, sistemak NIL itzultzen du. Estrategia horrekin aurkezten dituzten esperimentuek % 85eko asmatze-tasa inguruko portaera dute NIL kasuen aurrean, NIL ez direnetan baino 9 puntu gainetik.

Antzera gertatzen da TAC-KBP 2010 edizioko hirugarren sistema onenaren kasuan (Radford *et al.*, 2010), portaera hobea lortzen duela NIL kasuetan, adiera ezagutza-basean duten NEen agerpenekin baino. Horretarako, NEen agerpenen hautagaiak sortzean, ezagutza-baseko sarrerak diren zein ez diren hautagaiak sortzen dituzte. Hala, sistemak NIL itzuliko du, hautagaiei pisuak esleitu ondoren batek berak ere ez badu zero baino antzekotasun-pisu altuagorik eskuratzen, edota lehen hautagaia Wikipediako sarrera bat ez denean.

Oro har, TAC-KBP 2009an ataza partekatuaren laburpenean (McNamee eta Dang, 2009) ikus daitekeen bezala, aurkeztutako sistema guztien artean onenek behinik behin, NIL kasuak NIL ez direnak baino hobeto ebazten dituzte. Beti ere, azpimarratu nahi da tratamendu berezia eskaintzen zaiola erantzun mota horri, bai, erantzun posibleen artean gehituz, zein azken urrats batean landuz.

[5] Wikipediako sarrera azpi-multzo batekin osatutako baliabidea.

IV.3 Ingurune esperimentala

Aurreko atalean ikusi ahal izan den bezala NEen desanbiguazio sistemak eraikitzeko bide eta aukera desberdinak daude. Ondoren, tesi-lan honetan euskarazko NED tresna garatzeko aplikatu den estrategia aurkezten da. Estrategia horrek oinarrian, euskarazko testu batean agertzen den NEa Wikipediako sarrera batekin lotzea du helburu. Lotura hori sortzeko, Wikipedian definituta dauden eta bat etor daitezkeen sarreren artean egokiena aukeratu beharko du.

Prozesu hori burutzeko erabilitako baliabideak eta teknikak azaltzen dira jarraian. Esan behar da atal honetan azaltzen diren teknikak sinpleak direla, NED ataza ebazteko metodo konplexuagoak badaude ere, euskararako eskuragarri ditugun baliabide eskasak direla eta, metodo sinple eta ezgainbegiratuekin lan egitea erabaki baitugu.

IV.3.1 Baliabideak

NEen itzulpen atazan bezala (ikus III. kapitulua), adiera-desanbiguazioko ataza honetan ere abiapuntua euskarazko testuak eta bertan agertzen diren NEak dira. Euskal testuetako NEak identifikatzeko, II. kapituluan azaldutako *Eihera+* erabiliko da, euskaraz idatzitako sarrerako testuan NEak mugatu eta sailkatzen dituen tresna.

Jarraian, NED ataza ebazteko NEen agerpenak Wikipediako sarrerekin lotzeko erabili ditugun baliabideen deskribapena egiten da. Lehenik eta behin, euskarazko Wikipedia bera nola ustiatu dugun azaltzen da. Ondoren, NEen agerpenen hautagaiak definitzen dituen karaktere-kate mailako hiztegiaren deskribapena egiten da. Eta azkenik, Hermeseko *Euskaldunon Egunkaria*ko 2002 urteko berrien testuak ebaluazio-corpusak sortzeko nola erabili diren deskribatzen da.

IV.3.1.1 Euskarazko Wikipedia

Wikipedia HPko aplikazio ugaritan erabili izan da semantikoki aberatsa den baliabide gisa, galdera-erantzun sistemen (Ahn *et al.*, 2005) zein testu sailkapenerako (Gabrilovich eta Markovitch, 2006) sistemen hobekuntzetarako, besteak beste. NEen alorrean Wikipedia baliatzen duten Toral eta Muñoz (2006) eta Kazama eta Torisawa (2007) egileenak bezalako lan ugari publikatu dira, NEen identifikazio eta sailkapen atazan *gazetteer* zerrenda

espezializatuak sortzeko, zein kanpo-ezagutza bezala erabili dutenak, emaitzen hobekuntzak lortuz. Lan honetan, Wikipedia, eta konkretuki euskarazko Wikipedia euskarazko NED ataza ebazteko nola erabil daitekeen aztertu da.

Sarrerako testuetan mugatutako NEak, Wikipediako sarrerekin lotuz adiera-desanbiguazioa burutzea da euskarazko NED tresnaren helburu nagusia. Euskarazko Wikipedia, beraz, estrategia honen funtsezko baliabidea kontsidera daiteke.

Wikipedia momentu oro hazten doan baliabidea izanik, beharrezkoa da momentu konkretu bateko bertsioa hartu eta horrekin lanean aritzea sistema eta, batez ere, ebaluaziorako datu multzo egonkor batekin lan egin ahal izateko. Tesi-lan honetan azaltzen diren datu eta ebaluazio guztiak, euskarazko Wikipediaren 2006ko martxoko bertsioarekin egindakoak dira. Datu horiek `http://download.wikimedia.org/euwiki/` webgunetik XML formatuan deskargatu dira[6].

Wikipediako sarrera nagusiek tituluan adierazten den NEaren edo kontzeptuaren deskribapena izan ohi dute. Deskribapen horiek, edozeinek edita ditzakeen testu libreak izan arren, Wikipediaren hainbat arau[7] bete behar dituzte. Besteak beste, Wikipedian definituta dagoen kontzeptu bat aipatuz gero, lehenbiziko aipamena gutxienez, Wikipediako sarrera horrekin lotzea derrigorrezkoa da. Esteka horietan, beraz, testuan kontzeptuak duen forma eta dagokion Wikipedia sarrera nagusiaren titulua agertuko dira, biak berdinak direnean behin bakarrik azalduz informazioa. *Euskal Herria* sarrera kontsultatuz gero, euskarazko Wikipediaren XML bertsioan IV.3.1 adibideko testua aurkitzen dugu.

IV.3.1 Adibidea Wikipediako *Euskal Herria* sarrerako testua.

"Euskal Herria" [[Europa]]ko herrialdea da. Historikoki [[euskaldun]]en eta [[euskara]]ren lurraldea da, [[Pirinioak|Pirinio]] mendien mendebaldean kokatua, [[Frantzia]] eta [[Espainia]]ren arteko muga egiten duen mendilerroan, eta [[Bizkaiko Golkoa|Bizkaiko golkorantz]] zabaltzen dena.

Kakoen artean agertzen diren kontzeptuak Wikipedian sarrera duten kontzeptuak dira, hain zuzen ere. *[[Pirinioak|Pirinio]]* testu kateari erreparatuz, ikus daiteke testuan aipatzen den *Pirinio* kontzeptua, *Pirinioak* sarrerarekin lotuta dagoela.

[6] `http://ixa2.si.ehu.es/izaskunfernandez/euwiki-latest-pages-articles.xml.bz2` helbidean eskuragarri.

[7] `http://eu.wikipedia.org/wiki/Laguntza:Artikuluak_nola_aldatzen_diren`

Berbideratze motako orrietan, aldiz, tituluan agertzen den kontzeptuaren formarik usuena deskribatzen duen sarrera nagusiarekin lotzen duen esteka soilik egon ohi da. Estekak berbideratze motako esteka dela adierazita du. Horren adibide, Wikipediako *Pirinio* sarrera da. Orri honen deskribapen atalean, *Pirinioak* orrira zuzentzen duen berbideratze motako esteka besterik ez dago: *#REDIRECT [[Pirinioak]]*, modu horretan informazio errepikapena ekidinez.

Desanbiguazio-orriek, azkenik, tituluan agertzen den kontzeptuak erreferentzia ditzakeen Wikipediako sarrera nagusietako zerrenda bat izan ohi dute deskribapen gisa, non erreferentzia dezakeen sarrera nagusia aipatzen den horren definizio labur batekin batera. Esaterako, *Euskadi* sarrera kontsulatuz IV.3.2 adibideko informazioa eskuratzen da.

IV.3.2 Adibidea Wikipediako *Euskadi* sarrerako testua.

Euskadi honako bi esanahiekin erabili izan da:

- *[[Euskal Autonomia Erkidegoa]], Euskal Herriko mendebaldeko zatia, [[Espainia]]ko erkidego bat osatzen duena (erabilera arruntena).*
- *[[Euskal Herria]], historikoki [[euskaldun]]en eta [[euskara]]ren lurraldea.*

Hortaz, erraza da ikustea Wikipediaren arabera *Euskadi* kontzeptuak *Euskal Autonomia Erkidegoa* zein *Euskal Herria* sarrerei erreferentzia egin diezaiekeela. Wikipediaren arabera, hortaz, *Euskadi* espresioa, *Euskal Autonomia Erkidegoa* eta *Euskal Herria* adierak izan ditzakeen NE anbiguoa da.

Wikipedia eta bere egitura ustiatuz, beraz, NEen desanbiguaziorako sistema baten garapenerako oso baliagarri gerta daitekeen informazioa eskura daiteke.

IV.3.1.2 Karaktere-kate mailako hiztegia

Hitzen adiera-desanbiguazioan hitz anbiguo batek dituen adiera posibleak ezagutzeko hiztegi edo baliabide bat atzitzen den bezala, NEen desanbiguazio atazan antzeko urratsa behar da edozein NE anbiguo emanik, bere adiera posibleak identifikatzeko.

Kapitulu honetan aurkezten den NED tresnak NEaren espresio anbiguo bati dagokion euskarazko Wikipediako sarrera posibleen zerrenda eskuratzea izango du lehen helburu.

Zerrenda horren eskurapena burutzeko, euskarazko Wikipedian eta bere egituran oinarritutako karaktere-kate mailako hiztegi bat definitu da, Agirre

et al. (2009) egileen lanean proposatutako bidetik. Funtsean, sarrera nagusien tituluak, berbideratze- eta desanbiguazio-orrietako esteken informazioa eta, oro har, euskarazko Wikipediako sarrera bakoitzari erreferentziatzen dioten forma anbiguoak identifikatzeko deskribapen testuetan zehar agertzen diren Wikipedia sarreren esteketako informazioa erabili da. Edo beste modu batera esanda, NEaren forma anbiguo batek Wikipediako zein sarrera nagusiri egin diezaiokeen erreferentzia adierazten duen hiztegia definitu da. Karaktere-kate mailako hiztegia sortzeko prozesuan informazio mota bakoitzaren erabilera zehazten da ondoren:

- Wikipediako sarrera nagusi bakoitzeko titulua eta deskribapena erabiliz:
  - Tituluko forma, sarrera beraren forma anbiguoa izango balitz bezala biltegiratzen da. Horrez gain, titulua hitz anitzeko terminoa bada 2 urrats gehiago ematen dira:
    1. Hitz bakoitzeko 1. letra erabiliz azken hitzekoa ezik eta posizioak mantenduz, sor daitezkeen konbinazio guztiak biltegiratzen dira. Adibidez: *Juan Jose Ibarretxe* sarrerarako, *J. Jose Ibarretxe*; *J. J. Ibarretxe* eta *Juan J. Ibarretxe* formak sarrera honen forma anbiguo gisa biltegiratuko dira.
    2. Hitz osoak erabiliz eta posizioak mantenduz hitzen konbinazio guztiak biltegiratu. *Juan Jose Ibarretxe* sarreraren tituluarekin jarraituz, *Juan*, *Jose*, *Ibarretxe*, *Juan Jose*, eta *Jose Ibarretxe* forma anbiguo bezala biltegiratuko dira.
  - Deskribapenetako testuak erabiliz, beste sarreren testuetan erreferentzia egiten bazaio lantzen ari den sarrerari edo bertara berbideratutariko bati, forma hori ere aldaera gisa biltegiratzen da.
- Berbideratze motako orri bakoitzeko, tituluaren katea deskribapen atalean adierazten den sarreraren aldaera gisa biltegiratzen da .
- Desanbiguazio-orri bakoitzeko, tituluko forma deskribapenean agertzen den zerrendako sarrera bakoitzaren forma anbiguo bezala biltegiratzen da.

Prozesamendu horretatik, beraz, euskarazko karaktere-kate mailako hiztegi bat eskuratzen da, non NE anbiguo batetik desanbiguazio prozesuan

adiera gisa erabili behar diren euskarazko Wikipedia sarrerak gordetzen diren. Esaterako, *Ibarretxe* espresio anbiguoarentzat *Juan José Ibarretxe* eta *Esteban Ibarretxe* gorde dira.

Hiztegian guztira, gutxienez Wikipedia sarrera bat lotuta duten 230.000 forma anbiguo inguru biltegiratu dira.

Prozesamendurako, Milneren (2009) lanean deskribatzen den Wikipedia-Miner[8] tresna erabili da. Tresna horrek edozein Wikipedia XML bertsioko berbideratze, desanbiguazio eta oro har Wikipediako egitura zatitu eta fitxategi sinpleagoetan deskonposatzen du. Deskonposaketa horretan oinarrituz, hain zuzen ere, eraiki da hiztegia.

IV.3.1.3 Ebaluazio-corpusak

Euskarazko Wikipedian oinarritutako NEen desanbiguaziorako sistemaren ebaluazioa burutu ahal izateko, III. kapituluan deskribatutako Hermes proiektuan[9] etiketatutako 2002 urteko *Euskaldunon Egunkaria*ko corpusa erabili da. 40.648 artikuluk osatzen duten corpus horretan, 130.505 NE etiketatuta daude.

Kontuan hartzekoa da kazetariek berriak idazteko garaian pertsona, toki edo erakunde bat, alegia NE bat aipatzean, lehenbiziko aipamenean forma luzea edo osoa idatzi ohi dutela eta ondorengo aipamenetan, berriz, forma laburragoak. Esate baterako IV.3.3 adibidean agertzen den pilotari buruzko berri zatian, *Aimar Olaizola* pilotariari erreferentzia egiten dioten 3 aipamen desberdin topa daitezke.

IV.3.3 Adibidea *Aimar Olaizola* NEari aipamena egiten dion berri zatia.

Ajeak dira biak, baina ezberdinak oso arrakastarena eta porrotarena. ***Aimar Olaizolak*** *eta Abel Barriolak bestondo erabat ezberdina izan zuten atzo...* ***Olaizola*** *kontrario eta lagunaren dohainak goratu zituen: "Kirolean burua da garrantzitsuena, eta* ***Aimarrek*** *bere tokian dauka".*

Lehenbiziko aipamenean, izen osoa agertzen bada ere, ondorengo bi aipamenetan *Aimar Olaizola* erreferentziatzeko forma laburrak erabiltzen ditu kazetariak, *Olaizola* lehenbizi eta *Aimar* beranduago. Forma luzea bera anbiguoa izan badaiteke ere, laburrek anbiguotasun handiagoa sortu ohi dute. Adibidean esaterako, *Olaizola* formak *Aimar* zein *Asier* Olaizola bi anaietako edozeini erreferentzia egin diezaioke, lehenbiziko aipamena ezagutzen ez

[8] http://wikipedia-miner.sourceforge.net
[9] http://nlp.uned.es/hermes/

bada. Hortaz, berri sorta hau desanbiguazio ataza neurtzeko baliabide egokia izan daiteke, forma luzeko aipamenari ez ikusiarena eginez, forma laburren desanbiguazioa egiten saiatzeko.

Idazkera ohitura hori baliatuz eta forma labur horien forma luzea/desanbiguazio forma lehenengo aipamenekoa dela suposatuz, garapen eta ebaluazio-corpusak eratu ditugu. Garapeneko corpusa sistema balioztatzeko eta hobetzeko erabiliko da eta, behin sistema egonkorra denean, bere portaera ebaluazio-corpusarekin neurtuko da.

Berriak ebaluazio unitate gisa erabili ordez, paragrafo mailako zatiketa eta aukeraketa egin da, NEen forma laburrak eta luzeak aintzat harturik. Zatiketa irizpidea konkretuki ondoko hau izan da: batetik berria bere osotasunean harturik, NEen forma laburrak agertzen diren paragrafoak erauzi dira, aurretik forma labur horien NEaren forma luzea agertzen bada; eta bestetik, forma luzeagorik aurretik ez duten paragrafoak. Horrela, bi motatako paragrafo sortak bereizi dira.

Paragrafo bakoitza NE bat desanbiguatzeko erabiliko da. Lehenengo multzoko paragrafoetarako, beraz, aurretik bera baino forma luzeagoa duen NEaren agerpena, NE anbiguoaren desanbiguazio forma egokia dela suposatuko da, eta bigarren multzokoetarako desanbiguazio forma ezezaguna izango da.

Kapitulu honetan azaltzen den tresna, Wikipedian oinarritutako euskarazko NEen desanbiguazioa burutzen duen sistema izanik, alegia, testu batean agertzen den NEari zein Wikipedia sarrera dagokion ebazten duen sistema izanik, ondoko bi aukerak eman daitezke: desanbiguatu nahi den NEaren forma luzea edo NE desanbiguatuak *Wikipedian sarrera du*; edo NE anbiguoari dagokion desanbiguazio forma *ez da Wikipedian existitzen*.

Ebaluazioari begira informazio hori kontuan hartuz gero, NEen forma luze eta laburren irizpidean oinarrituz eskuratutako bi paragrafo motei erreparatuz gero, ondoko paragrafo multzoak lortzen dira:

- NE laburrak, beraien forma desanbiguatua kontsidera daitekeen NE luzeagoa berrian agertzen da, eta azken hori bat dator Wikipediako sarrera batekin.

- Aurrekoan bezala NE anbiguorako NE desanbiguatua kontsidera daitekeen NE luzeagoa badago berrian, baina azken horrek ez du Wikipedian sarrerarik.

- NE anbiguoak ez du bera baino luzeagoa den NE agerpenik aurretik.

Lehenbiziko multzoko paragrafoek, inolako eskuzko lanik egin gabe Wikipedian oinarritutako desanbiguazio atazan lan egiteko baliabidea osatzen dute, ezaguna baita sistemak NE anbiguorako desanbiguazio forma gisa itzuli behar duen Wikipedia sarrera.

Bigarren eta hirugarren multzoko paragrafoetako NEen kasuan, aldiz, ezin da gauza bera esan, alegia eskuzko lana beharrezkoa da desanbiguazio sistemaren egiaztapen eta ebaluaziorako erabili nahi izanez gero. Eskuzko lan horretan, egiaztatu beharko da desanbiguatu nahi den NEaren kontzeptua deskribatzen duen Wikipedia sarrerarik dagoen.

Bigarren multzokoen kasuan, berriko forma luzeak Wikipedian sarrerarik ez izateak, ez du esan nahi Wikipedian espresio motzaren desanbiguazio formarentzako sarrerarik ez dagoenik. Hori dela eta, eskuz aztertu behar da sarrera egokirik dagoen Wikipedian, eta baiezkoan definitu zein den. Sarrera egokirik ez dagoenean aldiz, ezingo da sarreraren esleipenik egin. Azken adibide horiek, etorkizunean sistemak NIL erantzunak itzultzen ikasteko adibide aproposak izan daitezke.

Hirugarren multzokoen kasuan, aztertu beharko da zein den desanbiguazio forma, berrian bera barne duen NE luzeagorik ez dagoenez, ezarritako irizpideen arabera ezin baita automatikoki desanbiguazio-forma identifikatu. Esaterako, IV.3.4 ondoko testuan bi eskultoreri egiten zaie erreferentzia eta aurretik ez dago bi horien izen osoak adierazten dituzten espresioak.

IV.3.4 Adibidea Testuan zehar aipamen luzeagorik ez duten bi eskultoreri buruzko berri zatia.

...Konparazio batera, gaurko lehen atalean honako hauek ekarriko dituzte gogora: Oiz mendian izan zen hegazkin istripua, Yoyesen hilketa, euskal txirrindularien garaipenak eta ***Oteiza*** *eta* ***Txillida*** *eskultoreen arteko besarkada...*

Oteiza eta *Txillida* NEak barne duten forma luzeagorik aurretik egon ez arren, testuinguruagatik ondoriozta daiteke *Jorge Oteiza* eta *Eduardo Txillida*ri egiten dietela erreferentzia, hurrenez hurren, eta eskultore horiek deskribatzen dituzten sarrerak badaude Wikipedian. Beraz, eskuzko lanean esleipen hori eskuz egingo da eskuz landutako corpusa osatuz.

Laburbilduz, batetik modu automatikoan eratutako paragrafo sorta bat lortzen da, non paragrafoko forma anbiguori dagokion forma ez-anbiguoa Wikipediako sarrera ezaguna den (corpusA). Eta bestetik, eskuzko berrikuspenaren ondoren, beste paragrafo sorta bat definitu dugu, non NE anbiguoaren desanbiguazio adiera Wikipedian existitzen zein existitzen ez diren paragra-

	Adib. Kop.	Adib. Anbiguo Kop.	NIL	Anb. Tasa
CorpusA	**6500**	**4376**	**0**	**% 67,32**
CorpusB-garapen	**532**	**295**	**70**	**% 55,45**
CorpusB-test	**500**	**300**	**63**	**% 60,00**

IV.1 Taula: Ebaluazio-corpusak

foak nahasten diren (corpusB). B corpus hori lagin txikia da, eskuzko lana minimizatzeko asmoz eratorria.

Bi corpusen izaera desberdinagatik, A corpusarekin sistemaren doitasuna neurtu ahalko dugula esan daiteke eta bigarrenarekin berriz, B corpusarekin, sistemaren estaldura ebaluatuko dugula. B corpusa bitan banatu da, garapen (corpusB-garapen) eta test (corpusB-test) multzoetan, hain zuzen ere. Tesi-lan honetan bi zatiak berdin erabili badira ere, etorkizunean sistema hobetzeko eta fintzeko banaketa interesgarria dela kontsideratu dugu.

IV.1 taulan laburbiltzen dira *Euskaldunon Egunkaria*n oinarrituta sortutako bi ebaluazio-corpusen banaketari buruzko informazioa. Konkretuki, corpus bakoitzeko zehazten dira adibide kopurua; adibide horietatik zenbatetan adibideko NEari Wikipedia artikulu posible bat baino gehiago egokitu dakiokeen, NE anbiguoa duten adibide kopurua (Adib Anbiguo Kop. zutabea), alegia; zenbat adibidetan NEaren desanbiguazio adiera egokiak euskarazko Wikipedian sarrerarik ez duen (NIL zutabea); eta, azken zutabean, corpusaren batez besteko anbiguotasun maila zehazten da.

IV.3.2 Metodoak

Adiera-desanbiguazioko atazan asko dira erabili izan diren metodoak, IV.2 atalean aurkeztu den bezala. Ondoren, euskarazko NEen desanbiguazioa burutzeko erabilitako metodo ez-gainbegiratuak deskribatzen dira.

Gogoratu lan honen helburua ez dela metodo konplexuak proposatzea, beste hizkuntzetarako sinpleak eta egokiak izan direnak euskarazko NED atazan probatzea, eta posible denean, hobekuntza txikiren bat proposatzea baizik. Hala izanik, ondorengo ataletan euskarazko NED ataza ebazteko erabilitako lau metodoen deskribapen orokorra egiten da: maiztasun handieneko adiera, bektore-espazioen eredua, ESA eta UKB algoritmoen deskribapena, hain zuzen ere.

IV.3.2.1 Maiztasun handieneko adiera (MFS)

Maiztasun handieneko adiera (*Most Frequent Sense*, MFS), izenak adierazten duen bezala, adiera guztien artean maitasun handienekoa hautatzen duen metodo sinplea da. Teknika hau aplikatu ahal izateko, NEak etiketatuta dituen corpusa behar da, non NEen agerpen bakoitzean adiera egokia zein den adierazita egon behar den. Horrela, adieren agerpenen kontaketa egin eta NE bakoitzerako maiztasun handieneko adiera kalkulatu ahal izango da.

Heuristiko honek testu berri bateko NE anbiguo baten agerpena, etiketatutako corpusaren kontaketetan oinarrituta ondorioztatutako maiztasun handieneko adierarekin desanbiguatuko du.

Oso sinplea den arren, metodo honen emaitzak hobetzea zaila izan da hainbat atazatan eta bera baino askoz konplexuagoak diren algoritmoek oso hobekuntza txikiarekin gainditzea lortu izan dute (Lopez de Lacalle, 2009).

IV.3.2.2 Bektore-espazioen eredua (VSM)

Bektore-espazioen eredua (*Vector Space Model*, VSM), eredu aljebraiko bat da. Ereduak, testuen edukiak bektoreetan adierazten ditu testuko hitzak, gako-hitzak (*keyword*) edo n-gramak eta horien maiztasunak bektore horietan egituratuz. Eredu honetan, testuen arteko antzekotasuna bektoreen arteko angeluaren kosinuaren (Baeza-Yates eta Ribeiro-Neto, 1999) bitartez neurtu ohi da.

Testu bakoitza, beraz, n dimentsioko bektore baten bidez errepresentatzen da, dimentsio bakoitza testuko hitz, gako-hitz (*keyword*) edo n-grama bati egokituz. Lan honetan dimentsio bakoitzean hitz edo lema baten informazioa gordeko da. Testuak errepresentatzeko modu honek, *bag-of-words* izena ere hartzen du.

Bektoreen arteko konparaketak egin ahal izateko, testu guztiak espazio berdinean errepresentatu beharko dira. Alegia, dimentsio-kopuru eta esanahi berdineko bektoreetan adierazita egon beharko dira testuak. Dimentsio-kopuru hori corpuseko hiztegiak mugatuko du, alegia, corpusa osatzen duten dokumentuetako termino desberdinen kopuruak. Hiztegi horren osaketan, dimentsio-kopurua murrizteko bidean, esanguratsuak ez diren hitzak ezabatu ohi dira. Hitz horiek *stopword* izenarekin ere ezagutzen dira eta zerrenda batean bilduta egon ohi dira. Dimentsio-kopurua beraz, *stopword* zerrendatan agertzen ez diren corpuseko dokumentuen termino desberdinen kopuruak mugatuko dute.

Termino bat dokumentu batean agertzen bada, dokumentuaren bektorean termino horri dagokion posizioko balioa zeroren desberdina izango da. Terminoen balioak edo pisuak kalkulatzeko metrika ugari daude arren, zabalduena *tf-idf* da, eta horixe erabili da tesi-lan honetan.

