LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA

INFORMATIKA FAKULTATEA

CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK

Euskararen etiketatzaile morfosintaktiko sendo eta malgua

Nerea Ezeiza Ramosek

Informatikan Doktore Titulua eskuratzeko aurkezturiko

TESI-TXOSTENA

LENGOAIA ETA SISTEMA INFORMATIKOEN SAILA

INFORMATIKA FAKULTATEA

CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK

Euskararen etiketatzaile morfosintaktiko sendo eta malgua

Nerea Ezeizak Iñaki Alegriaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

“Al final de este viaje en la vida quedarán
nuestros cuerpos hinchados de ir
a la muerte, al odio, al borde del mar.

Al final de este viaje en la vida, quedará
nuestro rastro invitando a vivir.
Por lo menos por eso es que estoy aquí.

Somos prehistoria que tendrá el futuro,
somos los anales remotos del hombre.
Estos años son
cierta agilidad con que el sol te dibuja
en el porvenir,
son la verdad o el fin,
son Dios.

Quedamos los que puedan sonreír
en medio de la muerte, en plena luz.

Al final de este viaje en la vida quedarán
nuestros cuerpos tendidos al sol,
como sábanas blancas después del amor.

Al final del viaje está el horizonte,
al final del viaje partiremos de nuevo,
al final del viaje comienza un camino,
otro buen camino que seguir descalzos,
contando la arena.

Al final del viaje
estamos tú y yo,
intactos.

Quedamos los que puedan sonreír
en medio de la muerte, en plena luz.”

Silvio Rodríguez

"Canciones de Mar"

1970eko urtarrilaren 28an

"Océano Pacífico" itsasontzian, Kubara bueltan iristeko egunean.

eskerrak ematen

• Senitartekoei, bereziki amari, erabakiak hartzen laguntzen didazuelako beti. Ilobei ere nere esker ona, haur baten moduan jolasteko aukera emateagatik, baita egiten dizkidazuen txantxa guztiengatik, Maripuri deituta ere.

• Lourdesi, bidaia luuuuuuuze honetan, ekaitza egon den bakoitzean, bai Azuquecan bai Ávilan zure portua (lehorrekoa, hori bai) eskaini didazulako, zure babesa, adiskidantza eta maitasuna emanez. Tesi honen fase guztien lekuko zuzena izanagatik, nere bigarren *alaba* honen ama ponteko ere bazarelako.

• Mameni, onerako eta txarrerako, beti alboan egoteagatik, eta Karmeleri, Guinness tertuliak antolatzeagatik.

• Itziarri, egiten dituen *mediku-bisita* guztietan nerekin egoteko astia hartzeagatik, irribarrea aurpegian eta zerbeza eskutan.

• Palakideoi, Aitzol, Izaskun eta Mikeli, stressa kentzen eta ondo pasatzen laguntzeagatik, gora palasamba! Eta, nola ez, muskideoi ere, hordago!

• Aitziberri, besteak beste, bere masajitoengatik.

• Ixakide guztioi, lan hau aurrera eramateko beharrezkoa zen laguntza emateagatik. Esker berezia Iñakiri, nerekin izan duen pazientziagatik. Hori da hori meritua zurea!

Gaurtik aurrera hasten den bidaia berrian bidelagun izango zaituztedalakoan, esker mila guztioi!

• Ahaztu gabe, eskerrak Murphy-ri (ala beste izenen bat dauka???), agerraldi batetik bestera lan pixka bat egiteko beta uzteagatik. Hala ere, eskertuko genizuke guztiok gutxiagotan etorriko bazina gurera.

Eskerrak..... amaitu dudan!

Aurkibidea

I Sarrera

I.1 Motibazioa eta helburuak

Euskararen prozesaketa automatikoan urrats garrantzitsu bat izan nahi du aurkezten dugun *Euskararen etiketatzaile morfosintaktiko sendo eta malgua* izeneko ikerketa-lan hau. IXA taldeak duen epe luzerako egitasmo zabalean kokatu behar da eta horretarako hizkuntzalari eta informatikarien artean osaturiko talde bat aritzen gara elkarlanean.

Lanaren helburu nagusia hainbat aplikaziotan erabiliko den euskararen lematizatzaile/etiketatzaile orokor eta sendoa egitea da. Euskara bezalako hizkuntza eranskari batean lan hori egiteko analisi morfologiko sendoa ezinbesteko urratsa da desanbiguazio-prozesuari ekin ahal izateko, hitz-zerrendetan oinarritutako metodoak ez baitira bideragarriak. Are gehiago, desanbiguazio-prozesua optimizatzeko hitzen morfologiatik hara joan behar da, flexiotik eta eratorpenetik hara behintzat, ezaguna baita hitz anitzeko unitateak analizatuz gero desanbiguazioaren emaitzak zehatzagoak izango direla.

Aurrekoaren ondorioz, lau urrats planteatzen dira tesi-lanaren garapenerako:

- Aurretik taldean garatutako analizatzailea morfologikoa, MORFEUS izenekoa, sendotzea eta doitzea. Hizkuntza estandarrerako emaitza onak ematen bazituen ere, doitasunari begira ahuleziak zituen desbiderapen linguistikoen aldetik zein lexikoan agertzen ez diren hitzen aldetik. Gainera, urrats honetan sortzen diren akatsek edo gehiegizko anbiguotasunek eragin kaltegarria dute ondoko urratsetan. Beraz, MORFEUSen emaitzak hobetzea doitasuna eta zehaztasuna galdu gabe da lehen helburua.
- Hitz anitzeko unitateen ezagutza. Lokuzioak, kolokazioak, datak, zenbakiak, entitateak etab. elementu interesgarriak dira hizkuntzaren ingeniaritzan, eta horietako ahalik eta gehienak identifikatzea eta analizatzea izango da lanaren bigarren helburua. Gainera, lan

horretan arrakasta baldin badugu, etiketatzailearen aplikazio-eremua zabaldu egingo da, informazioaren berreskurapena/erauzketa arloko aplikazioetan elementu hauen sailkapena funtsezkoa baita.

- Desanbiguazioa da proposatzen den tresnaren funtsezko urratsa. Txostenean ikusiko den bezala, euskararen eta proposatutako etiketatze-sistemaren ezaugarriak direla eta, anbiguotasun handia dago aurreko urratsen ondoren eta, horri aurre egiteko, erregeletan eta ikasketa automatikoan oinarritutako bi paradigmak konbinatu nahi dira emaitzak hobetzearren. Erregela-sistema taldeko linguistek garatu dute, eta berrerabiliko da lan honetan.
- Ebaluazioa eta aplikazioa. Garatutako tresna erabilgarria dela frogatzeko, batetik, ebaluazio sakona behar da, eta, bestetik, taldean bertan zein enpresetan erabilitako hainbat aplikaziotan integratzea.

I.2 Aurrekariak

Testuen desanbiguazio morfosintaktikoa punta-puntako ikerketa-gaia izan da aurreko hamarkadan, eta emaitzek eragin handia izan dute ikasketa automatikoak lengoaia naturalaren prozesamenduan (LNPan) izan duen arrakastan. Dena den, erregelen bidezko paradigman oinarritutako lan batzuek berdindu dute, eta kasu batzuetan hobetu, metodo estokastikoen emaitzak. Gaia erakargarria izan zen gure talderako eta teknologia hori euskarara aplikatzea erabaki genuen helburu bikoitzarekin: batetik, euskararako tresna erabilgarri bat lortzea, eta bestetik, ekarpen zientifikoa egitea hizkuntzaren ezaugarri bereziengatik zein aplikatzeko moduarengatik.

Lan honen hasiera 1995ean kokatu behar da. Garai hartan, IXA taldean oinarrizko analizatzaile morfologikoaren garapena (Alegria 1995) bukatu bezain laster, lehenago beharrezkotzat jotzen genuen lematizatzaile/etiketatzaile sendoa bideragarria zen, ezinbesteko tresna zen analizatzaile morfologikoa egina baitzegoen. Gure estrategia bikoitza izan zen. Alde linguistikoan, corpus desanbiguatu bat prestatzeari ekiteaz gain, desanbiguazio-erregelak idazteari ekin zitzaion (Aduriz 2000). Horrez gain, hitz anitzeko unitateen zerrenda bat bildu zen modu semiautomatikoan. Informatikari dagokion lana, berriz, tesi honetan azaltzen da, eta helburuak aurreko atalean laburbildu dira. *Euslem, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa* izeneko tesina da hasierako lanaren fruitua (Ezeiza 1997). II. kapituluan, lanaren kokapena egitean, sakontzen da tesiaren abiapuntuaz.

I.3 Metodologia

Ikerketa honi aurre egiteko garaian, IXA ikerketa-taldearen helburuak kontuan hartuz, irizpide metodologiko hauek zehaztu ziren, lanaren bideragarritasuna, kalitatea eta balioa ziurtatzearren:

- **Hizkuntza-ingeniaritzaren ikerkuntza-esparrua jorratzea**. Oinarrizko tresnetatik abiatuta, oinarri sendoa osatzen joatea, etorkizunean helburu zabalagoetara heltzeko. Honek informatikari eta hizkuntzalarien arteko elkarlana eskatzen du eta uztartze horretan taldeko hizkuntzalarien ekarpena funtsezkoa izan da.
- **Aplikazioa**. Ekarpen teorikoak baztertu gabe ikerkuntza aplikatua da taldearen lanaren helburu nagusia. Hala ere, eta beste kasuetan gertatu den legez, hizkuntza berrien aplikazioan arazo berriak sortzen dira eta, hortik abiaturik, teoria eta ekarpen berriak ere.
- **Eskala erreala**. Erabakiak hartzerakoan maketen eta antzekoen erabilgarritasuna kontutan hartuz, arazo eta eskala errealeko tresnen eraikuntza da helburu nagusia.
- **Berrerabilgarritasuna**. Burutzen diren aplikazioak berrerabilgarriak izan daitezela zentzu bikoitzean: batetik, tesi-lanean garatutako tresnak erabiliz aplikazio konplexuago eta osotuagoak eraiki ahal izatea, eta bestetik, irekiak izatea aplikazio hauek beste erabiltzaileen esku jarriz. Horrez gain, aurretik egindako lana, taldean zein komunitate zientifikoan, berrerabiliko da ahal den neurrian.
- **Corpus idatzietan oinarrituta**. Tresnen baliagarritasuna hizkuntzaren erabilera errealarekin alderatuz neurtu behar da. Beraz, corpusak ezagumenduaren iturria izateaz gain, ebaluatzeko ezinbesteko baliabidea izango da.
- **Sendotasuna eta malgutasuna**. Lan honetan planteatzen diren helburuak orokorrak dira, beraz, ez dira testu-tipologiaren batekin edo gairen batekin lotuta. Horretarako malgutasuna eta egokitzapena eskaintzen duten osagai batzuk proposatzeaz gain, tresnen sendotasuna berebiziko garrantzia izango dute.