Definizioz *tf-idf* pisua (Salton eta McGill, 1983), hitz batek corpus bateko dokumentu batean duen garrantzia neurtzen duen neurri estatistiko bat da. Termino baten pisua altua izango da dokumentu jakin batean agerpen ugari baditu eta, aldi berean, corpuseko dokumentu askotan ez bada termino hori agertzen. Alegia, zenbat eta dokumentu gehiagotan agertu, orduan eta pisu txikiagoa du. Haztaketa metodo horrek bi osagai nagusi ditu: osagai lokala, terminoak dokumentu jakin batean duen garrantzia neurtzen duen osagaia; eta bestetik osagai globala, dokumentu bilduman terminoaren garrantzia neurtzen duena. Terminoaren maiztasuna (tf) eta alderantzizko dokumentuaren maiztasuna (idf) dira, hurrenez hurren.

tfk, terminoak dokumentu batean duen agerpen kopurua definitzen du.

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|}$$

$n_{i,j}$ balioa, i dimentsioko terminoak j dokumentuan dituen agerpen kopurua izanik, eta zatitzailea, dokumentuko termino guztien agerpenen batura, dokumentuaren tamaina ($|d_j|$), alegia.

idfk bestetik, terminoak bilduman duen garrantzia definitzen du. Formalki,

$$idf_i = log(\frac{|D|}{1+|d:t_i \epsilon d|})$$

non $|D|$ corpuseko dokumentu kopurua adierazten duen, eta $|d : t_i \epsilon d|$, t_i terminoa agertzen den corpuseko dokumentu kopurua. Definizio horietan oinarrituta *tf-idf* formalki ondoko formularekin kalkulatzen da:

$$tf\text{-}idf_{i,j} = tf_{i,j} * idf_i$$

n termino desberdin dituen corpus bateko m termino desberdineko testu bat espazio bektorialean errepresentatzeko beraz, n dimentsioko bektore bat definitzen da, non n-m dimentsiotan zero balioa izango duen, eta testuan agertzen diren terminoen m dimentsioek aldiz, *tf-idf* pisua izango duten.

Corpus osoko dokumentuak n dimentsioko espazioan definituta, dokumentuen antzekotasuna bektoreen arteko eragiketen bidez egitea posible izango da. Bi bektoreen antzekotasuna neurtzeko teknika erabiliena, bektoreen arteko angeluaren kosinua da. n dimentsioko A eta B bektoreak emanik, kosinu antzekotasuna ondoko formularen bitartez kalkulatzen da:

$$antzekotasuna = cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} * \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Kosinuaren balioa zenbat eta 1etik hurbilago egon, dokumentuen arteko antzekotasun orduan eta handiagoa adierazten du. Eta kosinuak zero balioa duenean berriz, bi dokumentuen artean termino komunik ez dagoela adierazten du.

IV.3.2.3 Analisi semantiko esplizitua (ESA)

Analisi semantiko esplizitua, ingelesez *Explicit Semantic Analysis* (ESA) izenarekin ezagutzen den teknikak antzekotasun semantikoa lantzen duten atazetan azken urteetan indar handia hartu du. Oinarrian teknika honek, edozein testu Wikipediako kontzeptuen dimentsioen arabera adieraztea du helburu, dimentsio horretan dokumentuen arteko edo galdera bat eta dokumentu sorta baten antzekotasun semantikoa kalkulatu ahal izateko.

Gabrilovich eta Markovitchek (2007) teknika honen asmatzaileek azaltzen duten bezala, ESAren oinarrizko ideia edozein hitz Wikipediako sarreren bektore bezala errepresentatzea da, non bektoreko posizio edo dimentsio bakoitza Wikipediako kontzeptu bat izango den eta posizio horren pisua, hitzak Wikipediako sarreran duen garrantziak definituko du. Hitzak Wikipediako espazioan errepresentatzeko gai izanda, testuak errepresentatzeko aukera ere badago, testuko hitz guztien bektoreen konbinazioa egitea besterik ez baita behar. Algoritmoaren funtsa, beraz, hitz batek Wikipediako sarrera bakoitzean duen garrantzia adierazten duen pisua kalkulatzean datza, eta testu osoen errepresentaziorako, bektoreen konbinazioa burutzean. Haztaketa-sistemari *alderantzizko indize pisuduna* edo *weighted inverted index* izena ematen diote.

Alderantzizko indize horiek kalkulatzeko Wikipedia aurreprozesatu egin behar da, Wikipediako sarrera bakoitza, bektore-espazioen ereduan bezala, *bag-of-words* edo pisu-bektoreen formatuan adieraziz. Hitzen pisuak kalkulatzeko IV.3.2.2 atalean azaldutako *tf-idf* metrika bera erabiltzen da. Behin sarrera bakoitzaren bektore-espazioko pisu-bektoreak definituta, alderantzizko indizeak kalkulatzen dira, bektoreetako hitz bakoitzerako bere Wikipediako sarreretako agerpenak kontsideratuz. Alegia, hitz bakoitzeko Wikipediako kontzeptuen bektorea sortzen da, posizio bakoitzean, hitz horrek Wikipediako sarrera bakoitzean duen *tf-idf* balioa adieraziz. Formalki beraz, $p_{i,j}$ i hitzak j Wikipedia kontzeptu edo sarreran duen *tf-idf*a izanik, hitzaren bektorea ondoko litzateke:

$$hitza_i = (p_{i,1}, p_{i,2}, ..., p_{i,|W|})$$

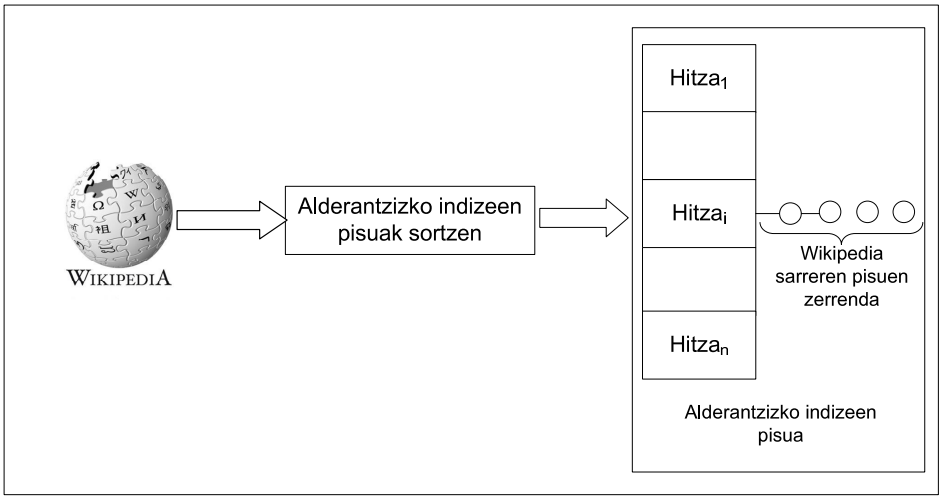
non $|W|$ Wikipediako artikulu kopurua den.

Hortaz, edozein testu Wikipediako kontzeptuen espazioan proiektatzeko, testu horretako hitzek Wikipediako kontzeptuetan duten pisuen batura besterik ez da egin behar. Alegia,

$$testua = (\sum_{k=1}^{|t|} p_{k,1}, \sum_{k=1}^{|t|} p_{k,2}, ..., \sum_{k=1}^{|t|} p_{k,|W|})$$

non $|t|$ testuko hitz kopurua den eta $|W|$ Wikipediako artikulu kopurua adierazten duen eta $p_{k,i}$ testuko $k.$ hitzak $i.$ Wikipedia artikuluan duen *tf-idf* balioa den.

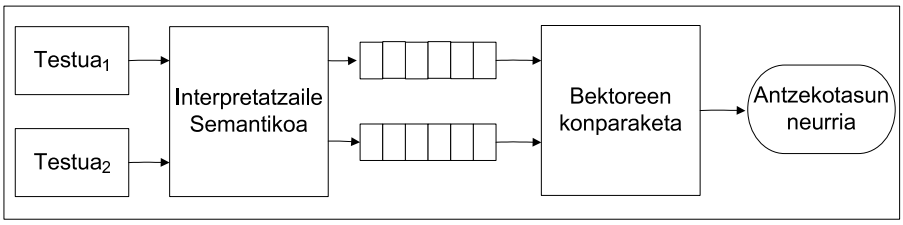
Edozein testu, Wikipediako espazioan proiektatzeko prozesu honi, Gabrilovich eta Markovitchek (2007) interpretatzaile semantikoa deitzen diote eta IV.2 irudian prozesua grafikoki ikus daiteke.



IV.1 Irudia: ESA algoritmoaren interpretatzaile semantikoa

IV.2 irudian agertzen den bezala ESAren bitartez bi hitz, bi dokumentu zein IR sistema bateko galdera eta dokumentu baten arteko antzekotasuna kalkulatzeko interpretatzaile semantikoarekin pisu-bektoreak sortu behar dira eta ondoren, antzekotasun-neurria lortzeko, bektoreen konparaketa burutu. Bektore-espazioen ereduan bezala, Gabrilovich eta Markovitchek (2007) bi bektoreen konparaketarako bektoreen arteko angeluaren kosinua erabiltzen dute. Sarrerako testuek Wikipediako sarrerekin duten antzekotasunean

oinarritutako antzekotasun-neurria lortzen denez, zenbat eta handiagoa izan balioa (0 eta 1 artekoa), orduan eta antzekoagoak izango dira bi testuak. IV.2 irudiak, ESAk bi testuren arteko antzekotasuna neurtzeko ematen dituen urratsak laburbiltzen dira.



IV.2 Irudia: ESA antzekotasun algoritmoa

ESA algoritmoa NEen desanbiguazio atazan erabiltzeko (Sorg eta Cimiano, 2008a) NE anbiguo bat desanbiguatzeko sarrerako testuaren bektoreak Wikipedia osoaren espazioan irudikatzea ez da beharrezkoa. Nahikoa da desanbiguazioaren emaitza izan daitezkeen Wikipediako sarreren espazioan proiektatzea sarrera, helburua ez baita bi testuren antzekotasuna neurtzea, sarrerako testu bat Wikipediako hainbat sarrerekin konparatzea baizik.

Azken batean problema beste ikuspuntu batetik planteatzea besterik ez da. Alegia, NE anbiguoa duen sarrerako testua, IR sistema bateko galdera bailitz kontsideratu eta bilaketa-corpusa Wikipediako artikuluak bailira. Horrela izanda, ESA algoritmoarekin galderaren erantzuna Wikipedian bilatzeko, Wikipediako artikulu antzekoena bilatzeko, alegia, sarrerako testuaren alderantzizko indizearen pisu-bektorea kalkulatu beharko da, non posizio bakoitzean testuak Wikipediako sarrera bakoitzarekin duen antzekotasun-pisua adieraziko den. Ondorioz, balio altuena lortzen duen posizioak Wikipediako artikulu antzekoena zein den adieraziko du.

NEen desanbiguazioan ordea, Wikipediako artikulu sorta osotik, soilik azpimultzo baten artikulu antzekoena zein den eskuratzea da funtsa. Zehazki, desanbiguazioaren emaitza izan daitezkeen artikuluen artean, sarrerako testu jakin baterako probableena zein den jakitea. Hortaz, ESAren alderantzizko indizearen pisu-bektorea kalkulatzeko garaian, soilik desanbiguazioaren emaitzak izan daitezkeen Wikipedia artikuluen antzekotasun-pisua kalkulatzea nahiko izango da, beste artikulu guztiekin kalkuluak baztertuz.

Problema sinplifikatuta, ESAren bektoreko posizio bakoitzeko pisua kalkulatzeko, alegia NE anbiguoaren testuingurua adierazten duen n hitzez osatutako $t = (h_1, h_2, ..., h_n)$ testua eta desanbiguazio-adiera izan daitekeen Wi-

kipediako a_j artikulu bakoitzaren arteko antzekotasuna neurtzeko, Sorg eta Cimianoren (2008a) arabera ESAn oinarritutako ondoko metrika aplikatzen da:

$$antzekotasuna(t, a_j) = (C_t) * \sqrt{|a_j|}^{-1} * \sum_{h_i \epsilon t} tf_{a_j}(h_i) * idf(h_i)$$

non, $C_t = \frac{1}{\sqrt{\sum_{h_i \epsilon t} idf(h_i)}}$ den eta $|W|$ Wikpipediaren artikulu kopurua.

$\sqrt{|a_j|}^{-1}$ Wikipedia artikuluko luzerarekiko normalizazio-faktorea izanik eta, C_t neurria sarrerako testuaren menpeko normalizazio-neurria izanik, Wikipediako artikuluek sarrerako testuarekin kalkulatzen den pisuan eta kosinuan eraginik ez dutela suposa daiteke.

Beraz, Gabrilovich eta Markovitchen (2007) antzekotasun formulara itzuliz, $t = (h_1, h_2, ..., h_n)$ sarrerako testuaren $V = (v_1, v_2, ..., v_n)$ *tf-idf* bektoretik abiatuta, Wikipediako a_j sarrerarekin antzekotasuna neurtzeko ondoko kalkulua burutuko da:

$$antzekotasuna(t, a_j) = \sum_{h_i \epsilon t} v_i * k_j$$

non k_j, h_i hitzak a_j artikuluan duen pisua da, *tf-idf* balioa, alegia.

IV.3.2.4 ESA orekatua

Aurreko atalean proposatutako soluzioan, antzekotasuna neurtzeko ez dira sarrerako testua eta Wikipediako sarreraren artean konpartitzen diren hitzak kontuan hartzen. Horrek, Wikipediako sarrera luzeek motzagoen aurrean antzekotasun-neurri baxuagoak izatea eragin dezake, artikulu luzeetako hitzen *tf-idf* neurriak laburragoetakoak baino balio baxuagoak izan ohi baitira. Hortaz, Gabrilovich eta Markovitch (2007) zein Sorg eta Cimianoren (2008a) lanetan aurkezten diren haztaketa-metodoak aplikatuz gero, gerta daiteke hitz komunen kopuru handigoa duen Wikipediako sarrera luze baten antzekotasun-neurria, hitz kopuru txikia konpartitzen dituen Wikipedia sarrera motz bateko antzekotasun-neurria baino baxuagoa izatea, eta hortaz sarrera luzeagoa duen desanbiguazio-hautagaia posizio okerragoan geratzea, aurrerago emaitzen atalean ikusi ahal izango den bezala.

Sarrera gehienek hitz kopuru minimo altua duten Wikipedia orekatu batean fenomeno hau ez litzake oso larria izango, sarrera guztietako hitzen *tf-idf* balioak antzekoak izango bailirateke. Tesi-lan honetan erabiltzen den Euskarazko Wikipedia (eta gure ustez beste hizkuntza batzuetako Wikipedia txiki

asko) ordea, ezin da corpus orekatu kontsideratu, oso luzera desberdinetako sarrerak baitaude, eta gainera asko oso motzak dira.

Artikulu laburrek kasu batzuetan informazio gutxi eta arazo handiak ematen dituzte. Horregatik, aurrez aipatu bezala (ikus IV.1 atala) hainbat lanetan sarrera laburrak ezabatu izan dira. Zoritxarrez, euskarazko baliabide honetan estrategia hori aplikatuz gero, esperimentuak burutzeko corpus txikiegia geratuko litzateke. Hortaz, desorekatua izan arren, luzera desberdinetako sarrerekin lan egitea erabaki da.

Dena dela, sarrera laburren eragina minimizatzeko asmoz, antzekotasun-neurrian normalizazio-faktore berri bat gehitu da, sarrerako testua eta Wikipediako sarreraren arteko hitz komun kopurua kontuan hartzen duen normalizazioa, hain zuzen ere.

Gabrilovich eta Markovitchen (2007) antzekotasun formulara itzuliz, eta normalizazio-faktore berria gehituz, $t = (h_1, h_2, ..., h_n)$ sarrerako testuaren $V = (v_1, v_2, ..., v_n)$ *tf-idf* bektoretik abiatuta, a_j Wikipedia sarrerarekin antzekotasuna neurtzeko ondoko kalkulua burutuko da:

$$antzekotasuna(t, a_j) = (\sum_{h_i \epsilon t} v_i * k_j) * \frac{Count_t}{|t|}$$

non k_j, h_i hitzak a_j artikuluan duen pisua, *tf-idf* balioa alegia, eta $Count_t$ t sarrerako testuaren eta a_j artikuluaren arteko hitz komunak den.

Ebaluazio-corpuseko testu guztiak, paragrafo mailakoak izanik, testu laburrak izan arren luzera antzekoa dute. Normalizaziorako faktore berriak luzera hori eta ez Wikipediako artikuluko luzera erabiltzean Wikipediaren menpekotasuna ekiditen du. Horrela, hitz komun kopurua txikia denean antzekotasuna txikiagoa izango da eta kopurua handiagoa denean antzekotasuna handiagoa, zarata sortzen duen Wikipediaren eragina txikitzea lortuz. Normalizazio-faktore berri honen eragina emaitzen atalean (ikus IV.5 atala) aztertuko dugu.

IV.3.2.5 UKB

UKB (Agirre eta Soroa, 2009) grafoetan oinarrituko algoritmoa da. Zehazki, algoritmo honek ezagutza-base bat grafo baten modura adierazten du, non ezagutza-baseko kontzeptuak grafoaren nodoak diren, eta kontzeptuen arteko erlazioak, berriz, grafoen arkuak. Grafo horretako nodoak pisatzeko, zorizko ibilbideetan (*random walks*) oinarritutako *PageRank* (Brin eta Page, 1998) algoritmoaren aldaera erabiltzen da.

PageRank algoritmoa webgune baten garrantzia ebaluatzeko pentsatutako sistema da. Garrantzia neurketa horretarako webguneen arteko estekak erabiltzen ditu. Hala, A orriak B orrirako duen esteka Ak Bri emandako boto gisa hartzen du. Eta zenbat eta A garrantzitsuagoa izan, orduan eta boto horren pisua altuagoa izango da.

Formalki, *PageRank* algoritmoak zorizko ibilbideen (*random walk*) eredua aplikatzen du. Ibilbidea zoriz hautatutako grafoko nodo batetik abiatzen da eta urrats bakoitzean, bertatik irteten den zorizko arku bat hautatzen duen ibilbideari jarraipena ematen dio. Edozein momentutan, arkuak jarraitu beharrean horiek alde batera utzi eta algoritmoa grafoko beste edozein nodotara joan daiteke ibilbidea osatzen jarraitzeko. Nodo baten pisua, beraz, ibilbide osaketan nodoa egoteko probabilitatea izango da, zorizko ibilbideak mugarik gabe jarraituko duela suposatuz. Beraz, N nodo $(n_1, ..., n_n)$ dituen G grafoa emanik, n_i nodoaren *PageRank* probabilitatea kalkulatzeko ondoko metrika aplikatzen du algoritmoak:

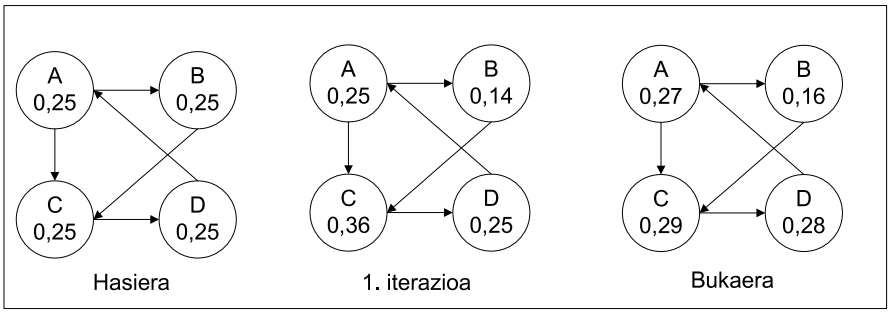
$$P(n_i) = c \sum_{n_j \epsilon In(n_i)} \frac{1}{d_j} P(n_j) + \frac{1-c}{N}$$

non c (*damping factor*) zenbaki erreal bat den, normalean 0,85-0,95 tarteko balioa du, eta urrats bakoitzean ibiltariak grafoa jarraitzeko edota zorizko saltoa egiteko probabilitatea adierazten du. $In(n_i)$, n_i nodora apuntatzen duten nodo multzoa da eta d_j, azkenik, n_j nodotik irteten diren arkuen kopurua.

Nodo bateko pisua, beraz, baturan parte hartzen duten bi faktorek definitzen dute. Lehenbiziko faktorean edozein n_j nodotik bere irteera arkuak jarraituz n_i nodora ailegatzeko probabilitatea balioesten da $(\frac{1}{d_j} P(n_j))$. Eta bigarren faktorean berriz, zorizko ibilbidean beste edozein nodotara zoriz salto egiteko probabilitatea nodo guztientzat bera dela adierazten da $(\frac{1-c}{N})$.

IV.3 irudian, lau nodotako grafo sinple baten nodoen pisuak aurkezten dira *PageRank*en 1. iterazioa eta azken iterazioa aplikatu eta gero. c faktorea enpirikoki 0,85ekin finkatuz gero, *PageRank* algoritmoaren iterazioa IV.4 irudian agertzen dena litzateke.

PageRank algoritmoaren arazoa ordea, testuinguruarekiko erabat independentea dela da. Alegia, grafo bat emanik, bere nodoen PageRank balioa finkoa da, nodoak grafoan duen garrantzi estrukturala soilik neurtzen baitu. *PageRank Pertsonalizatua* (*Personalized PageRank*) (Haveliwala, 2002) izeneko aldaerak, berriz, grafoko zenbait nodoei pisu handiagoa emateko aukera



IV.3 Irudia: Grafo bat eta alboan PageRank-en nodoen pisaketak

c $\Sigma_{nj \in IN(ni)} 1/d_j P(n_j)$	(1-c)/N
$P(A^1) = 0{,}85 \times P(D^0) \times 1{,}0$	+0,15 x 0,25 = 0,25
$P(B^1) = 0{,}85 \times P(A^o) \times 0{,}5$	+0,15 x 0,25 = 0,14
$P(C^1) = 0{,}85 \times (P(A^o) \times 0{,}5 + P(B^0) \times 1{,}0)$	+0,15 x 0,25 = 0,36
$P(D^1) = 0{,}85 \times P(C^0) \times 1{,}0$	+0,15 x 0,25 = 0,25

non $P(A^i)$, A nodoak i. iterazioan duen PageRank balioa den.

IV.4 Irudia: IV.3 irudiko grafoko *PageRank* metodoaren lehen iterazioa.

ematen du, nodo horien garrantzia handituz, eta pisu horiek grafoan zehar barreiatzea posible egiten du.

PageRank probabilitatearen kalkulua matrizeen bitartez adieraziz gero, ondoko formula definitu daiteke:

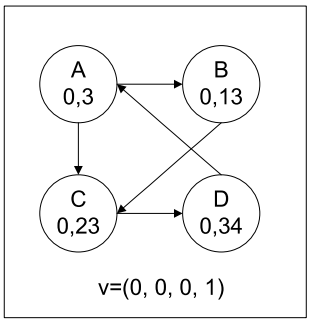
$$P = cMP + (1 - c)v$$

non M matrizea, $n \times n$ tamainako transizio matrizea den, $M_{ji} = \frac{1}{d_i}$ $n_i$tik $n_j$ra arkurik existitzen bada, eta zero bestela; eta v $n \times 1$ tamainako bektorea, non osagai guztien balioa $\frac{1}{n}$ den.

Esan bezala, *PageRank* algoritmoan v bektorea uniformeki banatua dago, zorizko jauzi bat gertatzen denean zorizko ibilbidea kalkulatzeko nodo guztiei probabilitate bera esleituz. *PageRank Pertsonalizatua*n, ordea, ez da banaketa uniformea egiten. v bektoreko osagaiek balio desberdinak izan ditzakete, zorizko jauzi baten aurrean, atazarako garrantzitsuagoak diren nodoetara

salto egiteko probabilitateak altuagoak izanik, algoritmoa adierazgarriagoak diren nodo horietara jauzi egitera bideratzea lortuz.

Beraz, IV.3 irudian azaldutako grafoaren gainean grafoko beste nodoetatik D nodoarekin lotuena zein den ebazteko, hasierako bektoreko banaketa aldatzen duen *PageRank Pertsonalizatua* aplikatuz gero, IV.5 irudian azaltzen diren pisuak lortuko lirateke. Horretarako, v bektorea D nodoa ez diren gainerako nodoetako balioak 0rekin hasieratzen dira, eta D nodoarena 1ekin. Horrela, probabilitateak kalkulatzean, D nodora iristen diren ibilbideei garrantzi handiagoa emanez, nodoaren pisuan eragina duten bi faktoreetako bigarrenean nodo guztiek jada ez baitute distribuzio bera izango.



IV.5 Irudia: PageRank Pertsonalizatuaren emaitza

UKB algoritmoa Agirre eta Soroa (2009) eta Yeh *et al.* (2009) egileen lanetan aurkezten den bezala, hitzen adiera-desanbiguazioa zein bi testuren antzekotasuna neurtzeko *PageRank Pertsonalizatua*n oinarritzen da. Oro har UKB erabili ahal izateko bi baliabide definitu behar dira:

- Ezagutza-basea, grafo moduan errepresentatuta.
- Hiztegia, non hitzen eta ezagutza-baseko nodoen arteko esteka definitzen den.

Hortaz, UKBk desanbiguazio probleman kontzeptu anbiguo baten testuingurua emanik, kontzeptu horri dagokion ezagutza-baseko nodoen artean adiera egokiena aukeratu beharko du. Horretarako, Agirre *et al.* (2010) egileek deskribatzen duten bezala, UKBk testuinguruan aipatzen diren kontzeptuei dagozkien ezagutza-baseko nodoak, hiztegian definitutako nodoak, alegia, v bektorean balio berdinarekin hasieratzen ditu (eta zero balioarekin bestelakoak), eta ondoren, *PageRank Pertsoalizatua* aplikatzen du. Sarrerako

kontzeptu anbiguoaren adiera diren nodoen artean *PageRank* balio altuena duen nodoa itzuliz ebazten du UKBk desanbiguazio ataza.

NEen desanbiguazioaren kasuan, NED sistemaren diseinu atalean (ikus IV.4 atala) azalduko dugu UKBk behar dituen baliabideak nola landu.

IV.3.3 Ebaluazio-neurriak

NEen desanbiguazio-sistema ebaluatzeko, IV.3.1 atalean deskribatutako bi ebaluazio-corpusak erabili dira. Biak *Euskaldunon Egunkaria* 2002ko corpusean oinarritutakoak dira, eta bietarako, automatikoki zein eskuz, NE anbiguoen agerpenei dagozkien Wikipediako sarrera egokiak lortu dira.

Edozein delarik ebaluazio-corpusa, NEen desanbiguazio sistema ebaluatzean bi ikuspegi neurtuko dira. Alde batetik, sistemak NE anbiguoak desanbiguatzeko gaitasuna, eta bestetik desanbiguazio adiera bat proposatzeko gai denean, zein asmatze-tasa duen. Horretarako, ohiko IR doitasun eta estaldura neurriak erabiliko dira, hurrenez hurren.

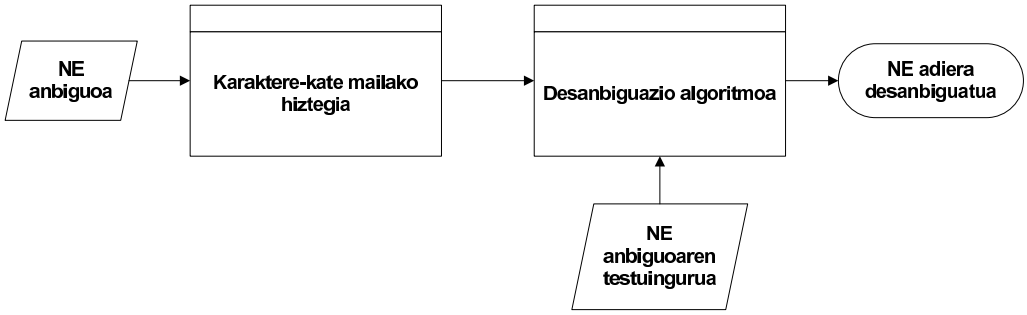
- $doitasuna = \frac{zuzen \quad desanbiguatutako \quad adibide \quad kopurua}{desanbiguatutako \quad adibide \quad kopurua}$
- $estaldura = \frac{zuzen \quad desanbiguatutako \quad adibide \quad kopurua}{sarrerako \quad adibide \quad kopurua}$
- $F_1 = \frac{2*Doitasuna*Estaldura}{(Doitasuna+Estaldura)}$

Azken neurriarekin estaldura eta doitasuna konbinatuz sistemaren ebaluazioaren ikuspegi globalagoa emateko aukera izango dugu.

IV.4 Euskarazko NED Sistemaren Diseinua

Wikipedian oinarritutako euskarazko NEen desanbiguazio sistema diseinatzean bi urrats nagusi definitu dira: lehenik NEaren agerpen bat emanik, bere adiera posibleak identifikatu; eta ondoren, hautagaien artean adiera egokiena zein den ebatzi. IV.6 irudian sistemaren arkitektura islatzen da.

Desanbiguazio urratsa algoritmoaren menpekoa bada ere, lehenbiziko urratsa edozein dela algoritmoa berdin ebatziko da, adieren hiztegi bat erabiliz. Horretarako baliabideen atalean (ikus IV.3.1.2 atala) azaldutako karaktere-kate mailako hiztegia erabiliko du sistemak. Horrela, NEaren agerpen bat sarrera izanik, hiztegian NEaren forma anbiguo horri lotutako Wikipediako sarrera nagusien multzo bat itzuliko da. Wikipediako berbideratze-



IV.6 Irudia: NED Sistemaren arkitektura

eta desanbiguazio-orriak ez dira multzo horretan egongo, hiztegi sorkuntzan ebatzi baitira horiekin lotutako anbiguotasunak.

Hiztegira beraz, *Abbadia* espresioarekin joz gero, *Anton Abbadia* eta *Abbadia jauregia* NE adierak eskuratuko dira. Aldiz, *Anton Abadia* batetik eta *Abbadia Gaztelua* eta *Abbadia Jauregia* bestetik baztertu izan dira, bi sarrera nagusietara bideratzen duten berbideratze-orriak baitira, hurrenez hurren, *Abadia* desanbiguazio-orria bezala.

Arestian esan bezala, desanbiguazio prozesua erabiltzen den algoritmoaren menpekoa denez, tesi-lan honetan prozesu hori deskribatzeko metodo bakoitzeko azpiatal bat erabili dugu.

Hautagaiak eta NEen agerpenaren testuinguruarekin, beraz, sistemaren bigarren moduluak desanbiguazioa burutzeko beharrezkoa den informazio guztia izango du. Testuingurua NEaren agerpenaren paragrafoak definitzen du.

Euskararen izaera eranskaria dela eta, paragrafoan dagoen informazioa prozesuetara pasatzean aukera anitz dago. Aukera horien artean, Wikipedia bera zein testuingurua definitzen duten testuak, lematizatu zein lematizatu gabe (soilik tokenizazio aplikatuz) erabiltzea erabaki da, informazio desberdin horien aurrean algoritmoen portaerak aztertu ahal izateko. Hori dela eta, algoritmoen menpekoak diren baliabideak sortzean bi estrategia horiek kontuan hartu behar dira, eta hainbat kasutan baliabideak bertsio bakoitzerako sortu beharko dira, ondorengo azpiataletan azaltzen den bezala.

Tokenizatu zein lematizatutako modalitatean, datuak prestatzeko garaian, bietan zaratatsuak diren eta informaziorik eskaintzen ez duten hitzak desanbiguazio prozesutik at utzi ditugu. Tokenizatutako esperimentuetara-

ko, *stopword* zerrenda[10] bat erabili dugu, non esanguratsuak ez diren hitzak zerrendatzen diren. Beraz, tokenizatutako esperimentuetarako datuak prestatzean, zerrenda horretan agertzen diren hitz guztiak ezabatuko dira, zaratatsuak diren osagaiek desanbiguazio prozesuan eragina izan ez dezaten.

Lematizatutako esperimentuetarako datuak prestatzean aldiz, *stopword* zerrenda bat erabili ordez, informazio morfosintaktikoan oinarritutako hitzen ezabaketa estrategia aplikatu dugu. Estrategia horretan, aditz, izen edo adjektibo ez den edozein hitz ezabatu egiten da, euskarazko testu sailkapenari buruzko Arregi eta Fernández (2002) egileen lanean egiten den antzera.

Oro har, edozein izanik desanbiguaziorako erabilitako algoritmoa, euskarazko NED sistemaren arkitektura beti berdina izango da: sarrerako NE anbiguorako adiera bat itzuli behar du, hautagaien artean bat aukeratzeko gai denean; eta adiera bat hautatzeko gai ez denean, bi funtzionalitate esperimentatu ditugu: isila eta hiztuna. Izaera isilean, adiera bat hautatzeko gai ez denean, sistemak ez du ezer itzultzen, alegia NIL erantzuna ematen du. Izaera hiztunean, aldiz, heuristiko bat aplikatuz hautagaietako bat aukeratu eta itzuliko du. Horrelakoetan gerta daiteke sistemak lehen urratsetik sarrerako NEaren agerpenerako adierarik ez eskuratzea. Ondorioz, izaera hiztunean kasu horietan NIL erantzuna itzuliko du sistemak.

Beraz, gure euskarazko NED sistemak bi egoeratan itzuliko du NIL erantzuna:

- Berdin du izaera, isila edo hiztuna izan, hiztegi atzipenetik adierarik eskuratzen ez denean, sistemak NIL erantzungo du.
- Izaera isilean, adiera egokienen artean berdinketa dagoenean sistemak erabakirik ez du hartuko eta NIL erantzungo du.

Esan behar da sistemak ez duela tratamendu berezirik egiten aukeratutako adiera egokia den edo ez aztertzeko. Alegia, adiera egokia ez duten NEen agerpenen kasuak identifikatzeko tratamendu berezirik ez dugu egiten. Esaterako sailkatzaile bitar bat aplikatuz behin desanbiguazioa burututa adiera egokia den edo ez ebatz daiteke, lan bibliografikoan aurkeztu diren hainbat sistematan egiten den bezala.

Kasu horien identifikaziorako hasierako ideia, metodo ez-gainbegiratu bat aplikatzea zen. Konkretuki, garapen-corpusetako adibideetan algoritmoen

[10] *IXA taldea*n hainbat proiektutan erabili den zerrenda da. Bertan maiztasun handieneko hitz funtzionalak biltzen dira.

desanbiguazio-probabilitateak aztertu eta atalase-balio minimo bat identifikatzea zen asmoa. Azterketa hori burutu bada ere, emaitzen atalean (ikus IV.5 atala) ikusiko dugun bezala, hain dira txikiak erabili ditugun baliabideak, ezen algoritmoen adibideen arteko desanbiguazio-probabilitate aldeak oso txikiak direla, ezinezkoa bihurtuz atalase minimo horren definizioa. Hortaz, etorkizunerako lan gisa utzi behar izan dugu NIL kasuen tratamendu sakonagoa egitea.

Edozein kasutan, NIL kasuen tratamendu zabala egin ez bada ere, esan bezala, sistemak NIL erantzunak itzultzen ditu. Eta NIL edo beste adiera bat itzultzeko erabakia ebazteko erabilitako algoritmo desberdinak nola aplikatu diren deskribatzen da jarraian.

IV.4.1 MFS

Maiztasun handieneko adiera itzultzeko heuristikoa euskarazko NEen desanbiguazioan erabili ahal izateko, ikasketarako corpus etiketatu bat beharrezkoa da, desanbiguazio-adiera bakoitzaren agerpen kopurua kontatu ahal izateko baliabidea, alegia. Esan bezala, tesi-lan honetarako 2006ko martxoko euskarazko Wikipedia erabili da. NEen adiera-desanbiguazio gisa desanbiguazio- eta berbideratze-orriak ez diren orrien tituluak erabili dira.

Wikipediako sarrera nagusietako titulu bakoitzeko, Wikipediako beste sarreretako zenbat testutatik erreferentziatzen den kontatuz kalkulatzen dira sarreren agerpenak. Kontaketa horretarako, beraz, Wikipediako sarrera testuetan agertzen diren estekak identifikatzea da lehen urratsa. Esteka horiek nagusiki bi formatutan topatuko dira: NEaren forma aipatzen duen NEaren Wikipediako sarreraren tituluarekin bat datorrenean, eta testuko forma eta sarrera nagusiko titulua bat ez datozenean.

Lehenbiziko eta sinpleenaren adibidea *Zazpiak Bat* sarrera da, *Anton Abbadia* erreferentziatzen duena (ikus IV.4.1 adibidea).

IV.4.1 Adibidea Wikipediako *Zazpiak Bat* sarrerako *Anton Abbadia* erreferentziatzen den testu zatia.

Zazpiak Bat [[Euskal Herria/Euskal Herriko]] zazpi [[Euskal Herriaren banaketa administratiboa/lurraldeen]] bateratzearen aldeko goiburua da: [[Araba]], [[Bizkaia]], [[Gipuzkoa]], [[Lapurdi]], [[Nafarroa]], [[Nafarroa Beherea]] eta [[Zuberoa]]. Lema ***[[Anton Abbadia]]*** *euskaltzaleak sortu zuen.*

Bigarrenaren adibidea *Zazpiak Bat* sarreran agertzen diren *Euskal Herria* eta *Euskal Herriaren banaketa administratiboa* NEekin gertatzen dena

da. Lehenbizi Wikipediako sarrera nagusiko titulua aipatuko da, eta ondoren testuan erakutsiko den aldaera. *Euskal Herria* sarreraren aldaera gisa *Euskal Herriko* agertuko da eta *Euskal Herriaren banaketa administratiboa* sarrerarena *lurraldeen* aldaera izango da.

Esteken egitura hori dela eta, baliabideen atalean (ikus IV.3.1.1 atala) aipatzen den *Wikipedia-Miner* tresnarekin lortutako fitxategietan oinarrituz kalkulatu dira kontaketak, sarrera nagusi bakoitzeko aipatua izan den kopurua gordez. Baliabide hori izango da MFS algoritmoaren oinarri. Kasu honetan, Wikipediak berak eskaintzen dituen esteken egituran oinarrituta egindako kontaketak direnez lematizazioak ez du eraginik (lotura bera lema izan ohi baita), eta beraz sarrerako testuingurua lematizatua egon ala ez baliabidea bera izango da.

Wikipedian oinarritutako maiztasun handieneko adiera itzultzen duen heuristikoak, beraz, sarrerako testuko NE anbiguorako, karaktere-kate mailako hiztegitik eskuratutako hautagaien artean adiera egokia itzultzeko agerpen kopuru handiena duen hautagaia hartuko du.

Adibide batekin ikusiko dugu. Demagun sarrerako testuan IV.4.2 adibideko *Abbadia* NEaren agerpena dagoela:

IV.4.2 Adibidea *Abbadia* NEaren agerpena duen testua.

Baina, ororen gainetik, euskaltzale amorratua zen Abbadia. Euskal literaturaren pizkundea, neurri handi batean, berari zor diogu.

Antton Abbadia eta *Abbadia jauregia* hautagaien artean erabakitzeko algoritmoak bi hautagai horiek euskarazko Wikipedian duten agerpen kopurua erabiliko du. Jakinik *Antton Abbadia* 23 agerpenekin maiztasun handieneko adiera dela, *Abbadia jauregia* sarreraren 13 agerpenen gainetik, heuristiko sinple horrek *Antton Abbadia* itzuliko du desanbiguazioa adieratzat, adiera egokia kasu honetan.

IV.4.2 VSM

Bektore-espazioen ereduak, IV.3.2.2 atalean deskribatzean aurkeztu den bezala, bi testuren antzekotasuna neurtzeko testuak dimentsio anitzeko bektore-espazioan irudikatzen ditu eta, behin espazioan irudikatuta, bektoreen arteko kosinua kalkulatzen du bi testuen antzekotasun-neurria balioesteko.

Bektore-espazioen eredua NEen desanbiguazioan aplikatu ahal izateko, beraz, testuak errepresentatzeko bektore-espazioaren dimentsioak definitzen

dituen hiztegia sortu behar da. Helburua euskarazko Wikipedian oinarritutako NED sistema bat garatzea izanik, Wikipedia bera erabili da bektoreen dimentsioak definitzeko.

Hiztegiaren sorkuntzarako, euskarazko Wikipediako sarreren testuak erabili dira. Lematizatu gabeko esperimentuetarako sarrera guztien tokenak, eta lematizazioa aplikatuko duten esperimentuetarako aldiz, Wikipedia lematizatuko sarrera guztietako lemak erabiliz.

Testu baten errepresentaziorako, bektorearen dimentsioak ezagutzeaz gain, testu bakoitzak dimentsio bakoitzean duen *tf-idf* balioa ezagutu behar da. NED sistemak, testuingurua definitzen duten testuak, bektoreak azken batean, Wikipedia sarrerekin konparatuko ditu behin eta berriz, azken horiek beti berdinak izango direlarik, Wikipedia sarrerak ez baitira aldatuko.

Beraz, euskarazko Wikipediako sarrera bakoitza bektore-espazioan errepresentatzeko, hiztegia sortu ahala, token edo lema bakoitzak testu bakoitzean duen garrantzia adierazten duen *tf-idf* balioa biltegiratuko da. Behin hiztegia eta *tf-idf* kalkuluak bukatuta, Wikipediako sarrera bakoitzari dagokion bektorea sortu eta gorde egingo da, dimentsio bakoitza adierazten duen token edo lema sarreran agerpenik baldin badu dagokion *tf-idf* balioa ezarriz, eta zero balioa esleituz bestela. Horrela aurreratzean, edozein testu Wikipediako sarrera batekin bektore-espazioan konparatu nahi izanez gero, testuari dagokion bektorea kalkulatu beharko da, Wikipediako sarrerako bektorea eskuratu besterik egin beharko ez delarik, inolako kalkulurik egin gabe.

Bide horretan, NE anbiguo baten testuingurua karaktere-kate mailako hiztegitik jasotako Wikipedia sarrera bakoitzarekin konparatu ahal izateko testua euskarazko Wikipediaren dimentsio anitzeko espazioan errepresentatzea izango da lehen urratsa. Horretarako, euskarazko Wikipediaren hiztegia erabiliko da, sarrerako testuko token edo lema bakoitza euskarazko Wikipedian dimentsio anitzeko espazioan errepresentatuta badago, sarrerako testuan tokenak edo lemak duen *tf-idf* balioa dagokion bektoreko posizioan esleitzeko. Euskarazko Wikipediaren hiztegian definituta dauden eta testuinguruko testuan agertzen ez diren tokenei edota lemei dagokien bektore posizioek, berriz, zero balioa izango dute. Azkenik, sarrerako testuan euskarazko Wikipedia espazioan errepresentatuta ez dagoen tokenik edota lemarik badago, informazio hori ez da erabiliko, ez baitu dimentsiorik definituta izango Wikipedia sarrera bektoreetan, eta konparaketarako informazio erabilgarria ez denez, ereduak baztertu egiten du.

Sarrerako testua eta desanbiguazioa-adierak izan daitezkeen Wikipediako sarrerak bektore-espazio berean adierazita izanik, sistemak sarrerako tes-

tuaren bektorearen eta Wikipedia sarrera horietako bakoitzaren bektorearen arteko kosinua kalkulatuko du, baliorik altuena lortzen duen bikoteko Wikipedia sarrera itzuliko delarik desanbiguazio adiera gisa. Wikipedia sarrera guztien konparaketan kosinuaren balioa zerokoa denean, alegia, sarrerako testuak eta Wikipediako sarrerek token edo lema komunik ez dutenean, sistemak ez du adierarik itzuliko, isilik geratuko da.

Adibide bat erabilita testu berri bat iristean, sistemak ematen dituen urratsak errepasatzen dira jarraian. Demagun sistema *Abbadia* NEaren agerpena desanbiguatzeko IV.4.3 adibideko testuingurutik abiatzen dela.

IV.4.3 Adibidea *Abbadia* NEaren agerpenaren testuingurua.

Abbadiak Etiopiara bidaian ikertutakoaren testigantza dugu gerora Rimbaudek erabiliko zuen Amaria hiztegia.

Sistemak sarrerako testuaren bektorea kalkulatu beharko du, euskarazko Wikipediaren hiztegia erabiliz. Suposa dezagun adibide sinplifikatzeko, *Etiopiara, bidaian,* eta *jauregia* tokenak dituen hiztegia erabiltzen duela sistemak, bektoreko posizioak 0, 1 eta 2 izango direlarik, hurrenez hurren. Testuinguruaren bektorea, beraz, $T_{Abbadia} = (tf\text{-}idf_{Etiopiara}, tf\text{-}idf_{bidaian}, 0)$ itxura izango du.

Bestetik *Abbadia* NE anbiguorako karaktere-kate mailako hiztegitik eskuratutako Wikipediako bi sarrerak IV.4.4 adibidekoak izango dira.

IV.4.4 Adibidea Karaktere-kate mailako hiztegian *Abbadia* aldaerarako definituta dauden Wikipediako bi sarreren testuak.

Anton Abbadia Wikipediako sarrerako testua: *Anton Abbadia Urrustoi geografo, astronomo eta euskal kulturaren eragilea izan zen. Etiopiara jo nahi zuen, baina lehenago Brasilera joan behar izan zuen, 1835an, Zientzia Akademiak eskatuta, bidaian lurraren magnetismoa aztertzera*

Abbadia Jauregia Wikipediako sarrerako testua: *Abbadia jauregia Hendaiako (Lapurdi) itsaslabarren gainean dago, Santa Ana muturreko muinoan.*

Sistemak inolako kalkulurik burutu gabe sarrera horiei dagozkien bektoreak eskuratuko ditu, adibideko hiztegian oinarrituz ondoko itxura izango dutenak:

- $W_{AntonAbbadia} = (tf\text{-}idf_{Etiopiara}, tf\text{-}idf_{bidaian}, 0)$
- $W_{Abbadiajauregia} = (0, 0, tf\text{-}idf_{jauregia})$

Hiru bektoreak definituta, sistemak bi Wikipedia sarreren artean egokiena aukeratzeko testuinguruko bektorea bi bektoreekin konparatuko du:

- $\cos(T_{Abbadia}, W_{AntonAbbadia})$
- $\cos(T_{Abbadia}, W_{Abbadiajauregia})$

Kosinu baliorik altuena lortzen duen kalkuluaren Wikipedia sarrera itzuliko duenez sistemak, kasu honetan *Anton Abbadia* itzuliko du desanbiguazio-adiera gisa.

IV.4.3 ESA eta ESA orekatua

ESAren oinarrizko ideia edozein hitz Wikipediako sarreren bektore bezala adieraztea da. Bektoreko posizio edo dimentsio bakoitza Wikipediako kontzeptu bat izango da eta posizio horren pisuak hitzak Wikipediako sarreran duen garrantzia adieraziko du. NEen desanbiguazio atazan aplikatzeko ordea, Wikipediako sarrera guztien bektorea sortzea ez da beharrezkoa izango, IV.3.2.3 atalean atalean azaldu den bezala. NE anbiguo jakin baterako karaktere-kate mailako hiztegitik eskuratzen diren Wikipedia sarreren bektoreak sortzea nahikoa da, horietako bat aukeratzea izango baita sistemaren helburua.

Edozein kasutan, ESA euskarazko Wikipediarekin erabili ahal izateko, euskarazko Wikipediako sarreretan agertzen den hitz bakoitzerako *tf-idf* balioa kalkulatu beharko da, hitz guztiekin hiztegi itxura duen baliabidea sortuz. Horretan oinarrituz, exekuzio-garaian erabakiko da NEaren desanbiguaziorako zeintzuk diren helburuko Wikipedia sarrerak. Aurrerantzean alderantzizko indize bezala aipatuko den baliabide horretan, beraz, hitz bakoitzari Wikipedia sarrera bakoitzean dagokion *tf-idf* balioa definituta egongo da. Aurretik esan bezala (ikus IV.3.2.3 atala) baliabidea behin bakarrik definitu beharko da euskarazko Wikipediarako, baina esperimentuetan Wikipediaren bertsio lematizatua eta ez lematizatua erabiliko denez, bertsio bakoitzerako alderantzizko indizeen baliabide bat sortu beharko dugu, lemekin eta tokenekin eraikiz, hurrenez hurren.

Sistemak NE anbiguo baten desanbiguazioa burutzeko, karaktere-kate mailan adiera gisa definitutako Wikipediako sarrera multzoa eta NEaren agerpenaren testuinguru testua jasoko ditu. Testuko token edo lema bakoitzeko, lematizatu gabeko edo lematizatutako alderantzizko indizea atzi-

tuz, hurrenez hurren, hitz bakoitzerako *tf-idf* bektorea sortuko du Wikipediako sarrera multzoaren dimentsioan. Bektore horien baturarekin, algoritmoak Wikipedia-bildumako sarrera bakoitzeko antzekotasun-neurria lortzen du, ESA antzekotasun-neurria, alegia. Azkenik, antzekotasun balio altuena lortzen duen sarrera itzuliko du adiera-desanbiguazio gisa.

Ingurune esperimentalean metodoen atalean azaldutakoaren ildotik (ikus IV.3.2 atala) esperimentuetan bi ESA desberdinen emaitzak aurkeztuko dira. Funtsean berdin funtzionatzen dute, bektoreen baturetan aplikatutako normalizazio-faktorean dago desberdintasun bakarra. ESA orekatuak (bESA aurrerantzean), alderantzizko indizeetan oinarritutako ESA antzekotasun-bektoreko posizio bakoitza faktore berri batekin normalizatzen du, hain zuzen ere. Faktore berri horrek normalizaziorako Wikipediako sarrera eta testuinguruko testuko token edo lema kopuru komuna eta testuinguruaren luzera kontuan hartzen ditu ingurune esperimentalean azaldu den bezala (ikus IV.3.2.4 atala).

Aurreko *Abbadia* NEaren agerpena desanbiguatzeko adibidea erabiliz, IV.4.3 adibideko agerpenaren testuingurua hain zuzen ere, sistemak ESA edo ESA orekatua aplikatuz ematen dituen urratsak errepasatuko ditugu.

Suposatuz, VSMren azalpenean bezala, *Abbadia* espresiorako *Anton Abbadia* eta *Abbadia Jauregia* direla karaktere-kate mailako hiztegian definituta dauden bi adierak, sistemak adiera bakoitzerako sarrerako testuarekin duen antzekotasuna balioetsiko du. Horretarako, Wikipediako bi sarrera horietako bakoitzeko, *Abbadia* testuinguruan agertzen diren hitzek Wikipediako sarreran duten *tf-idf* balioak eskuratuko ditu alderantzizko indizea atzituz, eta balio horien guztien batura eginez, ESA antzekotasun-neurria lortuko du. Bi antzekotasun balioak ESAren kasuan, beraz, ondokoak izango dira:

- $ESA(T_{Abbadia}, W_{AntonAbbadia}) = (tf\text{-}idf_{Etiopiara,T_{AntonAbbadia}} * tf\text{-}idf_{Etiopiara,W_{AntonAbbadia}}) + (tf\text{-}idf_{bidaian,T_{AntonAbbadia}} * tf\text{-}idf_{bidaian,W_{AntonAbbadia}})$
- $ESA(T_{Abbadia}, W_{Abbadiajauregia}) = 0$, ez baitago testuinguruan hitzik Wikipediako *Abbadia jauregia* sarrerarekin bat datorrenik.

ESA orekatuaren kasuan, antzekotasun balioaren kalkuluan, normalizazio-faktore berria kontuan hartuko da antzekotasuna balioestean:

- $bESA(T_{Abbadia}, W_{AntonAbbadia}) = ESA(T_{Abbadia}, W_{Abbadiajauregia}) * 2/9$[11]

[11] Kontuan izan sistemak *stopword*en zerrendan agertzen diren hitzak ezabatu egiten

- $bESA(T_{Abbadia}, W_{Abbadiajauregia}) = 0$

Edozein kasutan, ESA edo ESA orekatua aplikatuta desanbiguazio emaitza gisa antzekotasun baliorik handiena ematen denez, adibide honetan, bietan sistemak *Anton Abbadia* sarrera hautatuko du, txikia izanda ere antzekotasuna bi algoritmoetan 0 baino handiagoa izango baita.

IV.4.4 UKB

Euskarazko Wikipedia erabiliz UKB algoritmoaren bitartez NEen desanbiguazioa burutu ahal izateko, arestian aipatu den bezala (ikus IV.3.2.5 atala), bi oinarrizko baliabide sortu behar dira:

- Euskarazko Wikipedia grafo gisa errepresentatuta.
- Hitzen eta Wikipediaren grafoko nodoen arteko estekak definitzen dituen hiztegia.

Euskarazko Wikipediaren grafoan, Wikipediako sarreren tituluak nodoak izango dira, eta sarrera horien deskribapenean agertzen diren beste sarrera batzuetarako estekak berriz, arkuak izango dira. Grafo horrekin, euskarazko Wikipediaren egitura errepresentatuko da.