I.4 Tesiaren eskema eta argitalpenak

Txosten hau ondoren banan-banan azaltzen diren kapituluek osatzen dute. II. kapituluan lanaren kokapen orokorra egiten da, gainerako kapituluetan jorratzen diren gaien azalpen laburra eginez. Ondoren, lehenengo helburuari, MORFEUSen hobekuntzari alegia, aurre egiteko burututako lana III. eta IV. kapituluetan azaltzen da. III. kapituluan analizatzaile morfologikoaren hobekuntzak aurkezten dira, eraginkortasuna, estaldura eta doitasuna

areagotzeko helburua lortzeko burutu direnak. Baina estaldura osoa lortzeko, desbiderapen linguistikoen zein lexikoan agertzen ez diren hitzen kasuan, analizatzaileak hitz estandarretan baino interpretazio gehiago ematen ditu. Horregatik, IV. kapituluan aurkezten den hitz ez-estandarren tratamendua diseinatu da, analisi morfologikoaren zehaztasuna neurri batean hobetuz.

Bestalde, tratamendu automatikoan hitz anitzeko unitateak ere kontuan izan behar dira. Bigarren helburu gisa definitu den hitz anitzeko unitateen tratamenduari V. kapituluan eskaini zaio.

Behin unitate guztiak analizaturik, ezinbestekoa da bakoitzari interpretazio egokia ematea bere testuingurua kontuan hartuta. Horretarako, aipatu bezala, bi paradigma jorratu dira eta VI. kapituluan aurkezten da desanbiguazio morfosintaktikoan burututako lana.

VII. kapituluan tesi-lanean zehar aurkeztutako tresnen ebaluaziorako eta egiaztapenerako balio izan duten aplikazio batzuk aurkezten dira. Bukatzeko, VIII. kapituluan lanaren ondorioak eta etorkizunerako ikerlerroak aurkezten dira.

Sarrera hau amaitzeko I.1 eta I.2 taulek tesi honekin lotutako argitalpen nagusien berri ematen dute, bakoitza dagokion kapituluarekin lotuz.

Egileak	**Argitalpena**
Aduriz *et al*. 1995	Different Issues in the Design of a Lemmatizer/Tagger for Basque
Aduriz *et al*. 1996-a	EUSLEM: A Lemmatiser / Tagger for Basque
Aduriz *et al*. 1996-b	Del analizador morfológico al etiquetador/lematizador: unidades léxicas complejas y desambiguación
Aduriz *et al*. 1996-c	MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque
Ezeiza 1997	EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa
Ezeiza *et al*. 1998	Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages
Aizpurua *et al*. 2000	GaIn: un buscador Internet/Intranet avanzado para textos en euskera
Urizar *et al*. 2000	Morphosyntactic structure of terms in Basque for automatic terminology extraction
Alegria *et al*. 2001	Using Finite State Technology in Natural Language Processing of Basque
Alegria *et al*. 2002-a	Trabajos en el área de Recuperación de la Información del grupo IXA de la Universidad del País Vasco
Alegria *et al*. 2002-b	Robustness and customisation in an analyser/lemmatiser for Basque
Alegria *et al*. 2003 (*argitaratzear*)	Robustez y flexibilidad de un lematizador/etiquetador
Navas *et al*. 2002	Assigning Phrase Breaks Using CARTs for Basque TTS
López de Ipiña *et al*. 2002-a	Automatic morphological segmentation for continuous speech recognition of Basque
López de Ipiña *et al*. 2002-b	Morphological segmentation for speech processing in Basque

I.1 taula.- Tesiarekin lotutako argitalpen nagusiak.

Kapitulua	Argitalpena
II. Lanaren kokapena	Aduriz *et al*. 1995 Ezeiza 1997
III. Analizatzaile morfologikoaren doikuntza IV. Hitz ez-estandarren tratamenduaren hobekuntza	Alegria *et al*. 2001 Alegria *et al*. 2002-b
V. Hitz anitzeko unitateen tratamendua	Aduriz *et al*. 1996-ac
VI. Desanbiguazio morfosintaktikoa	Aduriz *et al*. 1996-b Ezeiza *et al*. 1998 Alegria *et al*. 2003 (*argitaratzear*)
VII. Lematizazio eta etiketatzearen aplikazioak	Aizpurua *et al*. 2000 Urizar *et al*. 2000 Alegria *et al*. 2002-a Navas *et al*. 2002 López de Ipiña *et al*. 2002-ab

I.2 taula.- Kapitulu bakoitzarekin lotutako argitalpen nagusiak[1].

[1] Argitalpen hauek IXA taldearen web-gunetik jaso daitezke: http://ixa.si.ehu.es.

II Lanaren kokapena

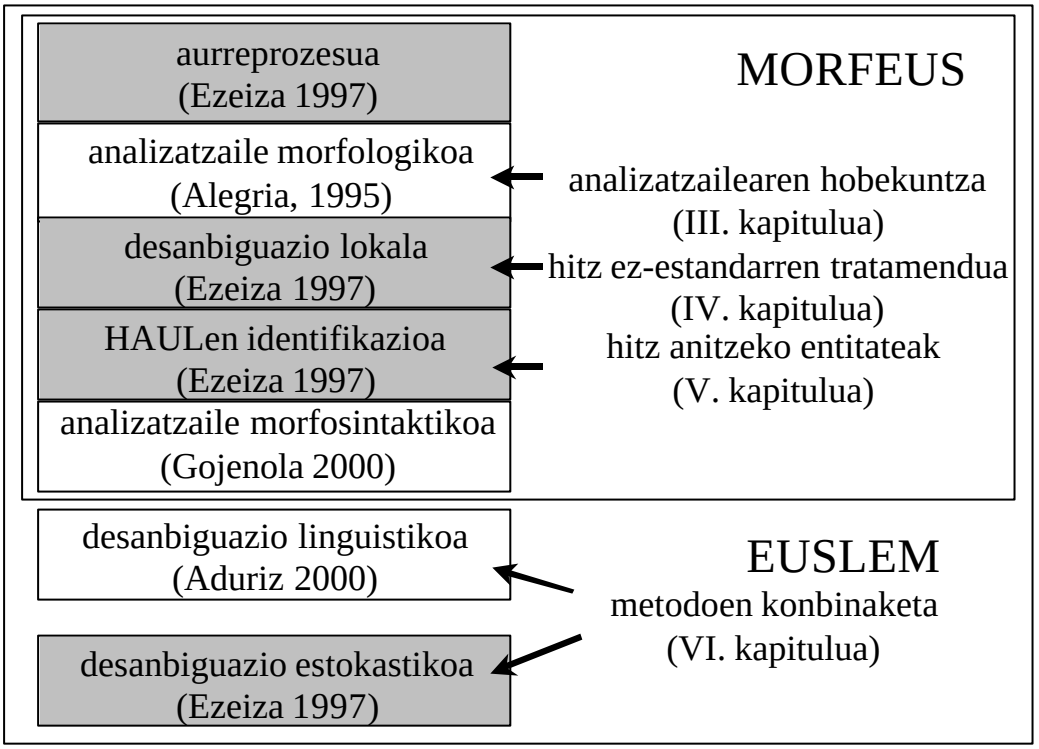
Tesi honen abiapuntua (Ezeiza 1997) tesina lanean deskribatzen den oinarrizko lematizatzaile/etiketatzailea da. Lematizatzaile/etiketatzaile batek sarrerako hitz-forma bakoitzeko, testuinguru horretan dagokion lema eta etiketa eman behar ditu. Etiketak ematen duen informazioa morfologikoa izango da, sintaxia eta semantika alde batera utziaz.

Aipatutako lanean, batetik, taldean garatutako analizatzaile morfologikoa (Alegria 1995) oinarritzat hartu eta analisi morfologiko sendoa osatzeko lehen urratsak eman ziren. Zehatzago esanda, token-ezagutzaile bat definitu, analizatzaileen arteko integrazioa egin, hitz ez-estandarren[1] anbiguotasuna mugatzeko lehen urratsak eman eta hitz anitzeko unitate lexikalen tratamendua diseinatu eta garatu ziren. Honekin batera, taldean beste lan batzuk burutu dira, hala nola, segmentatzaile morfologikoak emandako informazioa elaboratzeko analizatzaile morfosintaktikoa (Gojenola 2000) eta erregeletan oinarritutako desanbiguazio morfosintaktikoa (Aduriz 2000). Lan hauen guztien emaitzak MORFEUS analizatzaile morfosintaktikoa eta EUSLEM lematizatzaile/etiketatzailea dira, II.1 irudian azaldu bezala.

EUSLEMen lehenengo prototipoaren ebaluazioa egin zenean hainbat hobekuntzaren beharra ikusi zen. MORFEUSen hobekuntzei dagokionean, hasteko, analizatzaile morfologikoa LNParen beste hainbat tratamenduren oinarria izanik, bere eraginkortasuna hobetzea ezinbestekotzat jo zen. Horretarako, abiadura azkartu duen egoera finituko teknologian oinarritutako inplementazio berria burutu da (Alegria *et al.* 1997; Alegria *et al.* 2001). Bestalde, prozesuari berari erreparatuz gero, analizatzaile morfologikoaren emaitzetan hitz ez-estandarren anbiguotasun-neurri altuak EUSLEMen irteeran duen eragin negatiboa ikusita, hitz ez-estandarren tratamendua birplanteatu eta hedatu da. Gainera, hitz anitzeko

[1] Hitz ez-estandartzat aldaera dialektalak, gaitasun-desbideratzeak (ez-gaitasunak sortutakoak) eta hitz ezezagunak hartuko dira.

unitateen tratamendua beste unitate batzuetara hedatu eta tratamendua hobetu egin da. Hauen inguruan burututako lana tesi-lan honen III., IV. eta V. kapituluetan deskribatzen da.



II.1 irudia.- Aurrekariak eta lanaren kokapena.

Bestetik, desanbiguazio morfosintaktikoan, interpretazio guztien artean testuinguruari dagokiona aukeratzeko, teknika estokastikoak eta linguistikoak aplikatu ziren modu independentean eta ateratako ondorio garrantzitsuena honakoa izan zen: teknika estokastikoen kasuan, sarrerako anbiguotasuna handiegia dela teknika hauen bidez soilik emaitza egokiak lortzeko, eta, teknika linguistikoen bitartez emaitza egokiak lortu arren, ez zen testua erabat desanbiguatzen. Desanbiguazio linguistikoa erabilita burututako lanaren emaitzak Adurizen tesian (2000) aurkezten dira. Dena dela, lan hauen azken helburua teknika biak konbinatzea zen, ondorioz, integrazioa nola egin aztertu da eta VI. kapituluan lortutako emaitzak aurkezten dira.

Kapitulu honetan aldez aurretik egindako lana deskribatuko da. Lehenengoz, euskararen ezaugarri nagusiak azalduko dira eta euskararen prozesamendurako beharrezkoa den informazioa metatzeko erabiltzen den EDBL datu-basea deskribatuko da. Honekin batera, anbiguotasun morfosintaktikoari atal bat eskaintzen zaio, honek berebiziko garrantzia baitu emaitzetan. Gainera, anbiguotasuna neurtzeko erabiliko diren neurrien aurkezpena egingo da. Azkenik, MORFEUS eta EUSLEM tresnen aurkezpen orokorra egingo da atal banatan.