NE anbiguo baten testuingurua deskribatzen duen testu bati UKB aplikatu ahal izateko, beraz, testuan agertzen diren osagaiak grafoaren nodoekin lotu beharko dira. Horretarako, NEaren forma anbiguoak eta Wikipediako sarreraren estekak gordetzen dituen hiztegia erabiltzen da, IV.3.1.2 atalean definitutako baliabidea, hain zuzen ere. Azken batean, UKBk erabiltzen duen *PageRank Pertsonalizatua*k grafoaren konektibitatean oinarritzen da desanbiguazioa burutzeko, algoritmoa abiatzean aktibatuko diren nodoak testuinguruan agertzen diren osagaiei dagozkionak izanik.

Hortaz, NE anbiguo baten adiera egokiena zein den ebazteko, batetik UKBk NEaren agerpenaren testuinguruan agertzen diren osagaiak Wikipediako grafoko zein nodoekin lotuta dauden jakiteko karaktere-kate mailako hiztegia atzituko du. Horrela, osagai bakoitzeko hiztegia atzituko du eta bertan osagai horri lotuta dauden nodoak aktibatu egingo ditu grafoan.

---

dituela Wikipediako sarreren testuetatik zein testuinguru testuetatik, antzekotasun balioak ahalik eta zarata gutxienarekin balioesteko. Hori dela eta, sarrerako testuan 12 hitz badaude ere, soilik 9 erabiliko dira, 3 *stopword*ak izanik baztertu egiten baitira.

Eta bestetik, NEaren agerpenaren adierak zeintzuk diren identifikatzeko ere, karaktere-kate mailako hiztegia atzituko du. Irteeran, adiera horiek errepresentatzen dituzten nodoetako bat aukeratzea denez helburua, UKBk *PageRank Pertsonalizatuaren v* bektorearen hasieratze urratsean nodo horiei gainerakoei baino pisu handiagoak esleituko dizkie, sistemak adiera gisa horietako nodo bat eta ez beste edozein nodo proposatzen duela bermatzeko.

Nodo guztiak aktibatuta eta pisatuta, UKBk ibilbideak kalkulatuko ditu probabilitate altueneko ibilbideko helburu nodoa, desanbiguazio emaitza gisa proposatuz.

Jarraian azaltzen den emaitzen atalean ikusiko dugun bezala, beste metodo guztietan tokenizatutako esperimentuek lematizatutako esperimentuek baino emaitza hobeak lortzen dituztela ikusita, eta UKB algoritmoa lematizatutako datuekin ebaluatzeko aurkezten dituen zailtasunak direla eta, UKB soilik tokenizatutako datuekin ebaluatzea erabaki dugu. Hori dela eta, algoritmoak beharrezkoak dituen grafoa zein hiztegia soilik tokenizatutako esperimentuetarako prestatu ditugu.

IV.5 Emaitzak

Atal honetan aurreko ataletan azaldutako metodoak erabilita *Euskaldunon Egunkaria*ko artikuluetako NEen agerpenak euskarazko Wikipedia sarrera batekin lotzeko atazan egindako ebaluazioa aurkeztuko da.

IV.4 atalean azaldu den bezala, NE anbiguo baten desanbiguazio-prozesuan sistema egoera desberdinen aurrean aurki daiteke. Alegia, NEaren desanbiguazio-hautagaien bila karaktere-kate mailako hiztegira jotzean, hautagai bat edo gehiago eskuratzea gerta daiteke. Baina hautagairik ez aurkitzea ere gerta daiteke. Azken kasu horretan ez dago aukerarik eta sistemak ezingo du hautagairik proposatu. Horrelakoetan sistemak NIL itzuliko du.

Hautagai bat baino gehiago eskuratzean, sistemak hautagaiak pisatuko ditu eta metodoaren pisu altuena duena itzuliko du desanbiguazio-emaitza bezala. Pisu altuenen berdinketak eman daitezkeela ere kontuan izan behar da, ordea. Alegia, pisu altueneko hautagai bakarra beharrean, bi edo gehiago egotea. Horrelakoetan berdinketak ebazteko bi prozedura desberdin ebaluatu dira:

- Isila: Sistemak ez du erabakirik hartzen berdinketa kasuetan, eta beraz ez du hautagairik itzultzen. Alegia, NIL itzultzen du.

- Hiztuna: Berdinketa ebazten da eta hautagai irabazlea itzultzen da. Berdinketa hausteko, bi metodo ebaluatu dira: batetik zorizko hautapena eta bestetik maiztasun handieneko adiera itzultzea.

Jarraian sistema isilen zein hiztunen ebaluazioak aurkezten dira. Lehenbizi datu ez lematizatuekin egindako esperimentuak aurkezten dira eta, ondoren, datu lematizatuekin egindakoak.

IV.5.1 Sistema isilak

Baliabideen atalean deskribatu den bezala (ikus IV.3.1.3 atala), bi motatako ebaluazio-corpusak bereiztu ditugu: A corpusa (CorpusA tauletan), non testuetan agertzen diren NEen agerpen guztiek euskarazko Wikipedian beraien adierek sarrera definituta duten; eta B corpusa (CorpusB tauletan), non testuetako hainbat NEen agerpenen adierek Wikipedian sarrerarik ez duten.

Izaera desberdineko corpusak izanik, ezin dira corpus desberdinetako emaitzak elkarren artean konparatu, are gutxiago sistemaren izaera isila ebaluatzean. Hori dela eta, sistema isilen ebaluazioak corpusen arabera aurkezten ditugu jarraian.

IV.5.1.1 A corpusa

A corpusaren ebaluazioa IV.2 taulan aurkezten da. Ikus daitekeenez, ESA algoritmoa ebaluatutako sistema guztien artean MFS sistema sinpleena gainditzea lortzen ez duen bakarra da, doitasunean 3 puntu gainetik geratzen den arren, estaldura oso txikia lortzen baitu. Espero ez zen arren, VSM sistemak ere ESA sistema baino emaitza hobeak lortzen ditu.

ESA algoritmoari normalizazio-faktorea gehitutako bESA sistemak, aldiz, estalduran ere MFSek baino portaera okerragoa badu ere, doitasuna altuagoa duela eta, MFSek baino % 3ko F_1 hobea izatea lortzen du.

UKB sistema MFS sistemaren estaldura gainditu duen bakarra izan da. Doitasunean ere emaitzarik onenak lortu izanak, A corpusean euskarazko NED ataza hobekien ebatzi duen sistema bihurtzen du UKB, % 81,91ko F_1ekin.

Oro har, A corpus honetan edozein sistemak NIL erantzutean, alegia isilik geratzean, sistemaren hutsegitea izango da. Izan ere, aurretik esan bezala, A corpuseko adibideetako NE guztien agerpenen adierak Wikipedian

	Doitasuna	Estaldura	F_1
MFS	% 68,32	**% 68,32**	% 68,32
VSM	% 74,67	% 64,45	% 69,18
ESA	% 71,55	% 61,24	% 66,00
bESA	**% 76,78**	% 66,47	% 71,25
UKB	**% 82,85**	**% 81,01**	**% 81,91**

IV.2 Taula: A corpusaren gaineko berdinketetan isilik geratzen diren sistemen ebaluazioa

sarrera baitute. Hori dela eta, beti erantzuna ematen duen MFS ezik, sistema guztietan doitasuna estaldura baino hobea izango da. Doitasunean isilik geratzearen hutsegiteak ez baitira kontuan hartzen.

Sistema guztien artean UKB da, berriz ere, bi neurriak konparatzean portaera egonkorrena aurkezten duen metodoa. Horren arrazoia beste sistemek baino berdinketa kasu gutxiagoarekin topo egiten duela da. Hala, NIL erantzun gutxiagoa emanik besteak baino hutsegite gutxiago egiten ditu.

Ebaluazio honetako emaitzak aztertzean, NEaren agerpenarekin karaktere-kate mailako hiztegira jotzean sarri hautagai bakarra eskuratzen dela identifikatu da. A corpusaren NEen agerpenen desanbiguazio-adiera guztiek Wikipedian sarrera dutela kontuan hartuz, inolako desanbiguaziorik burutu gabe, adiera egokia eskuratzen dela ondoriozta daiteke.

Ebaluazio errealagoa aurkezteko asmotan, corpusetik adibide ez-anbiguo horiek kendu eta sistemak berriz ebaluatu ditugu isilik modalitatean. Esperimentu horien emaitzak IV.3 taulan laburtzen dira.

	Doitasuna	Estaldura	F_1
MFS - soilik anbiguoak	% 53,00	% 53,00	% 53,00
VSM - soilik anbiguoak	% 62,38	% 52,62	% 57,10
ESA - soilik anbiguoak	% 57,75	% 47,40	% 52,00
bESA - soilik anbiguoak	% 65,51	% 56,10	% 60,44
UKB - soilik anbiguoak	% 74,50	% 71,76	**% 73,10**

IV.3 Taula: A corpusetik NE ez-anbiguoak dituzten dokumentuak ezabatuta berdinketetan isilik geratzen diren sistemen ebaluazioa

Orokorrean sistemen portaerak 10 puntu inguru jaisten dira, baina espero bezala, aldeak areagotzen dira. UKB sistemak portaera ona izaten jarraitzen du % 73ko F_1ekin desanbiguatuz. bESA sistemak, bestetik, MFSen estaldura gainditzeaz gain, ESA sistemak baino 8,5 puntu hobeto ebazten du ataza.

Gehitutako normalizazio-faktoreak, beraz, euskara bezalako baliabide murritzetan oinarritzean, corpuseko dokumentu laburren eta luzeran ematen den desorekaren eragina neutralizatzeko estrategia egokia dela ematen du.

IV.5.1.2 B corpusa

B corpuseko NEen agerpenek adiera egokia Wikipedian sarrera definituta izatea ziurtatuta ez egoteak, corpusari sistemaren portaera eta bereziki NIL kasuen ebazpena fintzeko iturri egokia izateko izaera ematen dio. Hala izanik, B corpusa bi multzotan banatuta dago: garapen eta test multzoetan.

Tesi-lan honetan, fintze lan hori etorkizunerako lan gisa utzi denez (ikus IV.3.1.3 atala), metodo desberdinak ebaluatzeko bi corpusak erabili ditugu. IV.4 eta IV.5 tauletan laburbiltzen dira horien emaitzak. Gainera, tamaina antzeko eta ezaugarri berdineko corpusak izateak, emaitzak elkarren artean konparatu ahal izatea ahalbidetzen du.

	Doitasuna	**Estaldura**	F_1
MFS	% 72,00	% 72,00	% 72,00
VSM	% 71,42	% 66,72	% 69,00
ESA	% 61,57	% 57,51	% 59,47
bESA	% 69,01	% 64,47	% 66,66
UKB	% 75,37	% 76,10	% 75,73

IV.4 Taula: B-garapen corpusaren gaineko sistema isilen ebaluazioa

	Doitasuna	**Estaldura**	F_1
MFS	% 70,40	% 70,40	% 70,40
VSM	% 69,80	% 64,40	% 67,00
ESA	% 60,73	% 56,00	% 58,27
bESA	% 68,11	% 62,80	% 65,34
UKB	% 76,25	% 75,80	**% 76,02**

IV.5 Taula: B-test corpusaren gaineko sistema isilen ebaluazioa

Bi ebaluazioetan, aurrekoetan bezala, UKB da portaera onena duen sistema, eta bietan MFS sistema gainditzea lortzen duen bakarra izaten jarraitzen du. Bestetik, A corpusean ez bezala, VSM sistemak ESA zein bESA sistemak baino emaitza hobeak lortzen ditu, F_1 neurrian 9 eta 1,6 puntu inguru gainetik, hurrenez hurren.

Corpusen tamaina eta izaera desberdina dela eta, A eta B corpusetako emaitzak zuzenean konparatu ezin badira ere, sistema bakoitzaren doitasuna eta estaldura neurrien arteko aldeei erreparatu diegu. Bi neurrien azterketan ikus daiteke A corpusean sistemen neurrien arteko aldea, UKB eta MFSena ezik, 10 puntukoa dela eta B corpusetan, aldiz, aldeak 5 puntu ingurura jaisten direla. Beraz B corpusetako emaitzak A corpusekoak baino okerragoak badira ere, agertoki errealagoak irudikatzen duten B corpusetan sistemak egonkorragoak direla esan daiteke. Egonkortasun horren arrazoia NIL adibideetan izan dezake jatorria. Izan ere, B corpusean sistemaren portaera isilak, hainbat adibideetan asmatzea izan daiteke, A corpusean ez bezala.

IV.5.2 Sistema hiztunak

Orain arte aurkeztutako ebaluazioetan berdinketa kasuetan sistemak isilik geratzearen estrategiarekin ebaluatu dira. Jarraian, sistemak erantzun bat ematera behartuta daudenean, hiztuna modalitateko ebaluazioak aurkezten dira.

Modalitate honetan NE baten agerpen batentzat hiztegian hautagairik ez dagoenean ere sistemak zerbait itzuli behar duenez, NIL erantzuna itzuliko du. Hortaz, hiztuna modalitatean NIL erantzuna, edozein adiera bezala erantzun posible bezala kontsideratzen da.

Sistema hiztunek desanbiguazioan beti zerbait proposatzeak doitasun eta estaldura neurriak baliokide bihurtzen ditu.

Sistema isiletan egin dugun bezala, sistema hiztunen ebaluazioak A eta B corpusen gainean burutu ditugu.

IV.5.2.1 A corpusa

IV.7 taulan A corpusarekin sistema hiztunek zorizko zein MFS heuristikoekin lortutako emaitzak aurkezten dira. Bertan, sistema isilen emaitzak ere aurkezten dira, F_1 balioaren arabera, hiru estrategiak konparatu ahal izateko.

A corpuseko ebaluazio honetan, sistema hiztun guztiek MFS sistemaren emaitzak gainditzen dituzte. Zorizko eta MFS hautaketen artean emaitzetan oso alde handia egon ez arren, espero bezala bigarren aukera gailentzen da kasu guztietan. Sistema hiztunetan, aurreko ebaluazioetan bezala, UKB algoritmoan oinarritutako sistemak desanbiguaziorik onena lortzen du % 82,8ko estaldurarekin.

Sistema	Isilik	Zorizko hiztuna	MFS hiztuna
MFS	% 68,32	% 68,32	% 68,32
VSM	% 70,15	% 74,03	% 75,53
ESA	% 66,00	% 71,00	% 72,43
bESA	% 71,25	% 75,94	% 77,66
UKB	% 81,91	**% 82,80**	**% 82,80**

IV.6 Taula: A corpusean berdinketetako aukeraketa estrategia desberdinen ebaluazioa (isilen F_1)

Sistema isil eta berdinketak ebazteko MFS heurtistikoa aplikatzen duten sistema hiztunen portaerak konparatuz gero, hiztunek ataza isilek baino hobeto ebazten dutela ikus daiteke. Hobekuntza hori, corpus honetan, espero bezala oso nabaria da, multzo horretako desanbiguazio-adiera guztiak Wikipedian existitzen direlako ezaugarria betetzen da eta. Alegia, A corpuseko adibide guztietarako sistemek adiera egokia aurki dezakete Wikipedian, beraz, beti zerbait proposatzean isilik geratzean baino asmatze-tasa altuagoa lortzeko probabilitatea hazten da. Kasu okerrenean berdinketetako adibideetan proposatutako adiera guztiak okerrak izanda ere, sistema isilaren doitasun bera lortuko litzateke.

Sistema hiztunak corpus honetan ebaluatzean, beraz, NIL erantzuna emateko beharra ez dute izango, agerpen guztientzat gutxienez hiztegian Wikipedian sarrera duen adiera bat aurkituko baitute eta, gainera, bat baino gehiago eskuratu eta desanbiguazio-prozesuan berdinketa kasuetan heuristiko bat aplikatuz horietako bat hautatuko dute, NIL erantzuna ekidinez.

UKB algoritmoan oinarritutako sistema hiztun eta isilen arteko hobekuntza ordea, ez da hain nabarmena, arrazoi nagusiena algoritmoak aurkezten duen doitasun altua da. Doitasun altua berdinketa kasu gutxiren islapen kontsideratzen bada, UKBk berdinketa ebazpen gutxiago izanik sistemaren portaera bi estrategietan antzekoagoa izatea normala da.

ESAri dagokionez izaera isilean emaitzarik okerrenak agertzen dituen algoritmoa da, bere doitasun txikia dela eta MFS heuristikoaren emaitzen azpitik geratzen da ebaluazio guztietan. Arazoaren jatorria, euskarazko Wikipediako artikuluen luzera dela uste dugu. Alegia, euskarazko Wikipedian sarrera labur ugari egonik, ESAk NE anbiguoen testuinguru diren paragrafoekin alderatzean antzekotasun-neurri sendoak lortzea zaila bihurtzen da. Zailtasun hori gainditzeko asmoz definitutako normalizazio-faktorea baliatzen duen bESA algoritmoak, aldiz, oinarrizko MFS eta ESA bera gainditzeaz gain, VS-

Mek baino portaera hobea aurkezten du, izaera isilarekin ebaluatzean gertatu den bezala.

IV.5.2.2 B corpusa

B motako bi corpusen gainean ebaluatzean, sistemek IV.7 taulako emaitzak lortzen dituzte.

Ebaluazio-corpusa - Sistema	Isilik	Zorizko hiztuna	MFS hiztuna
Garapen - MFS	% 72,00	% 72,00	% 72,00
Garapen - VSM	% 69,00	% 70,30	% 70,48
Garapen - ESA	% 59,47	% 60,9	% 61,09
Garapen - bESA	% 66,66	% 67,85	% 68,00
Garapen - UKB	% 75,73	**% 75,95**	**% 75,95**
Test - MFS	% 70,40	% 70,40	% 70,40
Test - VSM	% 67,00	% 69,00	% 70,00
Test - ESA	% 58,27	% 61,20	% 61,60
Test - bESA	% 65,34	% 68,40	% 68,40
Test- UKB	% 76,02	**% 76,20**	**% 76,20**

IV.7 Taula: B corpusetan berdinketetako aukeraketa estrategia desberdinen ebaluazioa (isilen F_1)

Berriz ere, UKB bi multzoetan % 76 inguruko estaldurarekin portaera onena duen metodoa da. Bi corpusetako emaitzetan erreparatuz gero, MFS oinarrizko sistema gainditzen duen bakarra dela azpimarra daiteke. Beraz, UKBn oinarrituz garatutako sistemak, ebaluatutako guztien artean behinik behin, euskarazko NED atazarako egonkorrenak direla esan daiteke.

Sistema isil eta berdinketak ebazteko MFS heurtistikoa aplikatzen duten sistema hiztunen portaerak konparatuz gero, A corpusean gertatzen zen bezala, hiztunek ataza isilek baino hobeto ebazten dutela ikus daiteke, zehazki 3 puntu hobeto. Oraingoan ordea, hobekuntza ez da hain nabarmena. Horren arrazoi nagusia, NIL kasuen tratamendu sakon ezan du jatorria. Izan ere, B motako corpusetan badaude adiera egokia Wikipedian sarrera ez duten NEen agerpenak, eta sistemek soilik NIL proposatuko dute, agerpen horiekin hiztegira jotzean adierarik eskuratzen ez dutenean, inolako azterketa sakonagorik egin gabe, adierak eskuratzean horien artean egokia ote dagoen erabakitzen duen lagun dezakeen azterketa sakonik gabe, alegia. Beraz, sistemaren emaitzak hobetu eta portaera sendotzeko bidean, horrelako kasuak

ebazteko tratamendu sakonagoa egitea etorkizuneko lan interesgarria izan daiteke.

B corpuseko bi multzoetako emaitzak eta A corpuseko emaitzak zuzenean ezin badira konparatu ere, ESA, bESA eta UKBren emaitzak B corpusean okerragoak direla erraz ikus daiteke. Galera hori algoritmoen doitasun baxu eta NIL aukeraren esleipen okerrean jatorria dutela esan daiteke, aurretik aipatu bezala, azken batean sistemak ez baitira diseinatu NIL erantzunei behar duten tratamendu berezia emateko.

IV.5.3 Erroreen analisia

UKB izanik metodorik onena, gainerako metodoak hobetzeko ezaugarrien bila, metodo horrek ondo eta besteek gaizki ebazten dituzten adibideak eskuz aztertu ditugu. Konparaketarako metodo guztiak erabili beharrean, UKBren ostean portaera egonkorrena aurkeztu duen ESA orekatuaren emaitzak erabili ditugu.

Oro har, Wikipediak eskaintzen duen orrien arteko esteken informazioa baliatuz eta aipamen esplizituetan soilik ez oinarritzea UKBren abantailarik handiena dela ematen du. Izan ere, desanbiguatu nahi den NEaren agerpenaren testuinguruko testuan agertzen diren espresioak, Wikipediako hautagai orrian esplizituki agertzen ez badira, UKB ezik gainerako metodoek ez dute informazio hori erabiltzen. UKBk ordea, semantikaren ikuspuntutik pausu bat gehiago ematen du. Wikipedia osoa bere esteketan oinarrituz grafo gisa errepresentatuta erabiltzen duenez, espresio bat hautagai sarreran esplizituki adierazita egon ez arren, grafoan biak erlazionatzeko biderik badago, NEaren agerpenaren desanbiguazio prozesurako informazio baliagarria izango da. Bide hori existitzeko, Wikipedian zehar hautagai sarrerak deskribatzen duen NEaren aipamenen bat eta testuinguruko espresioa lotzen dituen aipamenen bat egongo da.

Horren adibide, IV.5.1 adibideko *Abel* NEaren agerpenaren testuingurua da.

IV.5.1 Adibidea *Abel* NEaren agerpena desanbiguatzeko testuingurua eta bi Wikipedia sarrerak.

Testuingurua: Sarrerak, oro har, garestiak dira oso. Esaterako, buruz burukoa kantxan ikusteko, 70 euro ordaindu beharko dira. Era berean, ostegunean jokatuko den binakako partidaren prezioa 50 eurokoa da. Abelek aintzat hartu zuen hori. 'Sarrerak 13.000 pezetan daude, eta guztiz osatuta ez banago, ez dut jokatuko; ez dut inor engainatu nahi. Eskua sendatuz gero, badakit oso zaila izango dela Mikel Goñiri irabaztea'.

*Abel*en agerpena desanbiguatzeko garaian sistemek *Abel Barriola* eta *Abel San Martin*[12] bi sarreren artean bat aukeratu behar dutenean, bESAk esaterako *Abel San Martin* itzuliko du adiera gisa, testuinguruaren eta Wikipediako sarreren artean soilik topatzen baitu hitz komun bakarra, *binakako* hitza, hain zuzen ere.

UKBk aldiz, Wikipediako esteken egitura baliatuz, *Abel San Martin*ekin ez bezala *Abel Barriola* eta *Mikel Goñi*ren arteko erlazioa/bidea dagoela ikusten du eta erabiltzen du. Informazio horri esker, *Abel San Martin* adiera baztertu eta *Abel Barriola* adiera egokia itzultzen du UKBk.

Argi dago, beraz, Wikipediako sarreren informazioa zabaldu egin behar dela gainerako metodoek UKBren informazio erabilpena simulatu ahal izateko. Hori lortzeko bide bat, sarrera batean beste sarrera bat erreferentziatzen bada, erlazionatuta daudela suposatu eta, bigarren sarrerako testua lehenbizikoa deskribatzeko ere erabiltzea izan daiteke.

Erroreak analizatzean, UKBren asmatzeak aztertzeaz gain, ESA metodoa hobetzeko bidean definitutako normalizazio-faktorea erabiltzen duen ESA orekatuaren (bESA) eta ohiko ESAren emaitzak aztertu ditugu.

Azterketa horretan, ESA algoritmoak kale eta bESA metodoak asmatzea ematen den kasu gehienetan, bESAk, normalizazio-faktoreari esker asmatzen duela ikusi dugu. Alegia, testuinguruko eta hautagai desberdinetako testuen arteko hitz komunen kopuruak antzekoak direnean, eta hautagaien testuak luzeran motzak eta desberdinak direnean behinik behin, ESAk testu laburrena duen hautagaia proposatzeko joera duela ikusi dugu. Izan ere, hitz bera eta bere agerpen kopuru bera duten luzera desberdineko bi Wikipedia sarreretan, hitz horren *tf-idf* balioa sarrera laburrenean balio altuagoa izango du. *idf*ren balioa bietan bera izanda ere, *tf*ren balioa oso desberdina izan daiteke, testu luzeraren arabera. Azken horretan, hitzak testuan dituen agerpenak testuaren luzerarekin zatitzen direnez, luzera handiko testuko *tf-idf* balioa, laburrekoan baino txikiagoa bihurtzen da.

ESAk, NE baten agerpen bat eta Wikipedia sarrera bateko antzekotasuna neurtzeko, funtsean, testuinguruko eta sarrerako testuan bat datozen hitzen Wikipediako *tf-idf* pisuen baturan oinarritzen dela kontuan hartuz, testuinguruarekin antzeko baldintzak betetzen dituzten bi hautagairen artean beti testu laburrena duen sarrera itzultzeko joera izango du.

ESA orekatuak, berriz, bi testuen artean dauden hitz komunen kopurua kontsideratzen duenez, luzera efektu hori murriztea lortzen du, joera oker

[12]Hautagai gehiago dauden arren, bi horiek aukeratu dira azalpena sinplifikatzeko.

hori zuzenduz eta ESAk arrazoi horregatik egiten dituen hutsegite gehienak zuzenduz. Horren adibide, IV.5.2 adibideko *Indurain* agerpenaren desanbiguazioa da.

IV.5.2 Adibidea *Indurain* NEaren agerpena desanbiguatzeko testuingurua eta bi Wikipedia sarrerak.

Testuingurua: Bi txirrindulari horiek ziren nagusiak eta beste guztiak morroiak. Morroi eskertuak, hala ere, bikote horrek (Indurain nafarrak, batez ere) urteetan pilatutako palmaresa ez baita nolanahikoa izan.

Wikipediako *Indurain* sarrera: Indurain Itzagaondoako kontzejua da. 2007ko urtarrilaren 1eko erroldaren arabera, herriak 24 biztanle ditu.

Wikipediako *Miguel Indurain* sarrera: Miguel Indurain Larraya (Atarrabia, 1964ko uztailaren 16a) txirrindulari nafarra da, historiako kirolari onenen artean sailkatua. Frantziako Tourra bost urtez jarraian irabazi zuen (1991tik 1995era) eta Italiako Giroa bitan (1992 eta 1993). 1992an Kiroletako Asturiasko Printzearekin saritua izan zen. 1984an egin zen txirrindulari profesional Reynolds taldearekin kontratua sinatzean. Aurretik zortzi urtez Villaves taldean aritua zen kategoria apalagoetan. Lehen urte honetan Los Angeleseko Olinpiar Jokoetan parte hartu zuen eta hamar garaipen lortu zituen...

Testuingurua eta Wikipediako bi sarreren testuak konparatzean, *Miguel Indurain* sarrerarekin *txirrindulari*, *palmaresa* eta *Indurain* espresioa bera bat datozela ikus daiteke eta *Indurain* sarrerarekin berriz, *Indurain* espresioa soilik. ESA aplikatzean *Miguel Indurain*eko hiru hitzen *tf-idf* balioen baturak, *Indurain* sarrerako osagai komun bakarraren *tf-idf* balioa baino txikiagoa denez, ESAk *Indurain* proposatuko du desanbiguazio emaitza gisa, hutsegitea sortuz. bESAk aldiz, testuen arteko antzekotasuna neurtzean osagai komunen kopurua kontuan hartzen duenez, *Miguel Indurain* proposatuko du adiera gisa, sistemaren asmatzea eraginez.

Hiru algoritmo horien emaitzak eskuz aztertzean, adibide askotako testuinguruak Wikipediako sarrerekin oso desberdinak eta askotan Wikipediako sarrerekin komunean dituzten hitzak oso generikoak edo adieren artean bat hautatzeko nahikoak ez direla ikusi dugu. Baina hitz komunen kopuruak hain eskasak dira, askotan sistemek hitz generiko edo informazio gutxi eskaintzen duten hitz horietan soilik oinarritzen direla desanbiguazioa ebazteko, noski askotan hutsegitea eraginez.

Esaterako IV.5.3 adibideko testuinguruarekin, sistemak Beltran agerpenaren adiera *Beltran Pagola* eta *Jaun Mari Beltran* Wikipedia sarreren artean aukeratu behar du.

IV.5.3 Adibidea *Beltran* NEaren agerpena desanbiguatzeko testuingurua.

Izan ere, herri musika erakutsi ez ezik, lan berriak sortzea bultzatu nahi du Beltranek. Jada 70 lagun inguru igaro dira ikerketa lanetarako informazio bila, eta herri musikan aditu den musikariak askotan proposatu izan die ikerketaren norabidea, 'eragile, zirikatzaile izan behar baitugu'.

Beltran Pagola eta *Jaun Mari Beltran* sarrerek bi musikari desberdin deskribatzen dituzte, eta testuinguruak biekin komunean duen hitz bakarra, *musika* hitza da. Argi dago, bi musikarien artean desanbiguatzeko musika hitzak ez duela informazio asko eskaintzen, hori da ordea, metodoek eskuragarri duten informazio guztia, oso erraza izanik sistemaren hutsegitea. Izan ere, horrelako kasuetan asmatzea ia-ia zorizkoa dela esan daiteke.

Bestetik, inolako eskuzko berrikuspenik gabe, guztiz automatikoki sortutako A corpusean (ikus IV.3.1.3 atala), badaude hainbat adibide, adiera egokia beharrean, Wikipedian sarrera duen beste adiera bat esleitu zaizkiena, eta sistemak adiera egokia aukeratzen duenean hutsegite kontsideratzen dena. IV.5.4 adibidea, ebaluazio-corpus horretako adibide bat da.

IV.5.4 Adibidea *Disney* NEaren agerpena desanbiguatzeko testuingurua.

Liburu askotan azaltzen da nolako basakeriak egiten zituzten animaliekin *Disneykoek*. 50eko eta 60ko hamarkadetan gertatzen zen hori. Ordutik hona asko aldatu da animalienganako begirunea eta errespetua.

Kasu konkretu horretan, berrian aurretik *Walt Disney* pertsonaia agertzen denez, *Disney* agerpenari pertsonaren eta ez *The Walt Disney Company* erakundearen adiera esleitzen zaio. Horrela izanda, sistemak adiera egokia esleitu arren hutsegite kontsideratuko da.

Errore asko, beraz, ebaluazio-corpusaren atazarako egokitasun faltan jatorria dutela esan daiteke. Agerikoa da beraz, atazari hobeto egokitzen zaion corpus batekin ebaluatzea ondorio garbiagoak lortzen lagunduko ligukeela.

IV.5.4 Lematizazioan oinarritutako emaitzak

Orain arte aurkeztutako ebaluazio guztiak, Wikipedia zein NEen agerpenen testuinguruko testu tokenizatuekin egindako esperimentuak dira. Euskara ordea, hizkuntza eranskaria izanik, HPko hainbat atazatan sarrerako testuak lematizatzeak batzuetan sistemak hobetzen lagundu izan dutenez, euskarazko NEen desanbiguazio atazarako ere testuen aurreprozesaketa hau burutu eta datu lematizatuekin esperimentuak errepikatzea erabaki da, sistemak

hobetzeko asmoz. Konkretuki sistemak modu hiztunean ebaluatu dira, berdinketak ebazteko portaera onena aurkeztu duen MFS heuristikoa erabiliz. Ebaluazioan lortutako emaitzak IV.8 taulan laburtzen dira.

	MFS	**VSM**	**ESA**	**bESA**
***A*-Tokenizatua**	% 68,32	% 75,53	% 72,43	% 77,66
***A*-Lematizatua**	% 68,32	% 66,04	% 58,58	% 73,96
$B_{Garapen}$**-Tokenizatua**	% 72,00	% 70,48	% 61,09	% 68,00
$B_{Garapen}$**-Lematizatua**	% 72,00	% 67,10	% 55,26	**% 70,86**
B_{Test}**-Tokenizatua**	% 70,40	% 70,00	% 61,60	% 68,40
B_{Test}**-Lematizatua**	% 70,04	% 69,00	% 54,40	**% 68,50**

IV.8 Taula: Corpus tokenizatu eta lematizatuen esperimentuen emaitzak (berdinketak MFSrekin ebatzita)

Orokorrean, lematizatutako esperimentuekin sistemek portaera okerragoa aurkezten dute. A corpuseko emaitzei erreparatuz, galera txikiena duen eta ebaluatutako sistemen artean egonkorren aurkezten den sistema bESA algoritmoan oinarritutakoa da, 4 puntuko galerarekin. Lematizatzean, testu normalizatuagoa eta hiztegi tamaina txikiagoa lortzen bada ere, bestelako erroreak ere sortzen dira. Besteak beste, analisi posible bat baino gehiago duten hitzen lema eskuratzeko *Eustagger*ek egindako analisi desanbiguazio edo aukeraketa okerraren ondorioz, lema okerra eskuratzea esaterako.

IV.5.5 Adibidea *Eustagger*en *Abbadia* osagaiaren lema eskurapen okerra.

Baina	baina LOT_JNT
euskaraz	euskara IZE_ARR
Anton	Anton IZE_IZB
Abbadia	***(abbadi) ADJ_IZO***
hobetsi	hobetsi ADI_SIN
zuen	*edun ADL

IV.5.5 adibidean esaterako, testua *Eustagger*ekin lematizatzean, *Abbadia* hitza *IZE_IZB* bezala lematizatu eta bere lema *Abbadia* eskuratu ordez, *ADJ_IZO* kategoria/azpikategoriarekin lematizatzen du, *abbadi* lema desegokia proposatuz.

B corpuseko bi multzoekin ebaluatzean, berriz, sistemen okertzea ez da hain nabarmena ESAn, eta bESAren kasuan, garapen-corpusean lematizatu gabeko esperimentuetan baino emaitza pixka bat hobeak lortzen dira, ia 3 puntu gainetik, eta test-corpusean mantendu egiten dira. Beraz, luzeraren

efektua minimizatzeko normalizazio-faktoreak ematen du sistema egonkorrago bat lortzeko bide egokia dela, behintzat baliabide txiki eta desorekatuekin lan egiteko garaian.

Lematizazioaren eragina bereziki nabarmena da adiera guztien antzekotasun neurria 0 balioa duten NEen agerpenen desanbiguazioetan, horrelako kasu gutxiago gertatutako baitira. Alegia, berdinketak ebazteko MFS heuristikoa gutxiagotan aplikatuko da sistemak datu lematizatuekin ebaluatzean, tokenen ordez lemak erabiltzean errazagoa izango baita gutxienez osagai bat izatea komunean testuinguruko testuak eta Wikipediako edozein sarrerak. Esaterako, demagun *Maialen Lujanbio*ren *Lujanbio* agerpen anbiguoaren desanbiguaziorako sistemak IV.5.6 adibideko testuingurua jasotzen duela:

IV.5.6 Adibidea *Lujanbio* NEaren agerpenaren testuingurua.

Egoera horren erantzule ez zituen bakarrik gai-jartzaileak jo Lujanbiok; bertsolariek ere badute erantzukizunaren zati handi bat, haren ustez. Bertsolariei dagokienez, haien jarrera nabarmendu zuen ***Lujanbiok*** *bere elkarrizketan, eta aipatu zuen bertsolarien artean badela joera, beraienak ez diren iritzi eta arrazoiak eurenak balira bezala plazaratzeko bertsotan.*

Sistemak *Maialen Lujanbio* eta *Jose Manuel Lujanbio* sarreren artean, adiera egokia zein den ebatzi beharko du. Lematizatu gabeko datuak erabiltzean adibideko agerpenaren testuingurua eta *Maialen Lujanbio* Wikipediako sarrera konparatzean ez da hitz komunik aurkituko. Hortaz, 0 antzekotasun balioa lortuko dute desanbiguaziorako proposatu ditugun metodo guztiek; denak MFS ezik, berdin duela hitz komunik egon edo ez, maiztasun handieneko adiera itzultzen baitu, *Maialen Lujanbio* adiera, hain zuzen ere. Agerpeneko testuingurua *Jose Manuel Lujanbio* sarrerarekin konparatzean, gauza bera gertatuko da. Hitz komunik ez dagoela aurkituko du sistemak. Bi adiera posibleek 0 antzekotasun-neurria izanik, sistemak berdinketa kasuko heuristikoa aplikatuz ebatzi beharko ditu berdinketa. Kasu honetan, MFS heuristikoa aplikatuz *Maialen Lujanbio* adiera itzuliko dute metodo guztiek, agerpena ondo desanbiguatuz.

Lematizatutako esperimentuan, bESAko emaitzetan zehazki, adiera egokia ere lortzen da, baina oraingoan, ez berdinketaren ostean MFS aplikatzen delako baizik eta formak bat ez etorri arren, badaude eta hainbat lema testuinguruaren eta Wikipediako sarrera egokiaren artean komunak direnak. Esaterako *bertsolari* eta *elkarrizketa* bi testuetan agertzen diren hitzak dira, baina desberdin deklinatuta. Hori dela eta, tokenizatutako datuekin lan

egitean ez datoz bat, baina bai lemekin lan egitean. Ondorioz zero baino antzekotasun balio handiagoa lortzen da.

Baina ez da beti horrela gertatzen, alegia, badaude beste adibide asko non lematizazio gabeko esperimentuetan berdinketa kasuak MFSekin ebatziz desanbiguazio egokia lortzen dutenak, eta lematizatzean ordea ez. Kasu honen adibide *Lance Armstrong* pertsonaiaren IV.5.7 adibideko agerpenaren desanbiguazioa burutzean ematen da.

IV.5.7 Adibidea *Armstrong* NEaren agerpenaren testuingurua.

Zazpi segundoko aldea du Gonzalez de Galdeanok sailkapen nagusian ***Armstrongekin****. Probako etapa menditsuena jokatuko da gaur. Saint-Clair pendiza igoko dute bitan, eta helmuga bigarren pasaeran izango da.*

Armstrong izanik desanbiguatu beharreko NEaren agerpena, *Lance Armstrong, Neil Armstrong* eta *Louis Armstrong* Wikipedia sarreren artean aukeratu beharko dute sistemek adiera egokiena. Lematizatu gabeko esperimentuetan testuinguru horretako hitz bat bera ere ez duenez bat egingo Wikipediako sarrerekin, zeroko antzekotasuna izango dute adiera guztiek eta ondorioz, MFS heuristikoarekin ebaztean, adiera egokia eskuratuko da, *Lance Armstrong*, alegia. Lematizazioko bertsioan aldiz, aditz bakan batzuk bat egiten dute testuinguruaren eta Wikipediako sarreren artean, bESAren kasuan, *Neil Armstrong* desanbiguazio adiera gisa itzultzea egiten dutenak, sistemaren hutsegitea eraginez.

Antzeko zerbait gertatzen da IV.5.8 adibidean, non lematizazioak *sail, egin* eta *ezberdin* bezalako hitz orokorren lemak erabiltzean Wikipedia sarreren eta testuinguruko testuaren arteko hitz gehiago egiten duten bat, eta agian desanbiguaziorako oso erabilgarriak izan ez arren, testu laburrekin ebaluatu denez, nahikoak izan daitezke desanbiguazio okerrera eramateko sistemak.

IV.5.8 Adibidea *Zinemaldia* NEaren agerpenaren testuingurua.

Gaurko inaugurazio ekitaldiaren osagaiei dagokionez, ezer gutxi ezagutzen da. Jakin denez, ***Zinemaldiaren*** *bost hamarkadetako ibilbidea gogoratuko da beste horrenbeste pertsonaiaren bidez, eta ohi bezala, sail ezberdinen eta atzera begirako zikloen gaineko errepaso azkarra egingo da.*

IV.5.8 adibide zehatz horretan, hitzekin *Donostiako Zinemaldia* adiera egokia proposatzen bada ere, lematizazioko datuekin bESAk *Canneseko Zinemaldia* ebazten du adiera egoki bezala, hutsegitea eraginez.

Esperimentu horiek lematizazioak laguntzen ez duela aurkitu duten lehen kasua ez badira ere, Arregi eta Fernández-ek (2002) testu sailkapenerako algoritmo guztiekin irabaziak lortzen ez zirela erakutsi baitzen, ezin da esan orokorrean NED atazarako lematizazioa lagungarri ez dela, ebaluazioetarako erabilitako corpusak ez baitira oso handiak, eta beraz, datu multzo handiagoekin berretsi beharreko emaitzak dira. Atazarako egokiagoak diren *stop-word* zerrendak definitu eta erabiltzen saiatzea ere emaitzak hobetzeko beste bide bat ere izan daiteke. Izan ere, azken adibidean ikusi dugun bezala, hitz orokorregien agerpenak desanbiguazio prozesuan lagundu baino hutsegiteak eragin baititzakete.

IV.6 Ondorioak

Euskaraz idatzitako erakunde-, pertsona- eta toki-izen anbiguoen agerpenak automatikoki desanbiguatzeko metodo ez-gainbegiratu desberdinenekin egindako esperimentuak aurkeztu ditugu kapitulu honetan. Desanbiguazio problema, NEaren agerpen bat ezagutza-base bateko sarrera batekin lotzeko problema bezala ulertzen da. Lan honetan, ezagutza-basea euskarazko Wikipedia izan da.

NEen desanbiguazioa ebazteko proposatu dugun estrategiak, lehenbizi ezagutza-basean dauden sarreren artean, NEaren agerpenaren adiera kontsidera daitezkeen hautagaiak identifikatzen ditu, eta ondoren, hautagaien artean egokiena zein den ebazten du. Euskarazko NED sistema, beraz, bi urrats nagusirekin diseinatu dugu:

1. Bilaketa fasea: NEaren agerpenaren adiera-hautagaien identifikazioa.
2. Desanbiguazioa fasea: adiera-hautagaien artean, adiera egokiena aukeratu.

Bilaketa urratsean, sistemak hiztegi antzeko bat erabiltzen du NEaren agerpenaren adiera posibleak identifikatzeko. Hiztegian, beraz, agerpen batentzat ezagutza-baseko sarrera sorta bat definituta egongo da. Hiztegi atzipenean, bi kasu posible daude: sistemak agerpenarentzat ezagutza-baseko sarrera bat edo gehiago eskuratzea; edo agerpena hiztegian definituta ez dagoelako adiera-hautagairik ez eskuratzea. Lehenbiziko kasuan, sistema desanbiguazio fasera pasako litzateke. Bigarren kasuan, aldiz, sistemak ezingo du adierarik proposatu eta ondorioz, NIL erantzuna emango du.

Desanbiguazio urratsean, sistemak lehenengo urratsetik eskuratutako hautagaiak pisatuko ditu, eta pisu altueneko hautagaia itzuliko du NEaren agerpenaren adiera gisa. Hautagai aukeraketa estrategia horretan ere, bi egoera eman daitezke: hautagaiak pisatzean berdinketa ematen da; edo hautagai guztietatik batek pisu altuena lortzen du. Azken kasuan, zalantzarik ez dago, eta sistemak hautagai hori itzuliko du adiera gisa. Lehenbiziko kasuan berriz, sistemak zerbait gehiago egin beharko du. Eta zerbait gehiago horretarako, sistema bi izaera desberdinekin ebaluatu dugu: isila eta hiztuna.

Izaera isilean, sistemak berdinketa kasuetan ez du ezer erabakitzen eta NIL erantzuten du. Izaera hiztunean, aldiz, sistemak berdinketa ebazteko heuristiko bat aplikatzen du: zorizko aukeraketa edo maiztasun handieneko adiera heuristikoa. Bi heuristikoak ebaluatu ostean, espero bezala, maiztasun handieneko adiera heuristikoak, zorizkoak baino emaitza hobeak lortu ditu.

Euskarazko NEak tesi-lan honen helburu nagusia izanik, esan bezala, euskarazko NEDerako ezagutza-base gisa Wikipediako euskarazko bertsioa erabili dugu. Euskarazko Wikipedia ingelesezko bertsioarekin konparatuz gero, bietan mota berdineko testuak eta sarrerak aurkitu arren, ezaugarri desberdinetako bertsioak direla ezin da ahaztu. Izan ere, euskarazko Wikipediaren tamaina nabarmenki txikiagoa izateaz gain, alegia, sarrera gutxiago izateaz gain, artikuluen testuak ere motzagoak baitira. NEen desanbiguazioan eragin zuzena duten ezaugarriak dira horiek: artikulu kopuruak agerpenen adiera kopurua mugatzen baitu, sarrera gutxiago izanda, agerpen gehiagoren adiera egokia ezagutza-basean ez egotearen arazoa ekarriz; eta bestetik, sarreren deskribapena izanik NEaren agerpenaren testuinguruarekin batera desanbiguazio prozesuko informazio-iturri nagusia, horiek laburrak izateak desanbiguaziorako datu gutxi izatea dakar, eta ondorioz problemaren zailtasuna handitzen dela uler daiteke.

Desanbiguazio faseko metodoei dagokionez, euskararako eskuragarri izan ditugun baliabide kopuru eta tamaina txikiak direla eta, NEen desanbiguaziorako portaera onena aurkezten duten metodoen artean, metodo ez-gainbegiratuak erabili ditugu. Metodo mota horiekin ingeleserako emaitza interesgarriak lortu izanaz gain, gainbegiratuak baino baliabide gutxiagorekin euskararako NEen desanbiguazio problema ebazteko bide bat eskaintzen dutelako aukeratu ditugu.

Metodo ez-gainbegiratuen artean, ingelesarekin portaera ona agertu duten izaera desberdineko metodoak erabili ditugu. Konkretuki, maiztasun handieneko adiera, bektore-espazioen eredua, ESA eta grafoetan oinarritutako UKB algoritmoak erabili ditugu.

Maiztasun handieneko adiera (MFS), ebaluatutako metodo sinpleena izateaz gain, hitz adiera-desanbiguazioan egin ohi den bezala, abiapuntuko emaitzak lortzeko erabili dugu.

Problemaren helburua NEen agerpen bat ezagutza-base bateko sarrera batekin lotzea izanik, ataza IR sistemen ikuspuntutik plantea daiteke: NEaren agerpena galdera izango da eta ezagutza-baseko sarrera bilatu beharreko erantzuna. Ikuspuntu horretatik, IR sistemetan oso erabilia den bektore-espazioen ereduaren portaera NED atazan ebaluatu dugu. Eta emaitzarik onenak lortu duen metodoa izan ez arren, hainbat esperimentutan MFSren abiapuntuko emaitzak gainditzea lortu duen bakarra izan da.

ESA, bestetik, semantika eta desanbiguazio alorrean oso emaitza onak lortu dituen algoritmoa izanik, euskarazko NED ataza ebazteko metodo interesgarria kontsideratu da hasieratik. Esperimentuak egitean ordea, Wikipedia txikiekin portaera asko okertzen dela ikusita, tamainaren eragina minimizatzeko ESArentzako normalizazio-faktore bat definitzea pentsatu da. Testuinguruko eta Wikipedia sarreren antzekotasuna neurtzean, bi testuek komunean duten hitz kopurua kontuan hartzen duen normalizazio-faktorea definitu da, hain zuzen ere. Normalizazio-faktore hori erabiltzen duen ESA orekatuarekin (bESA) ESAren emaitzak nabarmenki hobetzea lortu dugu.

Azkenik, grafoetan oinarritutako UKB algoritmoak, Wikipediako esteken egitura baliatuz desanbiguazioa ebazten duen metodoa erabili dugu. Gainerako metodoek ez bezala, UKBk egindako esperimentuetan, NED ataza euskararako ingeleserako baino hobeto ebazten duela esan dezakegu. Hobekuntza horren arrazoia, euskarazko Wikipedian, esteka gutxiago definituta egon arren semantikoki aberatsagoak izatean izan dezake bere jatorria. Hala ere, portaera hau, beste Wikipedia txikiekin ebaluatuz berretsi beharreko ondorioa da.

Ebaluaziorako, izaera desberdineko bi corpus erabili ditugu.

- A corpusa, non NEen agerpen guztiek adiera egokia euskarazko Wikipedian sarrera duten.
- B corpusa, non badauden NEen agerpen batzuk adiera egokia esperimentuetarako erabilitako Wikipedia bertsioan sarrera definituta ez duten.

Lehenbiziko corpusa, handiagoa bada ere, agertoki errealago baten ebaluazioa eskaintzen digu bigarrenak, eta aldi berean zailagoa ere. Zailtasun

horren islapena dira lortutako emaitzak, sistema guztietan betetzen baita A corpusean B corpusean baino emaitza hobeak lortzen direla.

B corpusean ematen diren errore askoren jatorria NIL kasuen ebazpen okerrean dute jatorria. Espero zitekeen zerbait da, azken batean sistema ez baita diseinatu NIL kasuek behar duten tratamendu berezia jasotzeko. Izan ere, sistemak NIL erantzuna soilik ematen baitu hautagairik aurkitzen ez duenean eta berdinketa kasuetan izaera isilean. Baina hautagai bat besteak baino pisu altuagoarekin eskuratzen duenean, ez du planteatzen benetan adiera egokia den. Eta sisteman planteamendu horren faltak, B corpuseko hainbat kasutan hutsegitea eragiten du. Beraz, etorkizunean sistema hobetzeko bidean eman beharreko lehen hobekuntzetako bat, NIL kasuen tratamendua aztertzea da. Edozein kasutan, metodo guztien artean, ebaluazio-corpus desberdinetan emaitzarik onenak eta portaera egonkorrena aurkeztu duen algoritmoa UKB izan da.

Azpimarratzekoa da ere, hobekuntza interesgarria eskaintzen duela ESAren inguruan definitutako normalizazio-faktoreak, egindako esperimentu guztietan ESA orekatuak ESAk baino emaitza hobeak lortu baititu. Etorkizunean euskararen tamaina antzekoa duen beste Wikipediekin ebaluatzea interesgarria izan daiteke, normalizazio-faktorearen sendotasuna neurtu ahal izateko.

Testuen ikuspuntutik, metodoak bi modutan ebaluatu dira: testu tokenizatuak eta testu lematizatuak erabiliz, hitzak eta lemak erabiliz, alegia. Lematizazioa askotan lagungarria gertatzen den aurreprozesaketa izan arren, egindako esperimentuetan behinik behin, NED ataza ebazteko lagungarria ez dela esan dezakegu, egindako esperimentu guztietan hitzak beharrean, lemak erabiltzean sistemaren portaerak okerrera jotzen baitu. Ebaluazio hori, UKBrekin ezik metodo guztiekin egin dira, besteetan lortutako emaitza eskasak eta UKBrentzat datuak prestatzeak suposatzen duen esfortzua dela eta, esperimentu hori ez egitea erabaki baita.

Hala ere, lematizazioak NED atazan lagungarri gertatzen ez dela orokortu ezin den ondorioa da, erabili ditugun baliabideak eta ebaluazio-corpusen tamainak txikiak izan baitira. Emaitzak berresteko esperimentuak datu multzo handiagoekin ebaluatu beharko dira etorkizunean.

V. KAPITULUA

Ondorioak eta etorkizuneko lanak

Kapitulu honetan, batetik tesi-lan honekin egin ditugun ekarpenak laburbiltzen dira eta, bestetik, atera ditugun ondorioak azaltzen ditugu. Bukatzeko, lan honi lotuta, etorkizunean egin daitezkeenak aztertuko ditugu.

V.1 Ekarpenak

Tesi-lan honetan euskarazko entitate-izenen inguruan egindako ikerketa azaldu dugu. Zehazki, I. kapituluan helburu gisa proposatutako euskarazko entitate-izenen identifikazio eta sailkapena (NERC), itzulpena (NET) eta desanbiguazio-atazak (NED) jorratu eta ebatzi dira. Guztietan, eta helburuak planteatu zirenean proposatu zen metodologiari jarraiki, eskuragarri izan ditugun baliabideak berrerabiliz eta metodo sinpleen erabileran eta horien konbinazioetan oinarritutako ebazpenak landu ditugu, horietan euskarazko ezaugarri morfologikoek duten eragina ere aztertuz.

Entitate-izenen tratamendua hedapen handiko lana izanik, alor guztietan ondo sakontzea ezinezko izan da. Hala, tesi-lan honetan landutako euskarazko NEen identifikazio-, sailkapen-, itzulpen- eta desanbiguazio-atazak baliabide urriko hizkuntzetarako agertoki errealetan kokatu ditugu. Eszenatoki horien barruko problematika aztertu eta ahal izan denean ebazpena eman diogu. Lan honetako ekarpenik garrantzitsuenak ebazpen horietan garatutako tresnak izan dira: euskarazko NEak identifikatzeko eta sailkatzeko tresna; NEen itzulpenak burutzeko hizkuntzen menpeko tresna eta hizkuntzekiko

erdi-independentea den tresna; eta azkenik, desanbiguazio-atazan euskarazko NEak Wikipedia sarrerekin lotzeko tresna.

Tresna horien guztien garapenak baliabide urriekin burutu dira, guztietan berrerabilpena oso kontuan izan delarik. Berrerabilpena baliabideen ikuspuntutik zein diseinuen hedagarritasunaren ikuspuntutik kontsideratu da, batetik, eskuragarri zeuden baliabide urriak ahalik eta hoberen ustiatu eta erabili ahal izateko, eta, bestetik, euskararako diseinatutako estrategiak baliabide urriko bestelako hizkuntzetarako ahal den heinean berrerabilgarri izan daitezen.