II.1 Euskara eta anbiguotasuna

Atal honetan aipatutako MORFEUS analizatzaile morfosintaktikoa garatzerakoan kontuan hartu diren euskararen ezaugarri nagusiak aipatuko dira, anbiguotasun morfosintaktikoa zertan datzan ulertu ahal izateko.

Euskararen kasuan morfologiaren eta sintaxiaren arteko lotura estua da. Izan ere, beste hizkuntza batzuetan sintaxi arloko fenomenoak direnak, euskararen kasuan hitzaren barruan gertatzen dira, aurrerago aipatuko den bezala. Horregatik, morfologiaz baino morfosintaxiaz hitz egin ohi da.

Lan honetarako adierazgarri diren euskararen ezaugarri morfosintaktiko nagusiak, honako puntu hauetan laburbil daitezke:

- Euskara hizkuntza eranskaria da, osagai guztiak elkarren segidan, independenteki eta deklinagaiak aldaketarik izan gabe metatzen dira ordena honetan: hiztegi-sarrera + determinatzailea + numeroa + kasua. Morfemak erantsiz sortzen dira sintagmak eta elementu bakoitzak bere informazioa mantentzen du. Informazio hori guztia formalki morfologikoa bada ere, sakonean morfosintaktikoa da, atzizkiek hitz-formaren funtzioaren berri ematen baitute.
- Lexiko sorkuntza aberatsa da euskaraz, eratorpena eta hitz-elkarketaren bidetik. Bi kasuotan lemaren kategoria alda daiteke, bereziki eratorpenean, eta kontuan hartu beharko da aplikazio konputazionalean.
- Hitz barruan izen-elipsia gerta daitekeela kontuan izan behar da azterketa morfosintaktikoa egiterakoan.
- Sintagma-buruak eskuineko posizioan jartzeko joera du. Hau da, mugatasuna, numeroa eta kasua sintagmaren azken hitzak hartzen du soilik. Horregatik, beste hizkuntzetan preposizioak direnak, euskaraz posposizioak dira, sintagmaren azken hitzean agertzen direlako.
- Euskara ordena libreko hizkuntza dela esaten da, esaldiko elementuek modu librean kokatzeko joera dutela, alegia. Hala ere, badago ordena neutrala deritzona, aditzaren aurretik doan elementua markatzen duena, eta gainerako ordenak informazio gehigarria emateko erabili ohi dira. Izen-sintagma mailan, berriz, elementuen ordena finkoagoa da.

Lehenengo hiru ezaugarriak segmentazio morfologikoa egiterakoan hartzen dira kontuan, hain zuzen ere, sarrerako hitza osagaietan banatu eta informazio morfologikoa esleitzerakoan. Aukera posible guztiak erabili behar izanagatik interpretazio kopurua asko handituko bada ere, ezaugarri guztiak aztertu behar dira testua morfosintaktikoki prozesatu nahi denean.

Sintagma barruko ordena zehatzak eta sintagmen arteko ordena libreak, berriz, garrantzi handia du morfosintaktikoki desanbiguatzerakoan. Izan ere, hitzaren testuinguruari erreparatu behar zaio, baina testuinguru hori noraino zabaldu erabakitzea zaila da. Gainera, desanbiguaziorako teknika estatistikoa erabiliz gero, testuingurua aurreko hitzera mugatzen da orokorrean, eta batzuetan interesgarriago litzateke ondokoaren erreferentzia ez-anbiguoa izatea.

Hurrengo atalean informazio guztia metatzen duen EDBL datu-basea aurkeztuko da, ondoren, lan honetan tratatuko den anbiguotasun morfosintaktikoaren arazoa zertan datzan azalduko da, eta honekin lotutako neurriak nola hartu ere deskribatuko da.

II.1.1 Euskararen Datu-Base Lexikala, EDBL

Euskararen Datu-Base Lexikala (EDBL) (Aldezabal *et al.* 1999-b) funtsezko ezagumendu-oinarria da euskararen prozesamendu automatikoaren arlo askotan, hala nola, analisi morfologikoa, sintaxia eta semantika. Eskala errealeko proiektu bati ekitean pentsaezina da dimentsio errealeko informazioa testu arruntetan edo fitxategi konbentzionaletan biltegiratzea, eta datu-basea da dudarik gabe dagokion errepresentazio-sistema.

Morfologia tratatzeko sortu bazen ere, EDBL gaur egun euskararen tratamendu automatikorako datu-base lexikal orokorra da. Horrexegatik, mota guztietako informazioa biltzen da bertan: morfologikoa eta sintaktikoa, eta semantikoa oraindik ez badago ere, homografo identifikatzaileak egoteak nolabaiteko hurbilpena adierazten du. Hala ere, momentuz inportanteena informazio lexikala da. Hirurogeita hamabost mila sarreratik gora ditu EDBLk gaur egun.

Informazio-iturriak anitz erabili dira EDBL landu eta eguneratzeko, besteak beste, Kintanaren *Hiztegia 80* (1984) eta *Hiztegia 2000* (2000), Sarasolaren *Hauta-Lanerako Euskal Hiztegia* (1984) eta *Euskal Hiztegia* (1996), *Elhuyar Hiztegia* (1996) elebiduna, Sarasolaren *Gaurko euskara idatziaren maiztasun-hiztegia* (1982), Etxebarria eta Mujikaren *Euskararen Oinarrizko Hiztegia, maiztasun eta prestasun azterketa* (1987) eta Euskaltzaindiaren *Hiztegi Batu*ko sarrerak eta arauak. Horretaz gain, *Euskaldunon Egunkaria*tik jasotako maiztasun handiko izen berezi eta siglak eta *Xuxen* zuzentzaile ortografikoaren erabileratik jasotakoak ere erabili dira datu-basea aberasteko.

Jakintza-arloen arteko talde-lana izatean, datu-basearen eguneratzea, zuzenketa eta mantentzea linguisten zeregina izan den bitartean, informatikariena izan da datu-basearen eta

interfazearen diseinua, esportaziorako prozeduren idazketa eta segurtasun- zein osotasun-egiaztapenerako murriztapenen definizioa.

Sarrera bakoitzari dagokion informazioa eremuetan biltzen da, eremu garrantzitsuenak honako hauek izanik:

- forma kanonikoa
- bi mailatako forma (morfologian erabiltzen den ereduari egokitua)
- itsats dakizkiokeen morfemei buruzko informazioa (jarraitze-klasea)
- homografo identifikatzailea
- iturburua
- iturburuko forma
- erabilpenaren adibidea(k)
- kategoria, azpikategoria eta aditz-mota
- flexioari buruzko informazioa: kasua, numeroa, mugatasuna, erlazioa, modu/denbora, pertsona
- kategoria erantsia
- funtzio sintaktikoak
- oharrak egiteko eremua
- maiztasuna
- eguneratze-data eta berau egin duen hizkuntzalaria
- arrarotasun eremua

Azken eremu horren beharra gehien bat zuzentzaile ortografikoaren emaitzak hobetzearren gehitu zen, zuzenketa proposamenak sailkatzeko baliagarri izan zedin. Aipatutako arrarotasun eremuan hiru balio desberdin erabiliko dira, hiru klasetako sarrera "arraroak" bereiztearren —esan gabe doa, bestelako hitzek eremu hau hutsik izango dutela—:

- *ANB*, analisietan sarrera-hitz bat halako kategoria batean arraroa dela adierazteko: esaterako, *hain* izenordain bezala interpreta daiteke, baina oso kasu gutxitan da hori interpretazio zuzena. Honela sailkatuko dira, halaber, kategoria bakarra izanik ere beste hitz baten flexio batekiko anbiguo suertatzen direnak: esate baterako *zela* izena, *zela* (*zen+la*) aditz-forma flexionatuarekiko anbiguo dena.
- *LEX* edo arrarotasun lexikala, hiztegietan "gutxi erabilia" oharraz horniturik dauden hitzak bereizteko erabiliko dena.
- *ABT* edo arrarotasun "absolutuak", *pa*, *no* eta *do* bezalako hitzen flexioen arrarotasuna markatzeko. Normalean hitz laburrak dira, eta zuzenketa-proposamenak ematerakoan zein analisi anbiguo batzuetan arazoak sortu ohi dituzte.

Bereziki aipatutako zuzentzaile ortografikoan pentsatuta gehitu bazen ere, III. kapituluan ikusiko denez, analisi morfologikoan aplikaziorik ere eman dakioke eremu honen informazioari.

Datu-basearen diseinu berrian (Aldezabal *et al*. 1999-b) forma sinpleak ez ezik hitz anitzeko unitateak ere kontuan hartu dira lexikoaren tratamenduan —hitz-elkarketa, kolokazioak eta lokuzioak besteak beste—. Hitz anitzeko unitateen tratamenduari eskaintzen zaion atalean —tesi-lan honetako V. kapituluan— aurkeztuko da unitate hauek lantzeko diseinatutako datu-basearen atala.

II.1.2 Anbiguotasun morfosintaktikoa

Anbiguotasuna lengoaia naturalaren prozesamenduaren maila guztietan agertzen da, bai maila morfologikoan baita maila sintaktiko, semantiko eta pragmatikoan ere, hizkuntza izatez baita anbiguo. Hori dela eta, lengoaia naturalaren prozesamenduan anbiguotasunaren ebazpena arazo nagusienetako bat da.

Lan hau maila morfosintaktikoan kokaturik dago eta anbiguotasun morfosintaktikoa aztertu eta ebaztea da helburu nagusia. Orokorrean hiru mota nagusi bereizten dira anbiguotasun morfosintaktikoaren barruan, ondoren azaltzen den bezala: kategoriari, morfema ez-askeei eta sintaxiari dagozkionak. Atal honen zati handiena hitzaren anbiguotasun morfosintaktikoari eskainiko zaio, hitz-formaren esparrukoari, alegia. Edozein kasutan, bukaeran sintaxi mailako anbiguotasuna ere laburki aipatuko da.

Anbiguotasun-tasa deskribapen linguistikoaren granularitatearekin batera aldatuz doa, baina analisi morfosintaktikoaren helburua beharrezkoak diren ezaugarri guztiak deskribatzea denez, desanbiguazio morfosintaktikoaren sarrera oso anbiguoa izango da, informazio morfosintaktiko guztia kontuan hartuz gero behinik behin.

Lehenengoz kategoria mailako anbiguotasuna kontsideratuko da, hau da, IZEN/ADITZ, ADITZ/ADJEKTIBO/ADBERBIO, etab. Anbiguotasun mota hau korapilotsuena da. Izan ere, anbiguotasun morfosintaktikoaren parterik inportanteena kategoria mailan aurki daiteke.