Baliabideei dagokienez ondo ezagunak diren *WordNet* eta Wikipedia bezalako baliabideak ustiatu ditugu, euskarazko NEen atazetan ahalik eta etekin handiena atera ahal izateko. Wikipediaren kasuan, baliabide horren erabileraren aitzindariak izan gara. Gaur egun oso zabaldua dagoen erabilera bada ere, hizkuntzaren prozesamenduan NEen itzulpenarekin baliabide hori ustiatzen lehenetarikoak izan ginen.

Estrategiei dagokienez, tresna horien diseinu eta garapena hizkuntzen prozesamenduko bi hurbilpen nagusiekin burutu ditugu: hizkuntza-ezagutza eta ikasketa automatikoarekin, hain zuzen ere. Bi hurbilpenak erabili izanak, ataza desberdinetan bi hurbilpenen portaerak konparatzea ahalbidetu digu. Posible izan denean ere, bi hurbilpenak konbinatu ditugu eta konbinazio horien portaerak bakarkakoekin konparatu ditugu.

Bestetik, ikasketa automatikoko hurbilpenetan bereziki, euskara hizkuntza eranskaria izanik, ikasketa automatikoan erabili ohi diren atributuez gain, informazio interesgarria eskaini dezaketen atributuen azterketa eta hautapenari arreta berezia eskaini diegu. Azterketa eta hautapen horiek bestelako hizkuntza eranskarietarako erabilgarri gerta daitezkeela uste dugu.

Oro har, tesi-lan honetan garatutako NEen tratamendurako tresnak HPko hainbat proiektu eta aplikazioetan erabilgarriak suertatu dira: *IXA taldea*n bertan zein *IXA talde*tik at.

Esaterako, euskarazko NEen identifikazio eta sailkapenerako tresna *IXA taldea*n corpus sorkuntza helburu duten *HERMES* (Verdejo *et al.*, 2003) eta *EPEC* (Aduriz *et al.*, 2006b) proiektuetan euskarazko etiketatze-prozesua automatizatzeko erabili da. Tresna bera taldean galdera-erantzun sisteman egindako hurbilpenetan (Alegria *et al.*, 2009) erabili izan da ere.

*IXA talde*tik at, itzulpen-automatikorako *OpenTrad*[1] proiektuaren barruan, itzulpen-memoriak publikatzeko egindako NEen garbiketa eta orokor-

[1] `http://www.opentrad.org/`

tze lanetan erabilgarriak izan dira tesi-lan honetako ekarpenak. UPV/EHUko GALAN taldekoek, berriz, ikas-materialetako domeinu identifikaziorako egindako HPko hurbilpenean erabili izan dute (Larrañaga *et al.*, 2003).

Tresnak tesi-lan honen ekarpen garrantzitsuenak izan badira ere, badira NEen inguruan landutako hiru ikerlerro nagusietatik aipagarriak diren bestelako ekarpenak ere. Horiek dira, hain zuzen ere, eginkizun bakoitzeko ondoko ataletan deskribatzen direnak.

V.1.1 Euskarazko entitate-izenen identifikazioa eta sailkapena

Euskarazko testuetan entitate-izenen multzoa osatzen duten pertsona-, toki- eta erakunde-izenak identifikatzen eta sailkatzen dituen NERC sistema erabilgarria sortu dugu (% 71,35eko F_1 emaitzekin). Sistema martxan da eta, arestian esan bezala, hainbat proiektu eta produktutan erabilgarria suertatu da.

NERC sistemaren garapenean, sistemen konbinazioak banakako sistemen emaitzak hobetzen lagun dezaketela berretsi dugu. Hizkuntza-ezagutzan oinarritutako hurbilpenak, giza baliabideez gain, baliabide gutxirekin NERC sistema sinple bat garatzeko aukera eskaintzen duela erakutsi dugu. Gainera, sistema horri esker, IArako baliabideak sortu eta instantzien errepresentazio ona egiteko atributu aukeraketan oso lagungarri suerta daitekeela frogatu dugu.

IAko hurbilpenekin egindako esperimentuek hizkuntza-ezagutzan oinarritutako NERC sistemaren emaitzak gainditzen badituzte ere, izaera desberdinetako hurbilpenak konbinatzeak sistema sendoagoa sortzeko bidea dela ikusi ahal izan dugu. Izan ere, hizkuntza-ezagutzan oinarritutako tresna eta IAko hurbilpenak konbinatzean, soilik IAko hurbilpen desberdinak konbinatzean baino emaitza hobeak lortzen dira; hizkuntza-ezagutzan oinarritutako sistemaren emaitzak baxuagoak izanda ere, konbinazioetarako osagarriak izan daitezkeela erakutsi da horrenbestez.

NERC atazaren ebazpenerako informazio semantikoaren ekarpena aztertu nahian, *EusWordNet* baliabidea ustiatu dugu. Zehazki NEen sailkapen-atazan aztertu dugu *EusWordNet*en eragina eta, oro har, baliabide horren hobekuntza emaitzetan mugatua dela ikusi ahal izan dugu. Kasurik onenean *EusWN*ek sistemaren F_1a % 1ean hobetu du.

V.1.2 Euskarazko entitate-izenen itzulpena

Entitate-izenen itzulpen-ataza ebazteko lanean, errebisio bibliografiko sakona egin dugu. Errebisio horretan, egile desberdinek egindako hurbilpenak aztertuz teknika eta baliabide erabilienak identifikatu ahal izan ditugu. Teknika eta baliabide horiek euskarazko NEen itzulpen sistema sortzeko nola ustiatu ere aztertu dugu. Hala, NEen itzulpen-sistema baten modulu nagusiak identifikatu ahal izan ditugu:

- NEaren osagaiz osagaiko itzulpen modulua: transliterazio-erregelak eta hiztegi elebidunak konbinatuz osagaiz osagaiko itzulpenaz arduratzen den modulua.
- NE osoaren itzulpen-hautagaiak eratzeko osagaien ordenaketa modulua osagaien itzulpen-proposamenetatik abiatuta: euskara-gaztelera, euskara-ingelesa eta gaztelera-ingelesa bikoteetako hizkuntzek jarraitzen duten egitura sintaktiko desberdina dela eta, abiapuntuko osagaien itzulpenak xedeko forman posizio egokian kokatzeaz arduratzen den modulua.
- NEen itzulpen-hautagaien aukeraketa: behin hautagaiak sortuta, guztien artean abiapuntuko NEerako proposamen egokiena aukeratzeaz arduratzen den modulua.

Transliterazioa NEen itzulpen hurbilpenen artean metodo zabalduena da, eta corpus moten artean corpus konparagarriak izanik eskuragarrienak, bi baliabide horiek konbinatuz euskarazko NEen itzulpen tresna garatzeko hurbilpen desberdinak egin ditugu: hizkuntza-ezagutzan oinarritutako hurbilpena batetik, non itzulpena ebazteko hizkuntza-bikotearekin guztiz menpekoa den sistema garatu den; eta bestetik, hizkuntza-bikotearekin menpekotasun hori ekiditeko hurbilpena. Bi hurbilpenetan eskuragarri izan ditugun corpus eta hiztegiak berrerabili ditugu, eta bien arteko desberdintasun nagusia aipatutako hiru moduluen artean lehenbiziko bien inplementazioa da. Hizkuntza-ezagutzan oinarritutako hurbilpenean hizkuntzen ezaugarri linguistikoetan oinarritutako erregelak erabiltzen dira. Bigarren hurbilpenean berriz, teknika orokorragoekin inplementatzen dira erregela horiek. Transliteraziorako edizio-distantzian oinarritutako erregelak erabili ditugu, eta ordenaziorako abiapuntuko NEaren osagaiak xede-hizkuntzako forman edozein posiziotan ager daitekeela kontsideratzen dituen gramatika erabili dugu, hain zuzen ere.

Bi hurbilpenekin egindako esperimentuek erakusten dute hizkuntzen ezagutza linguistikoa izan gabe NEak modu fidagarrian itzultzeko sistemak garatzeko hizkuntzekiko erdi-independentea den hurbilpena egokia dela. Hizkuntzekiko erdi-independente izateari esker, hurbilpen hori ere beste hizkuntza-bikoteetarako hedagarri dela ikusi dugu. Bikote berri baterako sistemaren garapena egiteko nahikoak dira corpus konparagarri bat eta hiztegi elebidun bat. Are gehiago, linguistikoki oso desberdinak diren bi hizkuntzen arteko NEen itzulpena egiteko (euskara-ingelesa) bi urratsetako itzulpena egin beharrean (euskara-gaztelera-ingelesa), alegia, tarteko hizkuntza erabiltzea baino metodologia erdi-independentearekin hizkuntza-bikote horretarako sistema sortzea aukera hobea dela erakutsi dugu.

Hizkuntzekiko erdi-independenteak diren hurbilpenetan, corpus konparagarrien egokitasuna bereziki garrantzitsua dela ikusi dugu. Izan ere, sistemaren funtsa da baliabide horiek eta horren egokitasunak emaitzetan eragin zuzena duela. Corpus konparagarrien ustiapen horretan, Wikipediaren ezaugarri estrukturalak eta corpus konparagarri gisa erabiltzen ere lehenetarikoak izan ginela esan dezakegu, esperimentuak egin genituen garaian, oso lan gutxi ezagutzen baitziren Wikipedia ustiatzen zutenak antzeko atazak ebazteko.

Horrez gain, Wikipedia ustiatuz sortutako euskara-ingelesa NEen itzulpen sistemak, espresio horiek euskaratik ingelesera itzultzeko gai izateaz gain, Wikipediako hizkuntzen arteko sarreren estekak (WILak) aberasteko erabilgarria izan daitekeela erakutsi dugu.

Hurbilpen desberdinak ebaluatzeko corpus desberdinak erabili badira ere, ataza honen ebaluazio-corpusen eraketa automatizatzeko ekarpena ere egin dugu; berriz ere, Wikipedia eta bere ezaugarri estrukturalak ustiatuz. Zehazki, edozein testu multzoko NEetatik abiatuz horiek Wikipedian bilatuz eta WILak ustiatuz ebaluazio-corpusa automatikoki eratzeko metodologia aurkeztu dugu.

NET sistemak hobetzeko bidean hiztegi elebidunen aberasketa automatikorako metodologia baten ekarpena ere egin dugu ataza honetan. Aberasketa prozesu automatiko horrek, NEen osagai diren hitz arruntek NE espresioetan agertzen direnean ohikoak ez diren itzulpen aldaerekin hiztegiak aberastea du helburu. Funtsean metodologiak, ohi bezala itzultzen ez diren osagaiak barne dituzten NE bikoteetatik abiatzen da. Hizkuntza-bikote horien hiztegi elebidun arrunt bat izanik, abiapuntu- eta xede-hizkuntzako NEaren osagaiak hiztegia erabiliz parekatzen saiatzen da, eta parekatzea lortzen ez dituen osagaien baliokidetza bilatzen saiatzen da.

Azkenik esan behar da lan honekin euskara-gaztelera, gaztelera-ingelesa eta euskara-ingelesa hizkuntza-bikoteetarako NEak itzultzeko sistemak martxan direla.

V.1.3 Euskarazko entitate-izenen desanbiguazioa

Entitate-izenak desanbiguatzeko ataza berri samarra den heinean, aurreko lan batzuk aztertu baditugu ere, bereziki azken bi urteetan TAC-KBP ataza-partekatuetan aurkeztutako lanen errepaso bibliografiko sakona egin dugu. Ataza horietan bezala, tesi-lan honetan ere NEen desanbiguazioa NEaren agerpen bat ezagutza-base bateko sarrera batekin lotzeko problema bezala planteatu dugu. Desberdintasun nagusia ezagutza-basea da, ingeleseko Wikipedia izan beharrean euskarazko Wikipedia izan da.

Euskarazko Wikipediaren tamaina ingelesekoarekin konparatzean, askoz txikiagoa dela ikustea erraza da. Eta NEen desanbiguazio-ebazpenean tamaina alde horrek duen eraginaren azterketa izan da, hain zuzen ere, tesi-lan honetako ekarpen aipagarrietako bat. Izan ere, lan bibliografikoan identifikatutako teknika hedatuenak eta NEen desanbiguazio-ataza hobekien ebazten dituzten estrategiak baliabide urriko hizkuntza batekin aztertu ditugu.

Azterketa horretan ezagutza-basearen garrantzia berretsi dugu, eta ingeleserako sistemarik onenak emaitzak euskararako mantentzen ez direla ikusi ahal izan dugu. Are gehiago, estrategia horiek baliabide urriko hizkuntzetan portaera hobetzeko ikerketa egin dugu. Bide horretan, azpimarragarria da euskararen emaitzak hobetzeko NED atazan ezaguna den ESA (*Explicit Semantic Analysis*) algoritmoari egindako egokitzapena (txostenean zehar bESA gisa aurkeztua). Egokitzapen horri esker euskararako egindako esperimentu guztietan jatorrizko ESA F_1 neurrian 5 eta 7 puntu artean hobetzea lortu dugu.

Hala ere, eta hobekuntza nabarmena bada ere, ez da hori lortutako sistemarik onena. Ingeleserako oso portaera ona erakutsi ez duen arren, UKB algoritmoan oinarritutako sistema izan da euskararen kasuan izaera egonkorrena erakutsi duen algoritmoa. Ebaluazio-corpusaren tamaina txikiak direla eta, ondorio definitiborik atera ezin bada ere, grafoetan oinarritutako algoritmo horrek besteen aurrean Wikipediaren ezaugarri estrukturalak hobeto ustiatzen dituela ematen du. Eta ustiapen estruktural horrek eskaintzen dion informazio semantikoari esker desanbiguazioa hobeto ebazteko gai dela ematen du.

Lematizazioarekin egindako esperimentuei begira, datuen normalizazio-

estrategia horrek kasu honetan lagungarria ez dela ikusi dugu. Normalizazioa egitean galtzen den informazioa desanbiguaziorako kaltegarria dela ematen du. Dena den, erabilitako ebaluazio-corpusen tamaina txikiak, berriz ere, ez du aukera ematen ondorio garbirik ateratzeko.

Aipatzekoa da artikulu laburrek arazo handiak ematen dituztela. Horregatik, ingeleserako egindako hainbat NED lanetan sarrera laburrak ezabatu izan dira. Euskarazko baliabide honetan estrategia hori aplikatuz gero, esperimentuak burutzeko corpus txikiegia geratuko litzateke eta emaitzak ez lirateke esanguratsuak izango. Hortaz, desorekatua izan arren, luzera desberdinetako sarrerekin lan egin behar izan dugu. Tamaina txiki eta oso desberdinekin lan egite horrek sortzen dituen efektu negatiboekin behin baino gehiagotan topo egin dugu. Lematizazioko esperimentuetan esaterako, artikuluen tamaina txikiak eta desberdinak izateak ez du ebaluazioan lagundu. ESAren kasuan ere, algoritmoaren egokitzapenaren beharraren arrazoietako bat tamaina horretan du jatorria.

Ebaluazio-corpusak txikiak badira ere, horiek sortzeko metodologia automatiko bat ere definitu dugu, baliabide urriko bestelako hizkuntzetarako lagungarria suerta daitekeena, hain zuzen ere. Funtsean, metodologia horrek NEak testuetan aipatzean testu bereko aipamenetan lehenbizikoan forma osoa eta ondorengoetan forma laburrak erabiltzen direneko hipotesian oinarritzen da. Horrela, forma laburrak NEen agerpen anbiguo kontsideratuz, eta forma desanbiguatua testu berean agertzen den forma luzea dela suposatuz, ebaluazio-corpusa automatikoki sor daiteke: forma labur bakoitzaren paragrafoa agerpenaren testuingurua izango da, eta desanbiguatzeko NEa agerpen laburra.

V.2 Ondorioak

Ekarpen nagusienak aurkeztuta, tesi-lan honekin lortutako ondorio nagusienak azalduko ditugu ondoko lerroetan.

Hizkuntza-ezagutza eta ikasketa automatikoan oinarritutako estrategiak, noiz zein aplikatu ez da erabaki erraza[2]. Atazaren inguruko ezagutza, horretarako eskuragarri dauden baliabideak, corpus tamainak eta horien egokitasunak erabaki horretan kontuan hartzeko ezaugarriak dira.

[2]Dan Yurafsky irakasleak 2011 urtean UPV/EHUn eman zuen hitzaldian agertu zuen bezala. `http://www.unibertsitatea.net/blogak/ixa/flirteatzen-ari-da`

Tesi-lan honetan, bi hurbilpen horiek entitate-izenen inguruko ataza desberdinetan aplikatu ditugu, eta atazaren arabera, eta noski baliabideen arabera, emaitzak aldatzen badira ere, hurbilpen desberdinetako sistemak konbinatzea sistemaren portaera hobetzeko bide ona dela berretsi dugu. Zenbat eta izaera desberdinagoa izan, orduan eta hobekuntza nabariagoa lortzen da, sistemen arteko osagarritasunari esker emaitzak hobetzea lortzen baita.

Edozein dela hurbilpena, atazak ebazteko teknika gainbegiratu, ez-gainbegiratu zein erdi-gainbegiratuak aplikatu ditugu. Ataza bera ebazteko bi hurbilpenak aplikatzeak bien portaera aztertzeko eta konparatzeko aukera eskaini digu. Azterketa hori bereziki NEen itzulpen-atazan egin dugu, eta bertan baliabide eta esfortzu txikiagoekin estrategia erdi-gainbegiratuek metodo gainbegiratuetan oinarritutako sistemen emaitzetatik gertu geratzen direla ondorioztatu ahal izan dugu.

Ahalegin eta baliabide gutxiagorekin sistema gainbegiratu baten pareko sistema garatzeko bidea izateaz gain, estrategia erdi-gainbegiratuek, metodo orokorrekin garatuz gero, hizkuntzarekin ia menpekotasunik ez duten eta beste hizkuntza batzuetara hedagarriak diren sistemak garatzeko bidea ireki dugu. NEen itzulpen-atazan esaterako, corpus konparagarriak eta Wikipedia bezalako baliabideak edizio-distantzian oinarritutako metodologiari helduz.

NEen desanbiguazioan ingeleserako, eta oro har baliabide aberats asko dituzten hizkuntzetarako, portaera ona duten estrategiak baliabide urriko hizkuntzetan aplikatzean lehenbizikoetatik lortutako ondorioak beti konfirmatzen ez direla ikusi dugu. Izan ere, tamaina eta ezaugarri desberdinetako Wikipedia erabiltzean NEen desanbiguazio emaitzak ez dira soilik okertzen, baizik eta metodoen portaerak zeharo aldatzen dira. Ingeleserako sistemarik onenak bezala kokatzen direnak euskararako okerrenetarikoak izatera pasatzen dira, ESAren kasuan bezala. Eta alderantziz ere, ingeleserako ez hain portaera ona aurkezten duten sistemek, euskararako oso emaitza onargarriak lortzen dituzte UKBren kasua bezala. Hala ere, erabilitako ebaluazio-corpus tamaina txikiak direla eta, ondorio hauek berresteko corpus handiagoekin esperimentuak errepikatzea ezinbestekoa kontsideratzen dugu.

Euskarazko HPko atazak ebazteko garaian lematizazioak ondoko galdera hau sortzen digu: lematizazioak eskaintzen duen informazioaren normalizaziorekin irabazten dena ez ote den prozesuan zehar sortzen diren errore eta zehaztasun galerekin galtzen. Izan ere, Arregi eta Fernández ek (2002) testuen sailkapen automatikorako aurkezten duten fenomeno horrekin NEen desanbiguazioa ebaztean topo egin dugu, lematizazioak anbiguotasuna ebazten lagundu baino sistemaren portaera okertzen duela ikusi baitugu.

Edozein kasutan, eta lehen esan bezala, azken ondorio horiek berretsi ahal izateko, esperimentuak ebaluazio-corpus handiagoekin errepikatzea beharrezkoa da.

V.3 Etorkizuneko lanak

HPrako sortutako tresnak ez dira inoiz perfektuak, eta, beraz, beti egin daiteke saiakeraren bat tresna horiek hobetzeko. Ez hainbeste hizkuntza-ezagutzan oinarritutako hurbilpenetan, bai ordea ikasketa automatikoan oinarritutako tresnetan. Algoritmo konplexuagoak erabiliz, atributu aukeraketa fintzen, eta oro har, ikasketa- zein ebaluazio-corpusak handituz Arrietak (2010) bere tesi-lanean aurkezten duen bezala, datu tamainarekin mota horretako sistemak hobetu ohi baitira. Beraz, tesi-lan honetan NEen atazak ebazteko proposatutako tresnen hobekuntzan, bereziki itzulpen- eta desanbiguazio-atazetan proposatutako estrategia erdi-gainbegiratuak hobetzeko bidea ireki da.

Wikipediaren eta bere ezaugarrien inguruan ikasitakoak NEen identifikazio- eta sailkapen-atazan aplikatzea interesgarria kontsideratzen dugu. Izan ere, ataza horretan baliabide hori ustiatzea NERC sistemetan eta bereziki sailkapen urratsean hobekuntzak ekarriko dituela uste dugu.

Sistemak hobetzeko estrategietan pentsatzean ezinbestekoa da aurreikustea NEen desanbiguazioan lortutako emaitzetan kolokan geratu diren zalantzak eta ondorioak argitzeko asmoz burututako esperimentuak datu sendoagoekin errepikatu behar direla. Gaur egun euskarazko Wikipedia handitzeko komunitatea egiten ari den esfortzuari esker, epe motzean esperimentu horiek Wikipedia handiago eta osoago batekin egitea bideragarria izango da. Hori, ordea, ez da nahikoa izango NED sistema sendo bat sortzeko. Izan ere, tesi-lan honetako NED sistemetan NIL kasuei (ezagutza-basean sarrera egokia definituta ez duten adibide anbiguoei) ez baitzaie behar duten tratamendua eman. Beraz, NIL kasu horien ebazpenerako estrategiaren diseinu eta inplementazioa eta horren integrazioa sisteman beharrezkoak izango dira euskarazko NED sistema egonkor eta sendo bat erabilgarria izateko. Azkenik, dimentsio antzeko beste hizkuntzetako Wikipediarekin NED esperimentuak errepikatzea ere interesgarria kontsideratzen dugu, horrek tesi-lan honetako ondorioak orokortzeko aukera emango bailiguke.

Tesi-lan honetan entitate-izenen identifikazioa/sailkapena, itzulpena eta

desanbiguazioa, hiru atazak nolabait independenteki ebatzi badira ere, batak besteentzat lagungarriak izan daitezke. Hala, itzulpenak eta desanbiguazioak, esaterako, NERC sistema baten emaitzak fintzeko, edo lantzeke geratu diren NE sendoen (NE labur bat habiatuta duen NEen) tratamendurako lagungarriak izan daitezke. Edota, NEen desanbiguaziorako NERC zein NET atazen emaitzak nola lagundu dezaketen aztertu beharko litzateke. Lagungarritasun eta osagarritasun hori tesi-lan honetatik kanpo geratu den heinean, etorkizuneko lan interesgarria da, ikerketa horretarako abiapuntuko tresnak tesi-lan honetatik sortutako NEen sistemak izan daitezkeelarik. Ikerketa-lerro horietatik lor daitezkeen ekarpenak dira, hain zuzen ere, alde teorikotik zein praktikotik, gure ustez, hobekuntza interesgarrienak lortzeko bideak. Hortaz, tesi-lan honetan landutako NEen inguruko hiru ataza nagusien arteko osagarritasuna irekita geratzen den ikerketa-lerroa da.

Bibliografia

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., eta Urizar R. A framework for the automatic processing of Basque. *Proceedings of the Workshop on Lexical Resources for Minority Languages*, 1998.

Aduriz I., Aranzabe M., Arriola J.M., eta Díaz de Ilarraza A. Sintaxi Partziala. *Andolin Gogoan: Essays in Honour of Professor Eguzkitza*. UPV/EHU Argitalpen Zerbitzua, 2006a.

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World.*, 1–15. Rodopi, 2006b.

Agirre E., Aldezabal I., eta Pociello E. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordnet. *GOGOA aldizkaria*, 237–266, 2006.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Urkia M., Maritxalar M., eta Sarasola K. XUXEN: A Spelling Checker/Corrector for Basque Based on Two-Level Morphology. *Proceedings of ANLP'92*, 119–125, 1992.

Agirre E., Chang A., Jurafsky D., Manning C., Spitkovsky V., eta Yeh E. Stanford-UBC at TAC-KBP (Knowledge Base Population). *In Proceedings of the NIST Text Analysis Conference (TAC2009)*, 2009.

Agirre E. eta Soroa A. Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.*, 33–41, 2009.

Agirre E., Soroa A., eta Stevenson M. Graph-based Word Sense Disambiguation of Biomedical Documents. *Bioinformatics. Oxford University Press. Bioinformatics*, 2889–2896, 2010.

Ahn D., Jijkoun V., Mishne G., Muller K., de Rijke M., eta Schlobach K. Using Wikipedia at the TREC QA Track. *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, 2005.

Al-Onaizan Y. eta Knight K. Machine transliteration of names in Arabic text. *Proceedings of the ACL-02 Workshop on Computational approaches to semitic languages*, page 13, 2002a.

Al-Onaizan Y. eta Knight K. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 408, 2002b.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop on Linguistic Databases.*, 2001.

Alegria I., Ansa O., Arregi X., Otegi A., eta Soraluze A. Ihardetsi: A Question Answering system for Basque built on reused linguistic processors. *SEPLN-SALTMIL 2009 Workshop: Information Retrieval and Information Extraction for Less Resourced Languages.*, 37–43, 2009.

Alegria I., Ceberio K., Ezeiza N., Soroa A., eta Hernandez G. Spelling Correction: from two-level morphology to open source. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008.

Alegria I., Etxeberria I., Hulden M., eta Maritxalar M. Porting Basque Morphological Grammars to foma, an Open-Source Tool. *Finite-State Methods and Natural Language Processing Lecture Notes in Computer Science*, 105–113, 2010.

Alegria I., Ezeiza N., Oronoz M., eta Urizar R. Extracción Automática de Terminología a partir de Etiquetado y Lematización. *VI Simposio Internacional de Comunicación Social.*, 1999.

Arévalo M., Carreras X., Márquez L., T. Martí L.P., eta Simon M. A Proposal for Wide-Coverage Spanish Named Entity Recognition. *SEPLN Journal "Procesamiento del Lenguaje Natural*, 63–80, 2002.

Arregi O. eta Fernández I. Clasificación de documentos escritos en euskara: impacto de la lematización. *I Jornadas de Tratamiento y Recuperación de Información, JOTRI*, 2002.

Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., García E., Lascurain V., Sarasola K., Soroa A., eta Uria L. Semiautomatic conversion of the *Euskal Hiztegia* Basque Dictionary to a queryable electronic form. *T.A.L. journal*, 107–124, 2003.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko teknikaren bidez: kateen eta perpausen identifikazioa eta bere erabilera komazuzentzaile baterako.* Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, 2010.

Artiles J., Gonzalo J., eta Sekine S. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. *Proceedings of the SemEval-2007*, 64–69, 2007.

Artiles J., Gonzalo J., eta Verdejo F. A testbed for people searching strategies in the www. *Proceedings of the 28th annual International ACM SIGIR Conference on Research and development in information retrieval*, 569–570, 2005.

Atserias J., Casas B., Comelles E., González M., Padró L., eta Padró M. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC06)*, 48–55, 2006.

Baeza-Yates R. eta Ribeiro-Neto B. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

Bagga A. eta Baldwin B. Entity-based cross-document coreferencing using the Vector Space Model. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 79–85, 1998.

Banko M. eta Etzion O. The tradeoffs between open and traditional relation extraction. *In Proceedings of 46th AnnualMeeting of the Association for Computational Linguistics: Human Language Technologies*, 28–36, 2008.

Beesley K. eta Karttunen L. *Finite state morphology.* The University of Chicago Press, 2003.

Bekkerman R. eta McCallum A. Disambiguating Web appearances of people in a social network. *Proceedings of the 14th International Conference on World Wide Web*, 463–470, 2005.

Bender O., Och F.J., eta Ney H. Maximum Entropy Models for Named Entity Recognition. *Proceedings of CoNLL-2003*, 148–151, 2003.

Bikel D., Schwartz R., eta Weischedel R. An algorithm that learns what's in a name. *Machine learning*, 211–231, 1999.

Borthwick A., Sterling J., Agichtein E., eta Grishman R. NYU: Description of the MENE Named Entity System as Used in MUC-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

Brighton H. eta Mellish C. Advances in instance selection for instance-based learning algorithms. *Data mining and knowledge discovery*, 153–172, 2002.

Brin S. eta Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 107–117, April 1998. ISSN 0169-7552.

Brown P., Pietra V., Pietra S., eta Mercer R. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist. 19*, 263–311, 1993.

Bunescu R. eta Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation. *Proceedings of 1th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 9–16, 2006.

Burger J., Henderson J., eta Morgan W. Statistical Named Entity Recognizer Adaptation. *Proceedings of CoNLL-2002*, 163–166, 2002.

Bysani P., Reddy K., Reddy V., Kovelamudi S., Pingali P., eta Varma V. IIIT Hyderabad in Guided Summarization and Knowledge Base Population. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

Carreras X., Màrques L., eta Padró L. Named Entity Extraction using AdaBoost. *Proceedings of CoNLL-2002*, 167–170, 2002.

Carreras X., Màrquez L., eta Padró L. A simple named entity extractor using AdaBoost. *Proceedings of the seventh Conference on Natural language learning at HLT-NAACL*, page 155, 2003.

Chang A., Spitkovsky V., Yeh E., Agirre E., eta Manning C. Stanford-UBC Entity Linking at TAC-KBP. *Proceedings of the Third Text Analysis Conference (TAC 2010)*, 2010.

Chen C. eta Chen H. A high-accurate chinese-english ne backward translation system combining both lexical information and web statistics. *Proceedings of the COLING/ACL on Main Conference poster sessions*, 81–88. Association for Computational Linguistics, 2006.

Chinchor N. Overview of MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

Collier N. eta Hirakawa H. Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using Katakana matching. *In Proceedings of the 3rd Natural Language Pacific Rim Symposium*, 309–314, 1997.

Cortes C. eta Vapnik V. Support-vector networks. *Machine Learning*, 273–297, 1995.

Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP07)*, 2007.

Dale R. Symbolic Approaches to Natural Language Processing. *Handbook of Natural Language Processing*. Marcel Dekker Inc, 2000.

Dale R. Language Technology. *In Slides of HCSNet Summer School course, Sydney, Australia*, 2007.

Daumè H. Support Vector Machines for Natural Language Processing. *CSCI 544, Notes to Accompany the Lecture*, 2004.

de Pablo-Sánchez C., Perea J., eta Martínez P. Combining similarities with regression based classifiers for Entity Linking at TAC 2010. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

Díaz de Ilarraza A., Gojenola K., eta Oronoz M. Reusability of a corpus and a treebank to enrich verb subcategorisation in a dictionary. *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP07)*, 280–284, 2007.

Euskaltzaindia. Euskaltzaindiaren 156. araua. 2009.

Ezeiza N., Alegria I., Arriola J.M., Urizar R., eta Aduriz I. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume*, 380–384, 1998.

Farkas R., Szarvas G., eta Ormándi R. Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web. *Advances in Data Mining. Theoretical Aspects and Applications*, 163–172, 2007.

Fellbaum C. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Fernández N., Fisteus J., Sánchez L., eta Martín E. WebTLab: A cooccurrence-based approach to KBP 2010 Entity-Linking task. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

Fleischman M. Fine grained classification of named entities. *Proceedings of the ACL Student Workshop*, 2001.

Fleischman M. eta Hovy E. Fine grained classification of named entities. *Proceedings of the 19th International Conference on Computational linguistics*, 1–7. Association for Computational Linguistics, 2002.

Florian R., Ittycheriah A., Jing H., eta Zhang T. Named Entity Recognition through Classifier Combination. In Daelemans W. eta Osborne M., editors, *Proceedings of CoNLL-2003*, 168–171. Edmonton, Canada, 2003.

Gabrilovich E. eta Markovitch S. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2006.

Gabrilovich E. eta Markovitch S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 6–12, 2007.

Gaizauskas R., Humphreys K., Cunningham H., eta Wilks Y. University of Sheffield: description of the LaSIE system as used for MUC-6. *Proceedings of the 6th Conference on Message understanding*, 207–220. Association for Computational Linguistics, 1995.

Gao W., Wong K., eta Lam W. Phoneme-Based Transliteration of Foreign Names for OOV Problem. *Natural Language Processing IJCNLP 2004*, Lecture Notes in Computer Science, 110–119. Springer Berlin / Heidelberg, 2005.

Gentile A., Zhang Z., Xia L., eta Iria J. Cultural Knowledge for Named Entity Disambiguation: A Graph-Based Semantic Relatedness Approach. *Serdica Journal of Computing*, 217–242, 2010.

Gooi C. eta Allan J. Cross-document coreference on a Large Scale Corpus. *HLT-NAACL 2004 Main Proceedings.*, 9–16, 2004.

Grishman R. eta Sundheim B. Message Understanding Conference-6: a brief history. *Proceedings of the 16th Conference on Computational linguistics*, 466–471. Association for Computational Linguistics, 1996.

Gurrutxaga A., Saralegi X., Ugartetxea S., eta Alegria I. Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa. *Mendebalde Kultur Alkartea, IX. Jardunaldiak: Euskera zientifiko-teknikoa. ISBN 84-931882-5-5*, 2005.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., eta Witten I. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 10–18, 2009.

Hassell J., Aleman-Meza B., eta Arpinar I. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text. *The Semantic Web - ISWC 2006*, Lecture Notes in Computer Science, 44–57. Springer Berlin / Heidelberg, 2006.

Haveliwala T. Topic-sensitive pagerank. *In WWW 2002*, 517–526, 2002.

Hobbs J., Bear J., Israel D., Kameyama M., Kehler A., Martin D., Myers K., eta Tyson M. SRI international FASTUS system MUC-6 test results and analysis. *In Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 237–248. NIST, Morgan-Kaufmann Publishers, 1995.

Huang F. Improved Named Entity Translation and Bilingual Named Entity Extraction. *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, 253–258, 2002.

Humphreys K., Gaizauskas R., Azzam S., Huyck C., Mitchell B., Cunningham H., eta Wilks Y. University Of Sheffield: Description Of The Lasie-II System As Used For MUC-7. *In Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Morgan, 1998.

Jaleel N. eta Larkey L. Statistical transliteration for english-arabic cross language information retrieval. *Proceedings of the twelfth International Conference on Information and knowledge management*, 139–146. ACM, 2003.

Ji H., Grishman R., Dang H., Griffitt K., eta Ellis J. Overview of the TAC 2010 Knowledge Base Population Track. *Proceedings of Text Analysis Conference(TAC 2010)*, 2010.

Jiang L., Zhou M., C. L.C., eta Niu. Named entity translation with web mining and transliteration. *Proc. of IJCAI*, 1629–1634, 2007.

Karimi S. *Machine transliteration of proper names between English and Persian*. Doktoretza-tesia, RMIT University, Melbourne, 2008.

Karimi S., Scholer F., eta Turpin A. Machine transliteration survey. *ACM Comput. Surv.*, 17:1–17:46, 2011.

Karttunen L., Chanod J., Grefenstette G., eta Schiller A. Regular Expressions for Language Engineering. *Natural Language Engineering*, 305–328, 1996.

Kazama J. eta Torisawa K. Exploiting Wikipedia as external knowledge for named entity recognition. *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 698–707, 2007.

Khalid M., Jijkoun V., eta de Rijke M. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. *Advances in Information Retrieval*, 705–710, 2008.

Kilgarriff A. eta Grefenstette G. Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 333–347, 2003.

Klein D., Smarr J., Nguyen H., eta Manning C.D. Named Entity Recognition with Character-Level Models. *Proceedings of CoNLL-2003*, 180–183, 2003.

Klementiev A. eta Roth D. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 817–824. Association for Computational Linguistics, 2006.

Kozareva Z. *Resolving Named Entity Problems: from Recognition and Discrimination to Semantic Class Learning.* Doktoretza-tesia, Department of Languages and Information Systems, University of Alicante, 2009.

Kukich K. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, page 439, 1992.

Kulkarni A. Unsupervised discrimination and labeling of ambiguous names. *Proceedings of the ACL Student Research Workshop.*, 2005.

Labaka G. *EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation.* Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila (UPV-EHU), 2010.

Lam W., Chan S., eta Huang R. Named entity translation matching and learning: With application for mining unseen translations. *ACM Transactions of Infotmation Systems (TOIS)*, 2007.

Larkey L. eta Croft W. Combining classifiers in text categorization. *Proceedings of the 19th annual International ACM SIGIR Conference on Research and development in information retrieval*, 289–297. ACM, 1996.

Larrañaga M., Rueda U., Elorriaga J., eta Arruarte A. Index Analysis: a Means to Acquire the Domain Module Structure. *In proceedings of 10th Conference of the Spanish Association for Artificial Intelligence and 5th Conference on Technology Transfer*, 339–342, 2003.

Lee C., Chang J., eta Jang J. Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. *ACM Transactions on Asian Language Information Processing (TALIP)*, 121–145, 2006.

Lee S. eta Lee G. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping. *Natural Language Processing IJCNLP 2005*, 658–669, 2005.

Leturia I., Gurrutxaga A., Alegria I., eta Ezeiza A. CorpEus, a web as corpus tool designed for the agglutinative nature of Basque. *WAC3 2007 (Web as a Corpus) Workshop*, 2007.

Li F., Zheng Z., Bu F., Tang Y., Zhu X., eta Huang M. THU QUANTA at TAC 2009 KBP and RTE Track. *In Proceedings of the NIST Text Analysis Conference (TAC2009)*, 2009.

Li H., Sim K., Kuo J., eta Dong M. Semantic transliteration of personal names. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, page 120127, 2007.

Li H., Zhang M., eta Su J. A joint source-channel model for machine transliteration. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 159166. Association for Computational Linguistics, 2004.

Lopez de Lacalle O. *Domain-Specific Word Sense Disambiguation.* Doktoretza-tesia, Lengoiaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia 2009ko Abenduaren 14ean., 2009.

Lopez de Lacalle O. eta Agirre E. Hitzen Adiera-desanbiguazioa. *EKAIA*, 127–156, 2010.

Magnini B., Negri M., Prevete R., eta Tanev H. A wordnet-based approach to named entities recognition. *COLING-02 on SEMANET: building and using semantic networks-Volume 11*, page 7, 2002.

Malouf R. Markov Models for language-independent named entity recognition. *Proceedings of CoNLL-2002*, 187–190, 2002.

Mann G. eta Yarowsky D. Unsupervised personal name disambiguation. *Proceedings of the seventh Conference on Natural language learning at HLT-NAACL 2003*, 33–40, 2003.

Manning C. eta Schutze H. *Foundations of statistical natural language processing.* The MIT Press, 2003.

Martinez D. *Supervised Word Sense Disambiguation: Facing Current Challenges.* Doktoretza-tesia, Informatika Fakultatea, UPV-EHU, 2004.

Matusov E., Ueffing N., eta Ney H. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. *In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 33–40, 2006.

Maynard D., Tablan V., Bontcheva K., eta Cunningham H. Rapid customization of an information extraction system for a surprise language. *ACM Transactions on Asian Language Information Processing (TALIP)*, 295–300, 2003.

McDonald D. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, 21–39, 1996.

McNamee P. HLTCOE Efforts in Entity Linking at TAC KBP 2010. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

McNamee P. eta Dang H. Overview of the TAC 2009 Knowledge Base Population Track. *Proceedings of the Second Text Analysis Conference(TAC 2009)*, 2009.

McNamee P., Dredze M., Gerber A., Garera N., Finin T., Mayfield J., Piatko C., Rao D., Yarowsky D., eta Dreyer M. HLTCOE Approaches to Knowledge Base Population at TAC 2009. *In Proceedings of the NIST Text Analysis Conference (TAC2009)*, 2009.

McNamee P. eta Mayfield J. Entity Extraction Without Language-Specific Resources. *Proceedings of CoNLL-2002*, 183–186, 2002.

Merchant R., Okurowski M., eta Chinchor N. The multilingual entity task (MET) overview. *Proceedings of a Workshop on held at Vienna, Virginia: May 6-8, 1996*, 445–447, 1996.

Mikheev A., Grover C., eta Moens M. Description of the LTG system used for MUC-7. *Proceedings of 7th Message Understanding Conference (MUC-7)*, 1998.

Mikheev A., Moens M., eta Grover C. Named entity recognition without gazetteers. *Proceedings of the ninth Conference on European Chapter of the Association for Computational Linguistics*, 1–8, 1999.

Milne D. An open-source toolkit for mining Wikipedia. *Proceedings of New Zealand Computer Science Research Student Conference, NZCSRSC09*, 2009.

Mitchell M., editor. *Machine Learning*. New York: McGraw-Hill, 1997.

Moore R. Learning translations of named-entity phrases from parallel corpora. *Proceedings of EACL*, 259–266, 2003.

Nadeau D. eta Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 3–26, 2007.

Niu C., Li W., eta R.K.Srihari. Weakly supervised learning for cross-document person name disambiguation supported by information extraction. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.

Nomoto T. Multi-engine machine translation with voted language model. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04. Association for Computational Linguistics, 2004.

Och F. eta Ney H. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 19–51, 2003.

Oh J., Choi K., eta Isahara H. Improving Machine Transliteration Performance by Using Multiple Transliteration Models. *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, Lecture Notes in Computer Science, 85–96. Springer Berlin / Heidelberg, 2006a.

Oh J., Choi K., eta Isahara H. A machine transliteration model based on correspondence between graphemes and phonemes. *ACM Transactions on Asian Language Information Processing (TALIP)*, 185–208, 2006b.

Oh J. eta Isahara H. Machine transliteration using multiple transliteration engines and hypothesis re-ranking. *In Proceedings of the 11th Machine Translation Summit*, 353–360, 2007.

Oronoz M. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura.* Doktoretzatesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, 2009.

Padró M. eta Padró L. Applying a Finite Automata Acquisition Algorithm to Named Entity Recognition. *Finite-State Methods and Natural Language Processing*, Lecture Notes in Computer Science, 203–214, 2006.

Palmer D. eta Day D. A Statistical Profile of the Named Entity Task. *Proceedings of the ACL Conference For Applied Natural Lalnguage Processing*, 190–193, 1997.

Patrick J., Whitelaw C., eta Munro R. SLINERC: The Sydney Language-Independent Named Entity Recogniser and Classifier. *Proceedings of CoNLL-2002*, 199–202, 2002.

Pedersen T. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 63–69. Morgan Kaufmann Publishers Inc., 2000.

Pociello E. *Euskararen ezagutza-base lexikala: Euskal WordNet.* Doktoretza-tesia, Euskal Filologia Saila (EHU/UPV). Leioa. 2008ko otsailaren 28an., 2008.

Pouliquen B., Steinberger R., Ignat C., Temnikova I., Widiger A., Zaghouani W., eta Zizka J. Multilingual person name recognition and transliteration. *Arxiv preprint cs/0609051*, 2006.

Quinlan J. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

Radford W., Nothman J., Honnibal M., eta Curran J. CMCRC at TAC10: Document-level Entity Linking with Graph-based Re-ranking. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

Rau L. Extracting company names from text. *Proceedings., Seventh IEEE Conference on In Artificial Intelligence Applications*, 29–32, 1991.

Reeder F., Miller K., Doyon K., eta White J. The Naming of Things and the Confusion of Tongues. *Proceedings of the 4th ISLE Evaluation Workshop, MT Summit VIII. Santiago de Compostela, Spain*, 2001.

Salton G. eta McGill M. *Introduction to Modern Information Retrieval.* McGraw Hill, 1983.

Sekine S. eta Nobata C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of Conference on Language Resources and Evaluation*, 2004.

Smith J., Quirk C., eta Toutanova K. Extracting parallel sentences from comparable corpora using document level alignment. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 403–411. Association for Computational Linguistics, 2010.

Soraluze A., Alegria I., Ansa O., Arregi O., eta Arregi X. Recognition and Classification of Numerical Entities in Basque. *Proceedings of Recent Advances in Natural Language Processing*, 2011.

Sorg P. eta Cimiano P. Cross-lingual information retrieval with explicit semantic analysis. *Working Notes of the Annual CLEF Meeting*, 2008a.

Sorg P. eta Cimiano P. Enriching the crosslingual link structure of Wikipedia-A classification-based approach. *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008b.

Sproat R., Tao T., eta Zhai C.X. Named entity transliteration with comparable corpora. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 73–80, 2006.

Srihari R., Niu C., eta Li W. A Hybrid Approach for Named Entity and Sub-Type Tagging. *Applied Natural Language Processing Conference*, 2001.

StatSoft. *Electronic Statistics Textbook*. Tulsa, OK: StatSoft, 2007.

Strube M. eta Ponzeto S. WikiRelate!:Computing semantic relatedness using Wikipedia. *Proceedings of the AAAI 2006*, 1419–1424, 2006.

Tanaka K. eta Iwasaki H. Extraction of lexical translations from non-aligned corpora. *Proceedings of the 16th Conference on Computational linguistics*, 580–585, 1996.

Tao T., Yoon S., Fister A., Sproat R., eta Zhai C. Unsupervised named entity transliteration using temporal and phonetic correlation. *Proceedings of EMNLP 2006*, 250–257, 2006.

Tjong Kim Sang E. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *Proceedings of CoNLL*, 155–158, 2002.

Tjong Kim Sang E. eta De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of CoNLL*, 2003.

Toral A. eta Muñoz R. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. *Proceedings of Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics.*, 2006.

Varma V., Bharat V., Kovelamundi S., Bysani P., GSK S., N K., Reddy K., Kumar K., eta Maganti N. IIIT Hyderabad at TAC 2009. *In Proceedings of the NIST Text Analysis Conference (TAC2009)*, 2009.

Verdejo M., Gonzalo J., Márquez L., Padró L., Rodríguez H., eta Agirre E. HERMES, Hemerotecas electrónicas: Recuperación multilingüe y extracción semántica. *Jornada de Seguimiento de Proyectos en Tecnologías del Software. Programa Nacional de Tecnologías de la Información y las Comunicaciones*, 2003.

Vincze V., Nagy T. I., eta Berend G. Multiword expressions and named entities in Wikipedia articles. *Proceedings of Recent Advances in Natural Language Processing*, 2011.

Virga P. eta Khudanpur S. Transliteration of proper names in cross-lingual information retrieval. *Proceedings of the ACL 2003 Workshop on Multilingual and mixed-language Named Entity Recognition*, 57–64. Association for Computational Linguistics, 2003.

Wan S. eta Verspoor C. Automatic english-chinese name transliteration for development of multilingual resources. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1352–1356. Association for Computational Linguistics, 1998.

Wentland W., Knopp J., Silberer C., eta Hartung M. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. *Proceedings of the Sixth International Language Resources and Evaluation (LREC08)*, 2008.

Whitelaw C. eta Patrick J. Named entity recognition using a character-based probabilistic approach. *Proceedings of the seventh Conference on Natural language learning at HLT-NAACL*, 2003.

Witten I. eta Frank E. *Data Mining. Practical Machine Learning Tools and Techniques with Java Inplementations*. Morgan Kaufmann Publishers, 2005.

Yeh E., Ramage D., Manning C., Agirre E., eta Soroa A. WikiWalk: Random walks on Wikipedia for Semantic Relatedness. *ACL Workshop TextGraphs-4: Graph-based Methods for Natural Language Processing*, 2009.

Zapirain B. *Rol Semantikoen Etiketatze Automatikoa: Rol Multzoak eta Hautapen Murriztapenak*. Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila (EHU-UPV). Informatika Fakultatea, 2011.

Zhang M., Li H., eta Su J. Direct orthographical mapping for machine transliteration. *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

Zhang W., Sim Y., Su J., eta Tan C. NUS-I2R: Learning a Combined System for Entity Linking. *In Proceedings of the NIST Text Analysis Conference (TAC2010)*, 2010.

A. ERANSKINA

Eiheraren euskarazko NEak identifikatzeko gramatika

Gramatika aplikatzeko testua jarraian azaltzen den formatuan idatzia egon behar du, non hitz bakoitzeko forma, lema eta informazio morfonsintaktikoa adierazita egon behar duen: *@S* karaktere-katearekin forma zein den adierazten da, *@L*rekin lema eta azkenik *@M*rekin informazio morfosintaktikoa. Informazio morfosntaktikoan hitzaren kategoria, azpikategoria eta deklinabidekasua adierazten dira. Esaterako, *Benito Lertxundi asko gustatzen zait baina berdin entzun dezaket Ismael Serrano* esaldia ondoko formatuan egon behar du bertako entitate-izenak gramatikarekin erauzi ahal izateko:

```
@SBenito@LBenito@MIZEIZBEZZ@@@SLertxundi@LLertxundi@MIZEIZBEZZ
@@@Sasko@Lasko@MADBADOARREZZ@@@Sgustatzen@Lgustatu@MADISININE
@@@Szait@Lizan@MADTA1EZZ@@@Sbaina@Lbaina@MLOTJNTEZZ
@@@Sberdin@Lberdin@MADBADOARREZZ@@@Sentzun@Lentzun@MADTMDNCEZZ
@@@Sdezaket@Lezan@MADLA5EZZ@@@SIsmael@LIsmael@MIZEIZBEZZ
@@@SSerrano@LSerrano@MIZEIZBEZZ@@
```