Euskaraz, definitu ditugun oinarrizko 20 etiketen arabera (Aduriz *et al*. 1995) —ikus A eranskina—, testu-hitzen %46-48 anbiguo dira, eta hitz anbiguo bakoitzak batez beste 2,3-2,4 interpretazio dauzka. Kategoria eta azpikategoria kontuan hartuta 45 etiketa definitu dira, eta testu-hitzen %50-55 dira anbiguo, horietako bakoitzak batez beste 2,5-2,6 interpretazio izanik. Ingeleserako, berriz, *Wall Street Journal* corpuseko milioi bat hitzetatik %34 anbiguo dira 45 etiketa dituen *Penn Treebank* etiketa-multzoaren arabera eta hitz anbiguo bakoitzak 2,4 interpretazio dauzka (Voutilainen 1994). Bestalde, gaztelerazko *LexEsp* corpuseko

5.500.000 hitzetako corpusean kategoria eta azpikategoria kontuan hartzen duen 62 etiketa erabilita hitz anbiguo bakoitzeko 2,63 interpretazio daude (Màrquez 1999).

Neurri hauei erreparatuz ikus daiteke kategoria mailako desanbiguazioa egitea gaitza dela, Karlsson-ek (Karlsson *et al.* (*eds.*) 1995:21) adierazi bezala:

> *"In English, categorial ambiguity (of parts of speech) is pervasive and one of the most serious problems facing anybody trying to construct a realistic and successful English parser"*

Baina orokorrean kategoria baino informazio gehiago behar izango da analisi morfologikoa oinarritzat hartzen duten aplikazioetan, hala nola, lematizazio/etiketatze morfologikoan, sintaxian, informazioaren erauzketa eta berreskurapenean, besteak beste. Hori dela eta, testu-hitz bakoitzari esleitzen zaizkion interpretazio morfosintaktiko posible guztiak kontsideratu behar dira, bai deklinabidea baita bestelako ezaugarri morfosintaktiko zein lexikalki ez-independenteak diren ezaugarriak, menpekoei dagozkien atzizkiak barne, interpretazioan sartuz. Informazio guztia kontuan hartuz gero, token bakoitzeko batez beste 3,2-4,2 analisi emango dira. Voutilainen-ek (1994) ingeleserako batez beste 2 interpretazio daudela dio[2], ondorioz, euskarak anbiguotasun handiagoa duela esan daiteke.

Horren arrazoietako bat, baina ez bakarra, lehen aipatutako preposizio falta da, testu hitz gutxiago izanik, bakoitzak bere baitan dituen morfema anbiguoen interpretazioen konbinaketa posible guztiak ematen baitira hitz bakarraren interpretazio gisa.

Beraz, euskararen kasuan, kategoria mailako anbiguotasunari atzizkien anbiguotasun morfosintaktikoa gehitzen zaio, kasu eta numero-mugatasun mailan aurkitzen dena, alegia. Esate baterako, *gizonak* hitzak bi irakurketa dauzka, IZEN-ABSOLUTIBO-PLURALA, objektu edo subjektu funtzioa bete dezakeena, eta IZEN-ERGATIBO-SINGULARRA, subjektu funtzioa bete dezakeena. Bi aukeren artean erabakitzeko, aditz mota eta esaldiko gainerako osagaien informazioa beharko da.

Beste adibide bat mendeko morfemen kasua da. Hauetan, lehendabizi erabaki behar da zein mendekori dagokion atzizkia eta, ondoren, mendekoari dagokion funtzio sintaktikoa zein den. Gai honen inguruko azterketa sakona Adurizen tesi-lanean (2000) aurki daiteke.

Adibideetan ikusten ahal da morfologia eta sintaxiak oso lotura estua dutela, anbiguotasun morfologikoa ebazteak askotan sintaxi mailako informazioa desanbiguatzea ere badakarrelako. Hala ere, hau ez da kasu guztietan gertatzen eta badira sintaxi mailan ebatzi behar diren anbiguotasunak. Aurreko adibidean, esate baterako, behin *gizonak*

[2] Corpus txiki baten gaineko neurriak ematen ditu Voutilainen-ek (van Halteren (*ed.*) 1999:241) eta hitzeko 1,8-1,9 inguru interpretazio daudela dio.

IZEN-ABSOLUTIBO-PLURALA dela erabaki denean, bere funtzio sintaktikoa objektu ala subjektu den erabaki beharko da[3].

II.1.3 Ebaluaziorako neurriak

Atal honen helburua tesi-lanean zehar agertuko diren neurrien berri ematea da. Izan ere, prozesu bakoitzaren emaitzak ebaluatzeko neurri desberdinak erabil daitezke. Bibliografian bi neurri-bikote erabiltzen dira gehien (van Halteren (*ed*.) 1999:81-82):

- *Correctness/Ambiguity* edo Zuzentasuna/Anbiguotasuna : lehenengo neurriak token guztietatik interpretazio zuzena zenbatek jasotzen duten adierazten du, eta bigarrenak, berriz, tokeneko batezbesteko interpretazio kopurua ematen du. Orokorrean, irteeran interpretazio bat baino gehiago geratzen direnean, bigarrena ere emango da.

$$Correctness = \frac{\text{etiketa zuzenen kopurua}}{\text{token kopurua}}$$

$$Ambiguity = \frac{\text{etiketa kopurua}}{\text{token kopurua}}$$

- *Recall/Precision* edo Zuzentasuna/Zehaztasuna: informazioaren berreskurapenean erabili ohi diren neurriak dira. Hala ere, bere erabilera beste arloetara ere hedatu da. Lehenengoak identifikatu beharreko interpretazio zuzen guztietatik zenbat izan diren identifikatuak eta bigarrenak eman diren interpretazio guztietatik ondo zenbat dauden neurtzen du. Bi neurri hauek portzentajeetan ematen dira.

$$Recall = \frac{\text{emandako interpretazio zuzen guztiak}}{\text{eman beharreko interpretazio zuzen guztiak}}$$

$$Precision = \frac{\text{emandako interpretazio zuzenak}}{\text{emandako interpretazio guztiak}}$$

Anbiguotasunak eta zehaztasunak desanbiguatzeko geratzen denaren neurria emango dute, zuzentasunak (*recall* eta *correctness*), berriz, galdu diren analisi zuzenen neurria emango dute. *Recall* eta *correctness* neurriak antzekoak dira baina ez berdinak. Lehenengoak hitzen berezko anbiguotasuna aurreikusten duen bitartean, bigarrenak ez. Hau da, hitz bat anbiguo

[3] Funtzio sintaktikoen anbiguotasuna tratatu den arren, lan honetan ez da kontuan izango.

izanik bi interpretazio posible baditu, lehenengoak biak agertzea eskatzen duen bitartean, bigarrenak nahikoa du bietako bat agertzea[4].

Recall/Precision neurri-bikotean bata handitzeak bestea jaistea dakarrela gertatu ohi da. Horrela, errazago da unitate guztietatik gutxi batzuk tratatuz gero oso emaitza onak ematea eta, alderantziz, guztiak tratatu nahi direnean, emaitzaren errore kopurua handiagoa gertatzea. Horregatik, bakoitza bere aldetik aztertuz gero, ez dute prozesuaren kalitatearen berri ematen.

Desanbiguazio-prozesuaren helburua token guztiak erabat desanbiguatzea denez, neurri biek balio bera emango lukete, beraz, bata zein bestea eman daiteke. Anbiguotasunik geratzen bada askotan *precision* edo zehaztasun neurria soilik eman ohi da. Hala ere, bestelako prozesuetan, informazioaren erauzketa eta berreskurapenean kasu, neurrietako bakarra emateak ez du informazio handirik ematen. Horregatik, helburua bien arteko oreka lortzea izaten da, ahalik eta gehien desanbiguatzea errore gutxien gehituta.

Bada oreka hori neurtzeko balio duen beste neurri bat, *f-score* deiturikoa. Honek zehaztasuna eta zuzentasuna erabilita, prozesuaren emaitza nolakoa den neurtzeko balio du. Bien artean oreka dagoenean, *f-score*k bi neurrien antzeko balioa hartzen du, baina neurrien artean alde handia dagoenean, bien erdiko balio bat ematen du, nolabait egin beharreko lana zein mailataraino bete den adieraziz.

$$f\text{-}score = \frac{2 * precision * recall}{precision + recall}$$

Tesi-lanean zehar aurkeztuko diren tauletan testuaren zenbait neurri ematen dira. Anbiguotasuna neurtzeko honako hiru datuok ematen dira:

- AR edo anbiguotasun-tasa, token edo testu-unitate guztietatik zenbatek duten interpretazio bat baino gehiago.
- I/A edo token anbiguo bakoitzak zenbat interpretazio dituen batez beste.
- I/T edo token bakoitzak zenbat interpretazio dituen batez beste, anbiguoak ez direnak ere kontuan hartuta.

Neurri hauekin erabateko desanbiguazioa burutzeko egin beharreko lanaren tamaina ikusi ahal izango da, %100 desanbiguatzeko tokeneko interpretazio bakarra utzi beharko baita — hau da, AR = 0, I/A=1, I/T=1.

[4] Lan honetan emaitzak lortzeko erabiltzen diren testuetan ez dago aipatutako berezko anbiguotasunik, horregatik, nahiz eta kontzeptualki *recall* eta *correctness* desberdinak izan, biak zuzentasun gisa itzuli dira.

Hauekin batera, zenbait kasutan metatutako erroreen tasa ere ematen da (AE), ordura arteko prozesu guztien artean eginikoak, alegia. Neurri honi erreparatuz jakin ahal da heuristiko bakoitzak sortzen duen errorea eta emaitzaren zuzentasunean (*correctness*) zenbaterainoko eragina duen, neurri osagarriak direlako (zuzentasuna = 100 – AE). Errore tasa ere portzentaje gisa ematen da.

$$\textit{Errore-tasa} = \frac{\text{etiketa okerren kopurua}}{\text{token kopurua}}$$

Azalduko den bezala, emaitzak ez dira parekoak izango hitza estandarra edo ez-estandarra bada, ez-estandarretan anbiguotasun handiagoa edota errore gehiago egongo baita. Ondorengo atalean azalduko denez, analisia egiterakoan lehenengoz hizkuntzaren arau estandarrei jarraitzen dieten edota lexikoan landuta dauden hitzak tratatuko dira, ondoren aldaera dialektal eta gaitasun-desbideratzeak eta azkenik, erabat ezezagunak diren hitzak. Horregatik, tokenak taldeka aztertuko dira, testu-hitzak hiru multzo hauetan banatuz:

- Estandar taldea, analizatzaile morfologikoaren lehen moduluak tratatzen dituenak.
- Aldaerak taldea, bigarren moduluak tratatzen dituenak.
- Ezezagunak taldea, aurreko bien bidez analizatu ezin izan direnak.

Hauek guztiek batera, sarrerako testu-hitzen multzoa eratzen dute, geratzen diren tokenak puntuazio-ikur eta bestelako bereizgarri eta identifikatzaileak direlarik. Hala ere, prozedura bakoitzari dagozkion emaitzak ematerakoan, helburu duten taldearen emaitzak emango dira soilik, gainerakoenak aldaketarik izango ez dutelako.

Neurri hauek guztiak testuari dagozkionak dira, hots, prozesua aurrera doan heinean anbiguotasuna jaisteko gehitzen doan errorea neurtzeko erabili direnak. Baina, horretaz gain, interesgarria gertatzen da prozesu bakoitzak burutzen duen lana neurtzea ere. Izan ere, anbiguotasuna gutxitzeko prozedura bat diseinatuz gero, ez da komeni neurririk gabeko errore kopurua gehitzea, ondorengo prozesuetarako ezin izango delako horiei dagokien informazio zuzena berreskuratu. Beraz, errore-tasa hori kontuan hartu da prozedurak integratu behar diren erabakitzerakoan.