Jarraian gramatikako osagaiak eta erregelak azaltzen dira:

```
# Lehenbizi NEetan ager daitezkeen karaktereak definitzen dira:

define AlfabetoSar [a | b | c | d | e | f | g | h | i | j | k |
```

```
       l | m | n |ñ | o | p | q | r | s | t | u | v | w | y | z |
       %- | x | %_ | %.]  ;

define AlfabetoKat [A | B | C | D | E | F | G | H | I | J | K |
       L | M | N |Ñ | O | P | Q | R | S | T | U | V | W | Y | Z |
       %- | X | %_ | %.]  ;

define Zifrak [%0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | %.  | %,] ;

# Ondoren, karaktere horien konbinazioekin mota desberdinetako
formak zein lemak:

define Forma [ "@S" [AlfabetoSar | AlfabetoKat | Zifrak]+ ];
       # adibidez @S Eibar

define FormaM [ "@S" AlfabetoKat [AlfabetoKat | AlfabetoSar]+ ];
       # adibidez @S EHUko

define FormaMM [ "@S" AlfabetoKat [AlfabetoKat]+ ];
       # adibidez @S EHU

define FormaIdent [ "@S" [AlfabetoKat|Zifrak] (%-)
       [AlfabetoKat|AlfabetoSar|Zifrak]+];
       # adibidez @S EITB2-ko

define Lema [ "@L" [AlfabetoSar | AlfabetoKat | Zifrak | "*"]+ ];
       # adibidez @L EHU

# Eustaggeren informazio morfosintaktikoa erabiliko denez,
deklinabide-kasuen etiketen definizioa burutzen da:

define Etik [AlfabetoKat AlfabetoKat AlfabetoKat];

define ABS [A B S];

define GEN [G E N | G E L];

define EZZ [E Z Z];

define DEK [ Etik - E Z Z ];

define DEKtart [DEK - G E L];

define DEK1 [DEKtart - G E N];
```

```
define EtikMorf [EZZ | DEK | GEN ];

# Testuko forma, lemak eta etiketen definizioekin, testuko
osagaien definizio orokorra burutzen da:

define Token [Forma Lema "@M" Etik+ "@@"];

define TokenMAIUSK [FormaMM Lema "@M" Etik+ "@@"];

define TokenIZENA [Forma Lema "@M" "IZE" Etik EtikMorf "@@"];

define TokenADJ [FormaM Lema "@M" "ADJ" "IZO" EtikMorf "@@"];

define TokenSIGLA [FormaM Lema "@M" "SIG" Etik EtikMorf "@@"];

define TokenIDENT [FormaIdent Lema "@M" "IDENT" "@@"];
       # ETB1, ETB2 eta horrelako egiturak identifikatzeko
       (etiketan IDENT azaltzen da)

define TokenERROM [FormaM Lema "@M" "ERROM" "@@"];
       # II.a eta horrelako identifikatzeko egitura
       (etiketan ERROM azaltzen da)

define TokenBAK [FormaM Lema "@M" "BAK" "@@"];

# Osagaien deklinabide-kasuei eta kategoria/azpikategoriak
kontuan hartuta, tokenen definizioak:

define TokenIZEEZ [FormaM Lema "@M" "IZE" Etik EZZ "@@"];

define TokenIZEGEN [FormaM Lema "@M" "IZE" Etik GEN "@@"];

define TokenAURGEN [Forma Lema "@M" [ "IZE" | "ADJ" | "ADI"]
       Etik "GEN" "@@"];

define TokenIZEDEK [FormaM Lema "@M" "IZE" Etik DEK1 "@@"];

define TokenADJDEK [FormaM Lema "@M" "ADJ" "IZO" DEK1 "@@"];

define TokenIZENB [FormaM Lema "@M" "IZE" "IZB" EtikMorf "@@"];

define TokenIZENLB [FormaM Lema "@M" "IZE" "LIB" EtikMorf "@@"];

define TokenIZEARR [FormaM Lema "@M" "IZE" "ARR" EtikMorf "@@"];
```

```
define TokenTarteko [Forma Lema "@M" "BST" EtikMorf "@@"];

define TokenTartekoM [FormaM Lema "@M" "BST" EtikMorf "@@"];

define TokenSIGLADEK [FormaM Lema "@M" "SIG" Etik DEK1 "@@"];

define TokenIZENBDEK [FormaM Lema "@M" "IZE" "IZB" DEK1 "@@"];

define TokenIZENLBDEK [FormaM Lema "@M" "IZE" "LIB" DEK1 "@@"];

define TokenDEK [TokenIZEDEK | TokenSIGLADEK | TokenIZENBDEK |
       TokenIZENLBDEK | TokenADJDEK] ;

define TokenIZEARREzk [FormaM Lema "@M" "IZE" "ARR" [EZZ|GEN] "@@"];

define TokenIZENBEzk [FormaM Lema "@M" "IZE" "IZB" [EZZ|GEN|ABS] "@@"];

define TokenIZENLBEzk [FormaM Lema "@M" "IZE" "LIB" [EZZ|GEN|ABS] "@@"];

define TokenADJEzk [FormaM Lema "@M" "ADJ" "IZO" [EZZ|GEN] "@@"];

define TokenSIGLAEzk [FormaM Lema "@M" "SIG" Etik [EZZ|GEN|ABS] "@@"];

define TokenAurreko [Forma Lema "@M" "IZE" Etik [EZZ|GEN] "@@"];

define TokenEZK [TokenIZEEZ | TokenIZEGEN | TokenSIGLAEzk |
       TokenIZENBEzk | TokenIZENLBEzk | TokenADJEzk ];

define TokenDeklinatuak [TokenSIGLADEK | TokenIZENBDEK | TokenIZENLBDEK ];

# NER atazarako interesgarriak diren Token guztiak definituta,
NEak definitzen dira:

# Osagai bakarreko NEen definizioa:

define PAT1 [TokenSIGLA | TokenIZENB | TokenIZENLB | TokenMAIUSK |
       TokenIZEARR | TokenIDENT | TokenBAK];

# Osagai bat baino gehiagoko entitatearen definizioa:

define PAT2 [[TokenIDENT | ([TokenTarteko | TokenTartekoM]*) TokenEZK |
      TokenBAK] [([TokenTarteko | TokenTartekoM]*) TokenEZK
      ([TokenTarteko | TokenTartekoM]*)| TokenBAK]*
      [([TokenTarteko | TokenTartekoM]*) TokenEZK | TokenDEK |
      TokenERROM | TokenBAK]];
```

```
# NEen definizioa, edozein delarik bere osagai kopurua:

define PAT [PAT1 | PAT2];

# NEak kakoekin mugatzeko erregela:

define MugaMuga [PAT @-> "[" ... "]"];

# NEen inguruan hitz eragilerik balego hau ere mugatzeko erregelak.

define PAT3 [ "["
      [TokenIDENT |([TokenTarteko | TokenTartekoM]*) TokenEZK | TokenBAK]
      [([TokenTarteko | TokenTartekoM]*) TokenEZK | TokenBAK]* "]"
      (TokenIZENA)];

define PAT4 [ "["
      [TokenIDENT | ([TokenTarteko | TokenTartekoM]*) TokenEZK | TokenBAK]
      [([TokenTarteko | TokenTartekoM]*) TokenEZK
      ([TokenTarteko | TokenTartekoM]*)| TokenBAK]*
      [TokenDEK | TokenERROM | TokenBAK] "]"];

define PAT5 [ "[" TokenDeklinatuak "]"];

define PAT9 [ "[" TokenMAIUSK "]" | "[" TokenIZEARR "]" |
       "[" TokenIDENT "]" | "[" TokenBAK "]"];

define PAT6 [PAT4 | PAT5 | PAT3 | PAT9];

define Mugatuta [PAT6 @-> ...   "#\n"];

define PAT7 [(TokenAurreko) "["
  [ TokenIDENT | ([TokenTarteko | TokenTartekoM]*) TokenEZK | TokenBAK]
  [([TokenTarteko | TokenTartekoM]*) TokenEZK
  ([TokenTarteko | TokenTartekoM]*)| TokenBAK]*
  ([ TokenDEK | TokenERROM | TokenBAK]) "]"];

define PAT8 [PAT7 | PAT5 | PAT9];

define MugatuAurre [ PAT8 @-> "\n#" ...];

# Sortutako tarteko ikurrak testutik ezabatzeko garbiketa erregelak:

define EzabatuSarEtik "@S" -> " " ;
```

```
define EzabatuLemEtik "@L" -> " ";

define EzabatuMorfEtik "@M" -> " " ;

define EzabatuAzkenmarka "@@" -> " " ;

# NER atazarako definitutako erregelak automata bihurtzen
dituzten espresioak:

define EzabatuEtiketak EzabatuSarEtik .o.  EzabatuLemEtik .o.
       EzabatuMorfEtik .o.  EzabatuAzkenmarka;

define MugatuEzabaketa MugatuAurre .o.  EzabatuEtiketak;

read regex Markak;

save stack Izb-LibGelBanatu.fst;

read regex MugaMuga;

save stack EntitateakMugatu.fst;

read regex Mugatuta;

save stack EskuinekoakMugatu.fst;

read regex MugatuEzabaketa;

save stack EzkerrekoakMugatu.fst;
```

B. eranskina

EDBLko kategoria sistema

Kategoria lexikalak:

IZENAK		
	IZE ARR	Arrunta (zuhaitz, bat)
	IZE IZB	Pertsona-izen berezia (Mikel)
	IZE LIB	Leku-izen berezia (Donostia)
ADJEKTIBOAK		
	ADJ IZL	Izen lagun (benetako,nongo)
	ADJ IZO	Izen ondo (handi)
ADITZAK		
	ADI SIN	Sinplea (ekarri)
	ADI ADK	Konposatua (lo egin)
	ADI ADP	Perifrasikoa (ahal izan)
	ADI FAK	Faktiboa (etorrarazi)
ADBERBIOAK		
	ADB ADOARR	Aditzondo arrunta (gaur)
	ADB ADOGAL	Aditzondo galdetzailea (noiz)
	ADB ALGARR	Aditz laguntzaile arrunta (negarrez)
	ADB ALGGAL	Aditz laguntzaile galdetzailea (nora)

DETERMINATZAILEAK		
	DET ERKARR	Erakusle arrunta (hau)
	DET ERKIND	Erakusle indartua (berori)
	DET NOLARR	Nolakotzaile arrunta (edozein)
	DET NOLGAL	Nolakotzaile galdetzailea(zein)
	DET DZH	Zenbatzaile zehaztua (bi)
	DET BAN	Zenbatzaile banatzailea (bina)
	DET ORD	Zenbatzaile ordinala (bigarren)
	DET DZG	Zenbatzaile zehaztugabea (zenbait)
	DET ORO	Zenbatzaile orokorra (guzti)
IZENORDAINAK		
	IOR PERARR	Pertsonal arrunta (ni)
	IOR PERIND	Pertsonal indartua (neu)
	IOR IZGMGB	Zehaztugabe mugagabea (norbait)
	IOR IZGGAL	Zehaztugabe galdetzailea (nor)
	IOR ELK	Zehaztugabe Elkarkaria (elkar)
LOTURAZKOAK		
	LOT LOK	Lokailua (hala ere)
	LOT JNT	Juntagailua (edo)
PARTIKULAK		
	PRT	(omen, ote, ...)
INTERJEKZIOAK		
	ITJ	(alajaina!)
BESTELAKOAK		
	BST	(baldin)
ADITZ LAGUNTZAILEAK		
	ADL	(du)
ADITZ SINTETIKOAK		
	ADT	(dator)
SIGLAK		
	SIG	(EHU)
SINBOLOAK		

	SNB	(km, cm, g,...)
LABURDURAK		
	LAB	(etab.)

Deklinabide-kasuak:

ABL	Ablatiboa
ABS	Absolutiboa
ABU	Adlatibo bukaerakoa
ABZ	Adlatibo bide zuzenekoa
ALA	Adlatiboa
BNK	Banakaria
DAT	Datiboa
DES	Destinatiboa
DESK	Deskribatzailea
ERG	Ergatiboa
GEL	Genitiboa (leku-denborazkoa)
GEN	Genitiboa
INE	Inesiboa
INS	Instrumentala
MOT	Motibatiboa
PAR	Partitiboa
PRO	Prolatiboa
SOZ	Soziatiboa

C. ERANSKINA

Euskara-Gaztelera NEen Transliterazio-Gramatika

Euskarazko hitzetatik abiatuta erregela fonologikoak aplikatuz gaztelerazko hitza lortzeko gramatikaren definizioa:

```
# Bokalak eta kontsonanteak definitzen

define Bokal [a | e | i | o | u | ü|A | E | I | O | U | Ü];

define Bokalm [a | e | i | o | u | ü];

define BokalM [A | E | I | O | U | Ü];

define Kontsonante [b | c | d | f | g | h | j | k | l | m | n |
       ñ | p | q | r | s | t | v | w | x | y | z | B | C | D | F
       | G | H | J | K | L | M | N | Ñ | P | Q | R | S | T | V |
       W | X | Y | Z];

define Maiuskulak [A | B | C | D | E | F | G | H | I | J | K |
       L | M | N | Ñ | O | P | Q | R | S | T | U | V | W | X |
       Y | Z];

define Minuskulak [a | b | c | d | e | f | g | h | i | j | k | l |
       m | n | ñ | o | p | q | r | s | t | u | v | w | x | y | z];

# Erregela bakoitza identifikatzeko erabiliko den
# kode multzoaren definizioa:
```

```
define Markak [ "*" R [ "0" | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9]+];

# Bokalen erregela sinpleak definitzen:

define Dieresiau [u (->) "*" R 1, U (->) "*" R 1||
       [g | G] (Markak) _ (Markak) [e | i | E | I]];
       # Ibarguenagoitia > Ibargenagoitia

define Bukaerakoo [u (->) o "*" R 2, U (->) O "*" R 2 || _ (Markak) .#.];
       # Artikulu > Articulo

# Diptongoak desegiteko erregela sinpleak definitzen:

define Diptongoak [e (->) i e "*" R 3, E (->) I E "*" R 3,
       o (->) u e "*" R 3,O (->) U E "*" R 3 ||
       Kontsonante (Markak) _ (Markak) Kontsonante];
       # gobernu -> gobierno , proba -> prueba

# Kontsonanteen bihurketetarako erregela sinpleak definitzen

define Rulek [k (->) c "*" R 4, K (->) C "*" R 4|| _ (Markak)
       [Kontsonante | a | o | u | A | O | U] ,, k (->) q u "*" R 4,
       K (->) Q u "*" R 4 || _ (Markak) [e | i] ,, K (->) Q U "*" R 4
       || _ (Markak) [E | I]]; # Barakaldo -> Baracaldo ,
       Eskozia -> Escocia , Joakin -> Joaquin

define Rulez [z (->) c "*" R 5, Z (->) C "*" R 5 ||
       [Bokal | c | C | .#.]  (Markak) _ (Markak) [e | i | E | I]];
       # Eskozia -> Escocia

define Rulej [j (->) g "*" R 6, J (->) G "*" R 6 ||
       _ (Markak) [e | i | E | I]];
       # jeneral -> general , erlijio -> religion

define Ruleg [g (->) g u "*" R 7, G (->) G u "*" R 7 ||
       _ (Markak) [e | i] ,, G (->) G U "*" R 7 ||
       _ (Markak) [E | I]]; # Arregi -> Arregui

define Rulex [x (->) j "*" R 8, X (->) J "*" R 8||
       Bokal (Markak) _ (Markak) [a | A | o | O | u | U] ,,
       x (->) g "*" R 8, X (->) G "*" R 8, x (->) j "*" R 8,
       X (->) J "*" R 8|| (Markak) _ (Markak) [e | E | i |I]];
       # kexa -> keja

define Rulenb [n b (->) m b "*" R 9, N B (->) M B "*" R 9];
```

```
       # Aranburu -> Aramburu

define Rulenp [n p (->) m p "*" R 1 "0" , N P (->) M P "*" R 1 "0"];
       # enpresa -> empresa

define Ruletx [t x (->) c h "*" R 1 1, T X (->) C H "*" R 1 1,
       T x (->) C h "*" R 1 1]; # Aretxaga -> Arechaga

define Ruletz [t z (->) c "*" R 1 2, T [Z | z] (->) C "*" R 1 2 ||
       (Markak) _ (Markak) [e | i | E | I] ,, t z (->) z "*" R 1 2,
       T [z | Z] (->) Z "*" R 1 2 || _ (Markak)
       [a | o | u | A | O | U | .#.]];
       # Arantzeta -> Aranceta , Iturgaitz -> Iturgaiz

define Rulets [t s (->) s "*" R 1 3, T [s |S] (->) S "*" R 1 3 ||
       [l | n | r |L | N | R] (Markak) _ ];
       # Alsasua -> Altsasua , inversiones -> inbertsio

define Ruleil [i l (->) l l "*" R 1 4, I l (->) L l "*" R 1 4,
       I L (->) L L "*" R 1 4 || Bokal (Markak) _ (Markak) Bokal .o.
       l (->) l l "*" R 1 4, L (->) L l "*" R 1 4||
       Bokalm (Markak) _ (Markak) Bokalm .o.  L (->) L L "*" R 1 4 ||
       BokalM (Markak) _ (Markak) BokalM];
       # milioi -> millon , ametrailadora -> ametralladora

define Rulein [i n (->) i ñ "*" R 1 5, I N (->) I Ñ "*" R 1 5 ||
       Kontsonante (Markak) _ (Markak) Bokal .o.  i n (->) ñ "*" R 1 5,
       I [n | N] (->) Ñ "*" R 1 5 || Bokal (Markak) _ (Markak) Bokal];
       # ikurrina -> ikurriña , kanpaina -> campaña

define RuleRhas [[e | a] r r (->) r "*" R 1 6,
       [E | A] [r | R] [r | R] (->) R "*" R 1 6
       || .#.  (Markak) _ ,, e r l (->) r e l "*" R 1 6 ,
       E r l (->) R e l "*" R 1 6, E R L (->) R E L "*" R 1 6 ||
       .#.  (Markak) _];
       # Erreferentzia -> Referencia, erlijio -> relijion

define Ruleb [b (->) v "*" R 1 7, B (->) V "*" R 1 7];
       # Bolivar -> Bolibar

define Ruleiy [i (->) y "*" R 1 8, I (->) Y "*" R 1 8||
       [[[q | Q | g |G ] (Markak) [Bokal - [u | U]]] |
       [[[Kontsonante - [q | Q | g |G ] - ErromUnitateak] | Bokal - I]
       (Markak) Bokal - I] | 0 (Markak) Bokal - I]
       (Markak) _ (Markak) [Bokal - I | .#.]];
```

```
       # Irigoien -> Irigoyen baina III -> IYI ez

define Rulejy [j (->) y "*" R 1 9, J (->) Y "*" R 1 9 ||
       .#.  (Markak) _];
# Jugoslavia -> Yugoslabia

# Atzizkien erregelak definitzen:

define Ruleon [i o (->) i o n "*" R 2 2, I O (->) I O N "*" R 2 2 ||
       [z | s | Z | S | c | C] (Markak) _ (Markak) .#.  .o.
       o i (->) o n "*" R 2 2 || _ (Markak) .#.];
       # fundazio -> fundacion , balkoi -> balcon

define Ruletu [t u (->) d o "*" R 2 3, T U (->) D O "*" R 2 3||
       [a | A] (Markak) _ (Markak) .#.];
       # homologatu -> homologado

define Ruleia [i a (->) j e "*" R 2 4, I A (->) J E "*" R 2 4 ||
       [a | A] (Markak) _ (Markak) .#.];
       # lengoaia -> lengoaje

define Ruledad [t a t e (->) d a d "*" R 2 5,
       T A T E (->) D A D "*" R 2 5 || _ (Markak) .#.];
       # unibertsitate -> universidad

define Ruledu [ d u (->) t o "*" R 2 6, D U (->) T O "*" R 2 6||
       [m | M] (Markak) [i | I] (Markak) [e | E] (Markak) [n | N]
       (Markak) _ (Markak) .#.];
       # Aurretik ie diptongoa ebatzita egon behar du.
       # establezimendu -> establecimiento

# Hobekuntzetarako proposatutako erregelak

define ErromUnitateak [I | V | X | L | C | D | M];

define Erromatarrak [ErromUnitateak]+;

define Ruleerrom [ Erromatarrak (->) ... "*" R 2 7 ||
       .#.  _ "." [Minuskulak]+ .#.
       .o. "*" R 2 7 "." [Minuskulak]+ (->) "*" R 2 7 || _ .#.];
       # I.a -> I

define Ruleñia [ i a (->) a "*" R 2 8 || ñ (Markak) _ .#.  ,,
       I A (->) A "*" R 2 8|| Ñ (Markak) _ (Markak) .#.];
       # Bretainia -> Bretañia -> Bretaña
```

```
define Ruleoaa [ o a (->) a "*" R 2 9 , O A (->) A "*" R 2 9 ||
       _ (Markak) .#.];
       # Subkarpatikoa -> Subcarpatico, Katolikoa -> Catolica

define Ruleoao [ o a (->) o "*" R 3 "O" , O A (->) O "*" R 3 "O" ||
       _ (Markak) .#.]; # Batikanoa -> Vaticano

define Ruleioo [ i o (->) o "*" R 3 1 , I O (->) O "*" R 3 1 ||
       _ (Markak) .#.]; # Akordio -> Acuerdo

define Ruleazkenk [ a k (->) s "*" R 3 2, A K (->) S "*" R 3 2 ||
       Bokal (Markak) _ (Markak) .#.  .o.  k (->) s "*" R 3 2 ,
       K (->) S "*" R 3 2 || _ (Markak) .#.];
       # Bahamak-> Bahamas, Eliseoak -> Eliseos

# Erregelak konposatuz transduktorea sortzen:

# Definitutako erregela guztiak ordena egokian biltzen dituen
# erregela edo transduktorea:

define Transduktorea [Dieresiau .o.  Bukaerakoo .o.  Diptongoak
       .o.  Rulek .o.  Rulez .o.  Rulej .o.  Ruleg .o.  Rulex .o.
       Rulenb .o.  Rulenp .o.  Ruletx .o.  Ruletz .o.  Rulets .o.
       Ruleil .o.  Rulein .o.  RuleRhas .o.  Ruleb .o.  Ruleiy .o.
       Rulejy .o.  Ruletu .o.  Ruleia .o.  Ruledad .o.  Ruledu .o.
       Ruleerrom .o.  Ruleia .o.  Ruleoaa .o.  Ruleoao .o.  Ruleioo
       .o.  Ruleazkenk .o.  Ruleon];

# Transduktorea programetatik erabiltzeko garaian behin
# bakarrik kargatu behar izateko alderantzizkoa kalkulatzen:

define Alderantzizkoa [Transduktorea].i;

# XFSTren pilan gorde transduktorea
# (urrats honetan da benetan transduktorea sortzen den unea):

read regex Alderantzizkoa;

# Sortu berri eta pilan gordeta dagoen transduktorea fitxategi
# batean gorde aurrerantzean erabilgarri izateko.
# Fitxategi horrek gordetzen duen transduktorea aplikatzeak,
# zehaztutako ordenan erregela guztien aplikazioa inplikatzen du:

save stack eus-gazt-fonologikoak-markekin-alderantzizkoa.fst;
```

D. ERANSKINA

Euskara-Gaztelera NEen Itzulpenerako Ordenazio Erregelak

NE osoa eraikitzeko orduan, osagaien arteko ordenazioa egiteko hartutako erabaki linguistikoak zerrendatuta eta dokumentatuta agertzen dira ondorengo lerroetan. Gogoratu erregelak osagaika edota osagai bikoteen informazioan oinarrituz definitu dira eta entitatearen azken osagaiaren deklinabidea ez dugula kontuan hartzen.

1. Bikoteko ezkerreko osagaiaren deklinabide-kasua eta numeroa kontsideratuta:

   (a) Ezkerreko osagaia genitibo pluralean deklinatuta dago. Orduan, osagaiak trukatu eta *de los* tartekatu. Adibidea: Ahateen Parkea itzultzeko lematizatzailearen informazioan oinarrituta,
   ```
   /<Ahateen>/<HAS_MAI>/
      ("ahate" IZE ARR DEK GEN NUMP MUGM)
   /<Parkea>/<HAS_MAI>/
      ("parke" IZE ARR DEK ABS NUMS MUGM)
   ```
   Itzulpena: Parque de los Patos

   (b) Ezkerreko osagaia genitibo edo genitibo lokatibo singularrean deklinatuta dago. Orduan, osagaiak trukatu eta *de* tartekatu Adibidea: Ekialdeko Timor
   ```
   /<Ekialdeko>/<HAS_MAI>/
      ("ekialde" IZE ARR DEK NUMS MUGM DEK GEL)
   ```

```
/<Timor>/<HAS_MAI>/
   ("Timor" IZE LIB PLU-)
```
Itzulpena: Timor de oriente

2. Hitz bikoteen kategoria, azpikategoria eta deklinabide-kasuak kontsideratuta, hiru azpimultzo nagusi bereiz daitezke: osagaien ordena trukatu eta tartean *de* partikula txertatu behar diren hitz bikoteak; ordena trukatu baina partikula tartekatzerik behar ez dutenak; eta, azkenik, euskaraz jarraitzen duten ordena bera gaztelerazko itzulpenean mantentzen duten patroiak.

   (a) Osagaiei buelta eman eta *de* partikula tartekatuta behar dituzten hitz bikoteen kasuistikarekin hasiko gara, bakoitzeko lematizaziotik lortzen den informaziotik beharrezko informazioa, eta adibide bana zehaztuz:

      i. Izen berezi edo leku-izen berezi bat eta jarraian izen arrunt bat dagoenean. Adibidea: Lizarra-Garazi Patua
      ```
      /<Lizarra-Garazi>/<DEN_MAI_DEK>/
         ("Lizarra-Garazi" IZE LIB PLU-)
      /<Patua>/
         ("patu" IZE ARR DEK ABS NUMS MUGM)
      ```
      Itzulpena: Pacto de Lizarra-Garazi

      ii. Bi izen arrunt jarraian eta lehenengoa deklinatu gabe dagoenean. Adibidea: Bake Akordio
      ```
      /<Bake>/<HAS_MAI>/
         ("bake" IZE ARR)
      /<Akordioetatik>/<HAS_MAI>/
         ("akordio" IZE ARR DEK NUMP MUGM DEK ABL)
      ```
      Itzulpena: Acuerdo de Paz

      iii. Leku-izen berezi bat eta ondoren deklinatu gabeko izen arrunt bat dagoenean. Adibidea: Valmy Gudua
      ```
      /<Valmy>/<HAS_MAI>/
         ("" /Valmy/ IZE LIB PLU-)
      /<Gudua>/
         ("gudu" IZE ARR DEK ABS NUMS MUGM)
      ```
      Itzulpena: Guerra de Valmy

   (b) Osagaiei buelta eman eta ezer tartekatzea behar ez duten hitz bikoteen kasuistikarekin jarraituko dugu, bakoitzeko lematizaziotik

lortzen den informaziotik beharrezko informazioa eta adibide bana zehaztuz:

i. Deklinatu gabeko adjektibo izen lagun bat eta jarraian izen arrunt bat dagoenean. Adibidea: Erabateko Independentzia

```
/<Erabateko>/<HAS_MAI>/
   ("erabateko" ADJ IZL)
<Independentziarako>/<HAS_MAI>/
   ("independentzia" IZE ARR DEK NUMS MUGM DEK
   ALA DEK GEL)
```

Itzulpena: Independencia Absoluta

(c) Euskarazko forman osagaiek duten ordena mantendu behar duten hitz bikoteen kasuistika aurkezten da jarraian:

i. Bi leku-izen berezi eta lehenengoa deklinatu gabe dagoenean. Adibidea: Sierra Leona

```
/<Sierra>/<HAS_MAI>/
   ("Sierra" IZE LIB PLU-)
/<Leonako>/<HAS_MAI>/
   ("Leon" IZE LIB PLU- DEK NUMS MUGM DEK ALA
   DEK GEL)
```

Itzulpena: Sierra Leona

ii. Bi izen berezi eta lehenengoa deklinatu gabe dagoenean. Adibidea: Balentin Berriotxoa

```
/<Balentin>/<HAS_MAI>/
   ("Balentin" Balentin IZE IZB PLU-)
/<Berriotxoa>/<HAS_MAI>/
   ("" /Berriotxoa/ IZE IZB PLU- DEK ABS MG)
```

Itzulpena: Valentin Berriochoa

iii. Deklinatu gabeko leku-izen berezi baten jarraian adjektibo izenondo bat agertzen denean. Adibidea: Zizur Nagusi

```
/<Zizur>/<HAS_MAI>/
   ("Zizur" IZE LIB PLU-)
/<Nagusi>/
   ("nagusi" ADJ IZO DEK ABS MG)
```

Itzulpena: Cizur Mayor

iv. Deklinatu gabeko izen arrunt baten jarraian adjektibo-izenondo bat agertzen denean. Adibidea: Parke Nazionala

```
/<Parke>/<HAS_MAI>/
   ("parke" IZE ARR)
/<Nazionala>/<HAS_MAI>/
   ("nazional" ADJ IZO DEK ABS NUMS MUGM)
```

Itzulpena: Parque Nacional

(d) Bukatzeko, gogorarazi ordena ere mantenduko dela itzulpena egitean hainbat izen arrunt eta ondoren beste izen arrunt bat dagoenean, lehenengo izen arrunta itzultzean adjektibo kategoria hartzen duten kasu berezietan. Adibidea: Giza Eskubideen Auzitegiak

```
/<Giza>/<HAS_MAI>/
   ("giza" IZE ARR)
/<Eskubideen>/<HAS_MAI>/
   ("eskubide" IZE ARR DEK GEN NUMP MUGM)
/<Auzitegiak>/<HAS_MAI>/
   ("auzitegi" IZE ARR DEK ABS NUMP MUGM)
```

Itzulpena: Juzgados de Derechos Humanos

Tesi honen idazketa
2012ko urtarrilaren 31n
bukatu zen.