Hortaz, desanbiguazio-prozesu bakoitzarekin batera, desanbiguazio-tasa eta prozeduraren errore-tasa ematen dira. Batetik, desanbiguazio-tasak sarreran soberan zeuden analisietatik[5] (*superfluous analyses*) zenbat baztertu diren, eta, bestetik, prozeduraren errore tasak baztertutako interpretazio horietatik zenbatetan huts egin duen adieraziko dute. Neurri hauek

[5] Sistemak hitzeko gutxienez interpretazio bat utziko duenez, soberan dauden analisiak kalkulatzeko analisi guztien kopuruari hitz kopurua kenduta kalkulatuko da, nahiz eta zenbait hitzek zuzena den interpretazioa izan ez.

askotan zuzentasun/zehaztasuna baino interesgarriagoak izaten dira sistemak erkatzeko, hasierako anbiguotasun-neurriak desberdinak izan arren, neurri hauen bidez egin beharreko lanaren zein proportzio eta lan hori nola egiten den neurtu daitekeelako.

$$\textit{Desanbiguazio-tasa} = \frac{\text{baztertutako aukerak}}{\text{soberako aukerak}}$$

$$\textit{Prozesuaren errore-tasa} = \frac{\text{oker baztertutako aukerak}}{\text{baztertutako aukerak}}$$

Desanbiguazio-prozesuaren hasierako egoera analizatzaile morfologikoaren irteerarena izango da. II.1 taulan ikusten denez, anbiguotasuna oso handia da, baina aldi berean emaitzak zuzentasun maila oso altua du. Lehenengo zutabean tokenen banaketa taldeka nola gauzatzen den ematen da (DT). Sarrerako hitz gehienak estandar gisa tratatu dira, baina III. kapituluan ikusiko denez, banaketa hau aldatuko da testu-motaren arabera.

Testu-hitzez gain, puntuazio-ikurrak eta bestelako identifikatzaileak ere agertzen dira testuan. Corpus honetan bestelako tokenen multzoa testu arruntetan baino handiagoa da[6], EEBS (Urkia eta Sagarna 1991) corpuseko identifikatzaileak agertzen direlako. Identifikatzaile horien bitartez, testuaren euskalkia, zein garaitakoa den, estiloa eta beste zenbait ezaugarri kodetzen dira. Azken hauek ez dute tratamendu morfologikoaren beharrik eta horregatik ez dira testu-hitzekin batera tratatzen.

II.1 taularen (a) atalean EEBS corpus orekatuko 27.000 inguru tokenek eta *Euskaldunon Egunkaria*ko 9.000 inguru tokenek osatzen duten erreferentzia-corpusaren anbiguotasun neurriak azaltzen dira. Taularen (b) atalean, berriz, prozesuen egokitasuna egiaztatzeko erabili den EEBSko 1.300 token inguru eta *Euskaldunon Egunkaria*ko 5.800 token inguruko corpusarenak aurkezten dira.

Egiaztapenerako corpus honetarako gehienbat *Euskaldunon Egunkaria*ko testuak aukeratu dira hiru arrazoi nagusi hauengatik:

- Gaur egungo euskararen erabilera estandarra islatzen dute. EEBS corpusean, corpus orekatua izanik, euskara batuaz zein euskalkietan idatzitako testuak daude eta, euskara batuaz idatzitakoak garai desberdinetakoak izanik, erabilera ez-estandar asko dituzte.

6 Orokorrean puntuazio-ikur eta bestelako bereizgarriek tokenen %15 inguru biltzen dute. Corpus honetan, aldiz, %20 inguru dira (ikus III.2 taula).

*Euskaldunon Egunkaria*koak, aldiz, estandartzat jo daitezke, tarteka erabilera dialektalak ere txertatzen diren arren.

- Izen berezi asko agertzen dira, gehienak erdaretakoak. Hitz hauek ezezagun gisa analizatuko dira gehienetan eta interpretazio kopurua bestelako hitz arruntena baino handiagoa izango da. Horien anbiguotasuna modu egokian murrizteak berebiziko garrantzia izango du hurrengo urratsetako emaitzetan. Horregatik diseinatu da izen berezietarako desanbiguazio-prozedura bat eta horren ebaluazio egokia egiteko, izen berezi asko duen corpusa behar da.
- Erabilera ez-estandar gutxiago dituzte. Aldaeren kopuruari erreparatuz gero, erreferentzia-corpusarekin alderatuz gero, erdia baino gutxiago dira. Gainera, gehienak EEBSko zatiari dagozkio.

(a)	**DT**	**AR**	**I/A**	**I/T**	**R**	**P**	**F**
estandar	%77,91	%80,72	3,81	3,27	**%99,73**	%30,53	46,75
aldaerak	%1,74	%81,83	4,47	3,84	**%92,26**	%24,05	38,16
ezezagunak	%2,65	%100	18,09	18,09	**%98,33**	%5,44	10,30
testu-hitzak	%82,30	%81,37	4,39	3,75	**%99,52**	%26,50	41,86
batez beste	%100	%66,96	4,39	3,27	**%99,61**	%30,49	46,68
(b)							
estandar	%78,76	%81,13	3,82	3,29	**%99,75**	%30,30	46,49
aldaerak	%0,70	%74,00	4,14	3,32	**%70,00**	%21,08	32,41
ezezagunak	%3,04	%100	19,42	19,42	**%99,54**	%5,13	9,75
testu-hitzak	%82,50	%81,76	4,53	3,89	**%99,51**	%25,61	40,73
batez beste	%100	%67,45	4,53	3,38	**%99,60**	%29,46	45,47

II.1 taula.- Anbiguotasun neurriak analizatzaile morfologikoaren irteeran[7].

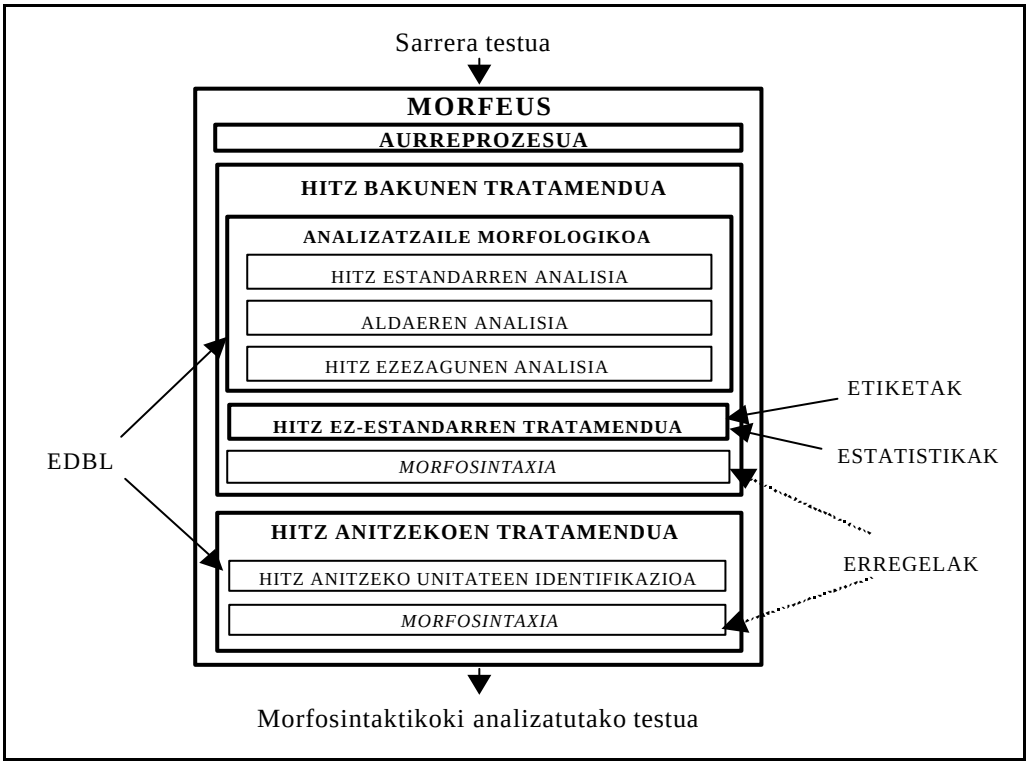
Hurrengo atalean, MORFEUSen deskribapena egin eta lanaren abiapuntua zein den aztertuko da.

[7] DT = Tokenen banaketa talde bakoitzean (*Distribution of Tokens*)
AR = Anbiguotasun-tasa (*Ambiguity Rate*)
I/A = Anbiguoen batezbesteko interpretazio kopurua (*Interpretation per Ambiguous token*)
I/T = Tokenen batezbesteko interp retazio kopurua (*Interpretation per Token*)
R = Zuzentasuna (*Recall*)
P = Zehaztasuna (*Precision*)
F = *f-score*

II.2 MORFEUS, analisi morfosintaktikoa

Sarrerako testua morfosintaktikoki analizatzeko erabiltzen da MORFEUS. Hori lortzeko hiru modulu nagusi ditu. Lehenengo moduluak, aurreprozesua deiturikoak, sarrerako testua unitate edo tokenetan banatzen du eta analizatzaile morfologikoari bidali behar zaizkionak bereizten ditu, hau da, testu-hitzak direnak. Bigarren moduluak, berriz, hitz bakunak morfosintaktikoki tratatzen ditu. Eta, hirugarrenak, hitz anitzeko unitateen prozesamendua egiten du.

"Hitz" kontzeptuaren definizio zehatza ematea ez da inondik ere zeregin erraza. Testu mailan hitza "zuriune biren arteko karaktere-kate" gisa definitzea litzateke irtenbiderik errazena (Fontenelle *et al*. 1994). Badira, noski, irizpide honekin bat datozen unitate lexikalak (*etxe*, *zuri*, *zakur*). Are gehiago, euskaraz, oso hizkuntza flexiboa izanik, beste zenbait hizkuntzetan hitz anitzeko liratekeen hainbat lokuziok, unitate tipografiko bakarra osatzen dute (*ziurrenik*, *'most probably'*; *aurrerantzean*, *'from now on'*; *aurretiaz*, *'in advance'*). Hala ere, unitate tipografikoa ezagutzea ez da berehalako lana. Funtzio hori tokenizatzaileak betetzen du eta aurrerago azalduko da zeintzuk izango diren MORFEUSen unitate tipografiko kontsideratu direnak.



II.2 irudia.- MORFEUSen egitura.

Baina, esan gabe doa, guztiz bestelako definizioa behar da lokuzioen edo osterantzeko hitz anitzeko unitate lexikalen kasuetarako (*an egin*, *hala eta guztiz ere*, *hutsaren hurrengoa*,

mila eta bostehun edo *2000ko maiatzaren 1ean*). Horrelakoak hitz anitzeko unitate gisa tratatu behar dira.

Ondorengo ataletan modulu bakoitzaren funtzioak azalduko dira, tesi-lan honen zenbait atal MORFEUSen kokatuz.

II.2.1 Aurreprozesua

Zuriune batez bereiziko hitz-elkarketa, lokuzioak eta orokorrean hitz anitzeko terminoak ez dira zuzenean morfologikoki analizatzen, hitz anitzeko unitateekin batera tratatuko direlako; beraz, analisirako tratamendu-unitatea hitza da, baina, esan bezala, formatua bakarrik kontuan hartzen bada hitza mugatzea ez da hain prozesu erraza.

Askotan tokenizazio-prozesuari ez zaio behar bezalako garrantzia ematen eta gaiaren inguruko eztabaida gutxi planteatzen da. Hala ere, Grefenstette-k (van Halteren (*ed.*) 1999:117) aipatzen duen bezala, tokenizatzaile baten eginkizunak definitzea ez da berehalako lana:

> *"Though rarely discussed, and quickly dismissed, tokenization in an automated text processing system poses a number of thorny questions, few of which have completely perfect answers."*

Kontuan izan behar da tokenizazioak gainerako LNPko prozesu asko elikatzen dituela, besteak beste, analisi morfologikoa, etiketatzea eta analisi sintaktikoa, eta lehen urrats honen doitasuna datu garrantzitsua izango da sistema osoaren ebaluazioan. Izan ere, tokena gaizki identifikatuz gero, ondorengo urratsetan tratatua izan daitekeen arren —tresna sendoak direnean behintzat—, tratamendu horren emaitzak inongo baliorik ez du izango.

Adibide baten bidez ikusteko, testuan *http://ixa.si.ehu.es* agertzen denean, tokena bere osotasunean eman beharrean hamaika token[8] ematen badira, analisi morfologikoak emandako interpretazioak ez dira baliagarriak izango.

Beraz, tokenizatzaile bat eraikitzea ez da berehalako lana. Tokenizatzaile sinple bat egiteko, berriz, ez da baliabide handirik erabili behar. UNIXeko *sed*, *awk* edota *lex* tresnak erabilita erraza da zuriuneen bidez banatuta dauden unitateak bereiztea. Hala ere, puntuazio ikurrak eta bestelako bereizgarriak ere kontuan izan behar dira. Aipatutako Grefenstetteren lanean hainbat adibide ematen dira tokenizatzaile sinpleenetik hasita nola hobe daitekeen erakusteko. Orokorrean, tokenizatzailean erabilitako baliabideak handitzen diren heinean,

8 Bestelako unitateak tratatzen diren moduan eginez gero honako tokenak emango lirateke: *'http'*, *':'*, *'/'*, *'/'*, *'ixa'*, *'.'*, *'si'*, *'.'*, *'ehu'*, *'.'* eta *'es'*.

tresnaren kalitatea ere areagotu egiten da, eta token-mota asko ezagutu nahi direnean ezinbestekoa da esfortzua handitzea.

Eman lezakeena baino lan neketsuagoa da euskararako halako ezagutzailea egitea, elementu batzuek —marra edo puntua adibidez— funtzio anitz dutelako. Puntuaren funtzioen artean aukeratzea bereziki problematikoa da, esaldi bukaera markatzeko balio duelako. Aurrerago atal bat eskaintzen zaio puntuaren anbiguotasunari.

Bestalde, beste hizkuntzetan formatuaren bidez oso erraz identifikatu eta tratatzen diren osagai batzuk tratamendurako prestatzea euskaraz ez da hain erraza, deklina daitezkeelako. Horien artean zenbakiak eta zenbakiz idatzitako datak (*1993.eko, %8,4ko*) daude. Horrelako tokenek bere baitan puntuazio-ikur eta bereizgarriak edota maiuskulaz zein minuskulaz idatzitako hizkiak izan ditzakete eta aukera guztiak aurreikusi behar dira tokenak ondo ezagutu nahi badira. Gainera, horiek guztiak egoki interpretatu behar dira analizatzaileari prozesatu beharrekoa emateko[9] (*1993.eko = mila, bederatzi ehun eta laurogeita hamahirugarreneko*). Baina, esan bezala, tokena ezagutzeaz gain, analisi morfologikoaren sarrera zein izan behar den ere sortu behar da, eta ondorengo urratsetan baliagarria izan daitekeen informazioa ere mantendu behar da. Erabilgarria izan daitekeenaren artean, informazio tipografikoa dago. Zenbait etiketa definitu dira modu berezian idatzita dauden hitzak markatzeko. Adibidez: HAS_MAI, hasiera maiuskulaz dutenetarako; DEN_MAI osorik maiuskulaz idatzita daudenetarako; SIGLAK puntuen bidez idatzirik dauden hitzetarako; ZEN_DEK zenbaki deklinatuetarako.

Azkenik, ezagututako token guztiak ez dira morfologikoki analizatuko, puntuazio-ikurrak esaterako, baina horiei etiketa bat esleitzen zaie nolako tokenak diren jakiteko. Puntuazio-ikurrak bereziki interesgarriak dira esaldiaren mugak ezarri ahal izateko. Hori dela eta, mugatzeko balio duten zeinuek etiketa bereziak izango dituzte[10], baina bestelako bereizgarriek etiketa orokor bat jasoko dute[11].

Prozesu honek tokenak ezagutu, hurrengo urratsetarako beharrezkoa den informazioa tipografikoa jaso eta tokena itzultzen duelako, tokenizatzaile ordez, aurreprozesu deitu zaio prozesu honi.

9 Analizatzaileak ez ditu zenbakiak zuzenean analizatzen. Itzuli behar dira, baina zenbaki osoa eman beharrean, zenbakiaren amaierari erreparatzen zaio. Kasu honetan *1993.eko* tokena *hamahirugarreneko* bihurtzen da.

10 PUNT_PUNTU punturako; PUNT_KOMA komarako, etab.

11 Gainerako bereizgarrien artean parentesia, komatxoak, etab. izango dira eta BEREIZ etiketa jasoko dute.

II.2.1.1 Puntuaren anbiguotasuna

Arestian aipatu bezala, puntuak funtzio anitz izan dezake. Horietan garrantzitsuenetakoa esaldiaren muga markatzea da, baina ez da bakarra eta zenbait kasutan funtzio bat baino gehiago bete dezake aldi berean. Horren adibidea laburdura bat agertzea da, ondorengo adibideetan bezala[12].

(1) "(…) beste zenbait sindrome ere; hala nola: kontrola-ezinezko berbalizazioa, desegituratutako sintaxia, ***etab.****"*

(2) "Edozein herritan abesbatza, musika-banda, txistulariak, soinujoleak, musika-jaialdiak ***etab.*** *aurkituko ditugu (...)"*

Horrelakoetan, prozesua tokena ezagutzen ari denean, erabaki beharko luke ea puntua hitzarekin batera doan ala ez, baina analisi morfologikoa egin aurretik ezin izango da jakin laburdura denentz. Dena dela, analisia jasotakoan laburdura dela jakinik puntuarekin bildu ahal izango da, puntua laburduraren parte delako, baina lehenengo adibidean esaldi bukaera ere adierazten duenez, puntuak bi funtzio beteko lituzke.

Hirugarren adibidean, aldiz, puntuaren ostean zuriunerik ez dagoenez, laburdura deklinatua dela jakin daiteke tokenizazio fasean bertan analisirik egin gabe ere, beti ere testuan akatsik ez dagoen bitartean.

(3) "Irakasleen arteko lan-mintegi, berriztapen-proiektu, ***etab.etarako****: Pedagogi Berrikuntzarako zuzendariak emandako egiaztagiriak."*

Antzeko arazoa suertatzen da zenbaki ordinalekin baina hurrengo urratsetan duen eragina askoz ere handiagoa da. Izan ere, puntuak *–garren* atzizki lexikala ordezkatzen du eta azpikategoria eratorriak determinatzaile zehaztugabetik determinatzaile ordinalera aldatzen du. Ondoko adibidean ikus daiteke puntua zenbakiarekin batera analizaturik lortzen den emaitza eta azken lerroan beltzez markaturik puntua banaturik eman behar lukeena.

(4) "Ikus: Euskal Autonomi Elkarteaz kanpoko euskal gizatalde eta etxeekiko harremanetarakoari buruzko legea. Vitoria-Gasteiz, Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia, ***1994.****"*

Beraz, zenbakiak hiru aldiz anbiguo bihurtzen dira esaldi amaieran: determinatzaile ordinala eta puntua, determinatzaile ordinala soilik edota determinatzaile zehaztugabea eta puntua.

12 Atal honetako adibide guztiak EEBS corpusekoak dira.

```
/<1994.>/<ZEN_DEK>/
 ("1994."  DET DZH NUMP + ATZ DET ORD + DEK ABS MG @OBJ @SUBJ @PRED)
 ("1994."  DET DZH NUMP + ATZ DET ORD)
 ("1994"   DET DZH NUMP + DEK ABS MG @OBJ @SUBJ @PRED)
```

Pentsa liteke hurrengo hitza maiuskulaz hasten den ala ez konprobatzea nahikoa dela puntua bereizi behar den jakiteko, (5) adibidean bezala, baina (6) kontradibidea da, kasu honetan determinatzaile ordinala da eta ez da esaldi amaierarik puntu horretan:

*(5) "Lehenengoa, **1993.** urtean argitara emandakoa alegia, mende hasieratik **1992.** urterarte argitaratutako bibliografiaren azalpena zen; oraingoak, berriz, 1993 eta 1994koak ematen dizkigu jakitera."*

*(6) "Uztailaren 16ko 21/1992 Industriako Legea (uztailaren 23ko **176.** Estatuko Aldizkari Ofiziala).*

Azkenik, (7) adibidean, ez dago argi zein den zenbakiei eman behar zaien interpretazioa. Izan ere, puntuaren ordez marratxoa ere jar daiteke horrelako adibide askotan, zerrenda zenbatu bat besterik ez delako. Kasu horretan ez litzateke ordinal gisa interpretatuko, baina puntua jarriz gero bi irakurketak onargarriak dira. Edozein kasutan, puntuak muga bat bereizten du eta ordinalaren irakurketa eginez gero puntuak bi funtzio izango lituzke.

*(7) "**1.** Ikuspuntu praktiko batetik begiraturik, gaur egun oso eskasean baino betetzen ez den gizarte-funtzio bat segurtaturik kausituko litzateke (…) **2**. Lanbide antropologikoari irtenbideak irekiko litzaizkioke."*

Dena dela, laburduretan bezala, zenbakietan ere badira tokenizazio fasean zuzenki tratatu daitezkeen kasuak, zenbaki deklinatuena kasu, puntuaren ostean zuriunerik gabe deklinabide-atzizkia datorrelako.

*(8) "Unamunoren obra hau, jatorriz gaztelaniaz idatzia izan arren, lehenbiziko aldiz frantsesez argitaratu zen Parisen **1925.ean**, bilbotar filosofoa Frantziako hiriburuan erbesteraturik zegoenean."*

Baina tokenizazioa horrela egiteak formatu akatsak daudenean tokenizazio akatsak ekarriko ditu, ez baita hain arraroa ere testuetan ordinalaren ostean zuriunerik utzi gabe hurrengo hitza agertzea:

*(9) "Halaber, Pertsona Fisikoen Errentaren gaineko Zergari buruzko abenduaren 27ko 13/1991 Foru Araueko **37.bi** artikuluan aurrikusitakoaren arabera, herri-aurrezki planak gauzatzeko onesten diren eragiketak ez dira aurreko parrafoan adierazitako mugapenera loturik egongo."*

*(10) "(..) ikus RODRIGUEZ, F. (1992): RAP, Trikitrixa zein bertsolaritza Georges Lapassaderen esku. Bertsolari, **7.zkia.** 12-18. orr."*

Puntuaren anbiguotasuna ebazteko erregela-multzo bat defini daiteke, gorago aipatu den hurrengo hitzaren grafia kontuan izanik, esaterako. Hala ere, puntuaren aurretik doan tokena laburdura dela jakitea komeni da puntuak funtzio bat baino gehiago betetzen duen jakiteko eta, hala bada, puntua bera bikoizteko —lehenengo puntua laburdurarekin batera token bat osatzeko eta bigarrena esaldi amaiera adierazteko.

Horrek gurpil zoroan sartzen du puntuaren anbiguotasuna ebaztea. Esate baterako, puntuaren ostean maiuskulaz idatzirik dagoen hitza izen berezia izan daiteke askotan, eta esaldi hasieran badago eta lexikoan landuta ez badago, beste aukera batzuk ere aurreikusi beharko lirateke analisi morfologikoa egiterakoan —edota etiketa-multzoa esleitzerakoan analisirik egiten ez denean. Horrek hitzaren desanbiguazioa zailago bihurtuko luke. Baina puntua laburduraren zatia balitz eta esaldia bertan amaituko ez balitz, maiuskulaz idatzitako hitz hori ia seguru izen berezi bat izango litzateke. Hortaz, ondo tokenizatu ahal izateko esaldia desanbiguatuta egoteak lagun dezake, baina, aldi berean, puntua ondo desanbiguatuta egoteak ondorengo hitza desanbiguatzen lagun dezake.

Gurpil zoro hori apurtzeko bi hurbilpen egin daitezke: lehenengoan puntuaren anbiguotasuna desanbiguazio-prozesuan ebaztea eta bigarrenean tokenizazio fasean bertan ebaztea.

Lehenengo multzoko hurbilpen bakarra aurkitu da bibliografian. Mikheev-ek (2000-b) puntua token bereizi gisa ematen du beti tokenizatzerakoan, puntua bikoiztu gabe. Ondoren, etiketatzean tokenen ezaugarriak erabiltzen ditu puntuaren funtzioak adierazteko. Egileak XML lengoaia baliatuz honako adierazpidea proposatzen du: puntu baten aurrean dauden tokenei atributu berri bat gehitzea laburdura direnentz jakiteko —*A* atributua, balio boolearra duena[13]— eta puntuaren etiketa ematerakoan —*C* atributuan ematen dena— hiru balio[14] posible izatea, ondorengo adibideetan ikusten den moduan kodetuta:

```
...<W C='RB' A='N'>soon</W> <W C='.'>.</W> ...
...<W C='NNP' A='Y'>Mr</W> <W C='A'>.</W> <W C='NNP'>Brown< /W> ...
...<W C=','>,</W> <W C='NNP' A='Y'>Tex</W> <W C='*'>.</W> ...
```

Lehenengo adibidean esaldi amaierako puntua besterik ez da, bigarrenean laburduraren puntua izango da eta hirugarrenean bi funtzioak betetzen ditu. Horrela, kodeketaren bidez eman daiteke funtzioen berri tokena bikoiztu gabe eta analisi morfologikoaren —edota desanbiguazioaren— eta tokenizazioaren arteko dependentzia apur daiteke. Beraz, puntuaren anbiguotasuna, gainerako anbiguotasunekin batera ebatz daiteke desanbiguazio-prozesuan aldaketa handirik egin gabe eta tokenizazioan informazio lexikala gehitu gabe.

Desanbiguazio-algoritmoak burutzen dituen aldaketak aplikazio-esparruari dagozkio. Orokorrean tratamendua esaldika aplikatu ohi da, baina, kasu honetan, esaldia non amaitu den

13 Puntu aurreko tokenaren *A* atributuaren balioak 'Y' (laburdura) eta 'N' (ez laburdura) izango dira.

14 Puntuaren *C* atributuaren balioak '.' (esaldi bukaerako puntua), 'A' (laburduraren puntua) eta '*' (bai laburduraren puntua baita esaldi bukaerako puntua) izango dira.

oraindik ezaguna ez denean, tratamendurako unitate izango den hitz-sekuentzia beste irizpide bati jarraituz aukeratuko da[15].

Hurbilpen honen emaitzak onak izan daitezen, laburdurak eta esaldi hasierako hitzak ondo desanbiguatzea oso garrantzitsua da —puntuaren inguruko hitzak, alegia— lexikoan agertzen ez badira ere. Horretarako, laburdura edota izen bereziak identifikatzeko *Document Centered Approach* deituriko hurbilpena aplikatzen du (Mikheev 1999, 2000-a). Modu horretan dokumentuan puntu aurretik dagoen hitza beste posizio batean punturik gabe agertzen bada seguruenik ez da laburdura izango, baina beti puntu baten aurretik agertzen bada, orduan laburdura izan daiteke dokumentu horretan. Puntu ondoren agertzen diren hitzetarako antzera aplikatzen da, hitza dokumentuaren beste posizioetan minuskulaz idatzita agertzen bada seguruenik hitz arrunta izango da, baina beti maiuskulaz aurkitzen bada, orduan izen berezia izango da. Tratamendu hauek guztiak desanbiguazio-prozesuan integratuz, esaldi bukaeren identifikazioan %99,7 —*Wall Street Journal* corpusean— eta %99,8 —*Brown Corpus* erabilita— ondo tratatzea lortzen da[16].

Bigarren hurbilpen-multzoan bibliografian aurkitutako ebazpide gehienak sartzen dira. Gehienetan tokenizazio fasean erabaki ahal izateko maiztasun handiko laburduren zerrenda erabiltzen da puntua laburdura baten parte den ala ez erabakitzeko. Horren desabantaila nagusiena zerrendaren luzeran datza, izan ere, corpuseko —edota lexikoko— laburdura guztiak erabiliz gero, eraginkortasunean galtzen da, baina maiztasun handienekoak bakarrik erabiliz gero, gainerako laburduren berri izan gabe akatsak ere egin daitezke. Dena dela, lexikoan ez dauden laburdurak ere ager daitezke, beraz, tokenizazio mailan ezin izango da guztien berri izan.

Aurreko hitzaren informazioaz gain, bestelako ezaugarriak ere kontuan izan behar dira puntu bat esaldi amaiera adierazten duen ala ez erabakitzeko. Arestian esan bezala, irtenbiderik sinpleena hurrengo hitzaren grafiari erreparatzea da. Maiuskulaz idatzirik badago, orduan esaldi amaiera gisa markatu daiteke puntua eta bestela ez. Baina aurreko hurbilpenean ikusi den bezala, izen berezi bat izan daiteke. Beraz, erregela-multzo hori hedatuz puntuaren funtzioa modu zuzenago batean erabakitzea lor daiteke.

Adierazpen erregularren bidez erregela-multzo konplexua definitu eta laburduren zerrendak erabiliz gero puntuaren anbiguotasuna nahiko zuzenki ebatz daiteke. *Alembic* sisteman (Aberdeen *et al*. 1995) 100 bat adierazpen erregularren bitartez eta 70 bat laburdura erabilita nahiko emaitza onak lortzen dira —%99,1 *Wall Street Journal* corpusean. Hala ere,

15 Mikheevek aipatzen duenez, 30-40 luzerako hitz-sekuentziak erabil daitezke, nahikoa da bigrametan oinarritutako desanbiguazioan azken hitza eta trigrametan oinarritutakoetan azken bi hitzak ez-anbiguo izatea.

16 Emaitzetan '.' eta '*' etiketen arteko nahasketak ez dira akats kontsideratu, biek esaldi amaiera adierazi nahi dutelako.

kasu hauetan definizioa eskuz egin behar izaten da, eta orokorrean erreferentzia gisa erabilitako corpusari egokitutakoak izaten dira.

Gainerako prozesuetan gertatu ohi den bezala, erregelak eskuz idaztea neketsuago denez, esaldi amaiera desanbiguatzeko informazioa ere modu automatikoan erauztea da soluziorik egokiena. Horretarako teknika desberdinak erabili izan dira, hala nola erabaki-zuhaitzak (Riley 1989; Palmer eta Hearst 1997), sare neuronalak (Palmer eta Hearst 1997), eredu estatistiko elaboratuak —*maximum entropy*— (Reynar eta Ratnaparkhi 1997) edota datu estatistiko gordinagoak (Schmid 2000).

Riley-k (1989) erabaki-zuhaitzen informazioa lortzeko eskuz markatutako 25 milioi hitzeko corpusa erabili zuen. Aurreko eta ondoko hitzei buruzko ezaugarriak hartzen zituen kontuan eta %99,8 ondo identifikatzea lortu zuen —*Brown Corpus* erabilita. Hala ere, zuhaitzak eratzeko erabilitako corpusa itzela zen, eta hizkuntza gehienetan ez dago horrelako baliabiderik erabiltzerik. Hori dela eta, ondoren burututako hurbilpenetan erabili beharreko baliabideak minimizatzen saiatu dira.

Palmer eta Hearst-ek (1997) ere erabaki-zuhaitzen bidezko hurbilpena aurkeztu dute, zuhaitzean aurreko eta ondoko hitzei buruzko ezaugarriak ematen direlarik. Zuhaitzak eratzeko 622 adibide erabilita %98,4 lortu dute *Wall Street Journal* corpuseko 19.000 adibideetan. Zuhaitzen ikasketarako adibide kopurua 6.300era handituz %99 lortzen dutela diote. Sare neuronalen bidez, berriz, testuinguruaren luzera anitzekin saiatu ondoren, 6ko testuingurua aukeratuta —aurreko eta ondoko 3 hitzak, alegia— %98,5 zuzenki identifikatzea lortzen dute.

Testuinguruko hitzei dagozkien ezaugarriak hiztegitik lortzen dira, bai erabaki-zuhaitzetarako baita sare neuronaletarako ere. Horrela, hitz bakoitzari dagozkion etiketa guztiak lortzen dira. Testua desanbiguatuta ez dagoenez, etiketa-multzoak erabiltzen dira puntuari buruz erabakitzeko, baina etiketa bakoitzari buruzko maiztasunak ere kontuan hartzen dira. Horrek esan nahi du lexikoaz gain hitz bakoitzaren etiketen banaketa estatistikoa ere ezagutu behar dela. Dena dela, informazio hori ikasketarako corpusetik erauz daiteke. Hiztegiaren tamaina desberdinekin eta ikasketarako corpusaren tamaina aldatuz lortutako emaitzarik onena %99koa da ingeleserako eta alemanerako eta %99,6koa frantseserako, eta txarrenak %98 ingurukoak dira.

Reynar eta Ratnaparkhi-k (1997) desanbiguazio morfosintaktikorako erabiltzen duten eredu berbera proposatzen dute —*Maximum Entropy Approach*— puntuaren funtzioa erabakitzeko. Kasu honetan ere aurreko eta ondoko hitzen ezaugarriak erabiltzen dituzte eta emaitzarik onenak %98,8koa —*Wall Street Journal*eko 20.000 esaldiekin— eta %97,9koa —*Brown Corpus*eko 52.000 esaldiekin— dira. Emaitza hauek ez dira aurrekoak bezain onak,

baina puntua desanbiguatuta duen esaldi-multzo bat besterik ez dute behar datu estatistikoak lortzeko. Egileen arabera, sistema hau oso erraz egokitu daiteke beste hizkuntzetara eskuzko lan minimoa eginik eta lexiko zein etiketen beharrik gabe.

Azkenik, Schmid-ek (2000) hurbilpen berrienetakoa aurkezten du. Egileak dioenez, ez da testutik kanpoko inongo informazioren beharrik esaldi amaierak markatu ahal izateko. Berak proposatutako sistemak testua bi aldiz korritzen du, lehenengoan beharrezkoak dituen datu estatistiko guztiak erauzten ditu eta bigarrenean, datu horiek erabilita, puntuaren funtzioa identifikatzen du. Puntuaren funtzioa erabakitzeko bost kasu bereizten ditu[17] eta bakoitzaren inguruko azterketa teorikoa burutu ostean kasu bakoitzerako kalkulatu beharreko datu estatistikoak lortzen ditu. Azken finean, puntuaren aurreko eta ondoko hitzei buruzko maiztasun neurriak besterik ez ditu behar erabaki ahal izateko. Maiztasunak kalkulatzeko, Mikheevek laburdura berrietarako erabiltzen duen teknika berbera aplikatzen du, baina Schmiden kasuan puntu aurreko eta ondorengo hitz guztietarako egiten du.

Emaitzei erreparatuz ingeleserako emaitza oso onak lortzen dituela ikus daiteke, %99,79 ondo identifikatuz *Brown Corpus* tratatzerakoan eta %99,62 *Wall Street Journal* corpusean. Alemanerako 3, 4 eta 5 luzerako atzizkien —hitz-bukaeren— analisia ere gehitzen du emaitzak hobetzeko. Egilearen arabera, nahiko emaitza onak lortzen dira alemanerako, baina ez zegoen esaldi amaiera eskuz markatuta zuen corpusik ebaluaziorako. Hori dela eta, atzizkien informazioaren erabilera ebaluatzeko informazio hau ingeleserako erabili eta %99,9ko zuzentasuna lortzen du *Brown Corpus*ean.

Euskararen aurreprozesuari dagokionean, tokenizazio mailan erabakia hartzen da eta, ebaluaketa zehatzik egin ez den arren, kasurik gehienetan egoki desanbiguatzen da. Hala ere, badira kasuak, puntuz amaitutako zenbakiak esaterako, gaizki tokenizatuz gero ondorengo urratsetan prozesu akatsak sor ditzaketenak, (4) adibidean bezala, eta laburduren kasuan, puntua bereizi uztean esaterako, (2) adibidean bezala. Ondorioz, hasieratik sortutako erroretzat hartzen dira, eta gainerako atazen ebaluazioari begira, analisi egokia eskuz gehitu zaie. Hori dela eta, etorkizunean gorago aipatutako tresnen antzekoa prestatuko da puntuaren anbiguotasuna ahalik eta egokien ebazteko.

[17] (1) laburdura. + maiuskulaz hasitako hitza
(2) hitz normala + . + maiuskulaz hasitako hitza
(3) hitz normala + . + minuskulaz hasitako hitza
(4) laburdura. + . + maiuskulaz hasitako hitza
(5) laburdura. + . + minuskulaz hasitako hitza

II.2.1.2 Tokenizatzailearen deskribapena

Ezaguna denez tokenizatzailea LNParen lehen urratsa da, sarrerako testua unitate edo tokenetan zatitzen duena. Bere emaitzak gainerako prozesuak elikatzen ditu, analisi morfologikoa kasu. Bere eginkizunen artean ondoko elementu hauen identifikazioa eta tratamendua du gure kasuan:

- zenbakiak, arruntak edo erromatarrak, dagokien deklinabidearekin
- laburdurak eta siglak dagokien deklinabidearekin
- lerro-bukaeran hitza banatzen duen marratxoa (*hyphenation*)
- zuriuneak eta puntuazio zeinuak, hitzen arteko bereizgarriak direlako
- gainontzeko markak eta karaktere bereziak
- maiuskulaz idatzitako hasierako letrak, zatiak eta izenburuak
- posta elektronikoen helbideak eta web-guneen helbideak
- corpusetan agertu ohi diren testu-identifikazioak —urtea, testu-mota, idazlea, etab. zehazten dutena—, orri-zenbakiak, beste hizkuntzen aipamenak, etab.

Konplexutasun horren aurrean eta beste ezagutzaile batzuen bidetik, automata bat da identifikazioaz arduratzen den tresna. Aipatutako token mota guztiak ezagutzeko definitutako automatak 48 egoera eta 17 karaktere-multzo ditu, automata txikia da, beraz, 48x17ko trantsizio-taula duena. Taula hori datu-fitxategi batean gordetzen da egoera bakoitzari dagokion informazio guztiarekin batera. Informazio horretan egoera bukaerakoa ala ez-bukaerakoa den, zein token-mota ezagutzen den eta tokenak prozesu gehigarririk behar duen ala ez kodetzen da.

Horrela, esate baterako, zenbaki deklinatu bat ezagutu denean, ZEN_DEK etiketa eman behar zaio eta letraz idatzitako bere baliokidea kalkulatu behar dela adierazten zaio aurreprozesuari. Honek, prozesu gehigarri horren kodearen arabera funtzio bat edo beste erabiliko du, analizatzaile morfologikoak behar duen informazio guztia lortzeko. Beraz, token-mota berri bat gehitu nahi baldin bada, edota karaktere-multzoa zabaldu nahi izanez gero, ez da programa aldatu behar, ez bada prozesu gehigarri berri bat erabili behar.

Oraingoz sarrerako formatu bakarra onartzen du, testu gordina hain zuzen ere, baina etorkizunean HTML, XML edota SGML formatuak onartzeko beharrezkoak diren aldaketak burutuko dira sarrera tratatzen duen moduluan. Irteerari dagokionean, bi modutan funtziona dezake: tokenizatzaile hutsaren funtzioa egiteko, token guztiak —edota testu-hitzak bakarrik— emanik, edota analizatzaile morfologikoaren sarrera izateko, token bakoitza beharrezkoa duen informazio gehigarriarekin hornituta. Edozein kasutan SGMLz ere sor dezake irteera, kapitulu honen azken atalean ikusiko den formatua sortuz.

II.2.2 Hitz bakunen tratamendua

Hitz bakunen tratamenduak sarrerako testu-hitz bakoitzeko interpretazio posible guztiak identifikatzen ditu, emaitza gisa hitzaren analisi morfosintaktiko guztien multzoa emanik. Lemak eta morfemak lexikotik datozkien ezaugarriekin lotzen dira. Ezaugarri horiek kategoria, azpikategoria, deklinabide atzizkia, numeroa eta mugatasuna dira izen eta adjektiboen kasuan, eta modu/denbora eta aspektua aditzen kasuan. Hauekin batera funtzio sintaktikoak eta zenbait ezaugarri semantiko ere ematen dira.

Analizatzaile morfologikoaren eraikuntza ondoko bi faseetan izan zen burutua. Lehenengoan, euskara estandarrerako prozesadore morfologikoa eraiki zen, baina honek ezagutzen duen hitz-multzoa —estaldura-tasa— %95 ingurukoa besterik ez zen. Emaitza horren arrazoi nagusienetako bat euskalkien erabilera hedatua izanik, bigarren fasean analizatzailea hitz ez-estandarrak tratatzeko zabaldu zen, prozesadore morfologiko sendoa lortuz.

Horretarako bi analizatzaile berri sortu ziren: batak, aldaera dialektalak eta gaitasun-desbideratzeak tratatzen ditu eta, besteak, analizatzaile orokor edo lexikorik gabeko analizatzaileak (*guesser*), bere lema lexikoan ez duten hitzak prozesatzen ditu. Gainera, tratamendua osatzeko, erabiltzailearen lexikoen kudeaketarako hedakuntza ere burutu da.

Hiru analizatzaile hauek bata bestearen atzetik aplikatzen dira, II.2 irudian arestian azaldu den bezala, eta modu inkrementalean egiten dute lan: lehenengo pausoan hitzak estandar gisa interpretatzen saiatuko da analizatzaile estandarra; honek analisirik ematen ez badu, aldaera dialektal edota gaitasun-desbideratze gisa prozesatzen saiatuko da bigarren analizatzailea; azkenik, aurreko bi urratsek emaitzarik ematen ez badute, lexikorik gabeko analizatzaileak tratatuko ditu hitzak.

Horrenbestez, analizatzailearen estaldura %100ekoa izatea lortu da eta tratamendu-multzo horrekin aberasturiko analizatzaileak honako ezaugarriak ditu:

- **Orokorra**: euskara estandarraren forma gehienak analizatzeko eta sortzeko gai da.
- **Malgua**: ez-orokorrak edo ez-estandarrak diren formen ezagutza erabiltzailearen lexikoek eta aldaeren tratamenduak bideratzen dute, prozesadore morfologikoari malgutasuna emanez.
- **Sendoa**: Lexikorik gabeko lematizazioari esker beste urratsetan ezagutzen ez ziren hitzen analisia bideratzen da, sistemari sendotasun handiagoa emanez.

Hala ere, analizatzailearen ebaluazioan sendotasuna lortzeaz gain beste hainbat hobekuntzaren beharra ikusi zen:

- **Eraginkortasuna**: analizatzaile morfologikoa LNParen beste hainbat aplikazioen oinarria da, hala nola, *Xuxen* zuzentzaile ortografikoa (Aldezabal *et al*. 1999-a), GaIn Intraneterako bilatzailea (Aizpurua *et al*. 2000) edota EUSLEM lematizatzailea bera. Aplikazio hauek denbora errealean lan egitea garrantzitsua da erabilgarriak izango badira. Eraginkortasuna hobetzeko egoera finituko tresnen inplementazio azkarragoa erabili da (Alegria *et al*. 1997; Alegria *et al*. 2001).
- **Zuzentasuna**: zuzentasuna hobetzeko bide bat lexikoa aberastea da, eta lan hori etengabe burutzen da EDBLn sarrera berriak gehituz. Baina beste bide garrantzitsu bat akatsak ekiditea da. Analisi morfologikoak eginiko akatsak aztertu dira eta erroreen multzo handia hiru analizatzaileak modu inkrementalean aplikatzeagatik datorrela ikusi da. Hau da, analizatzaile batek emaitzarik sortzen duenean aukera gehiago ez begiratzearen murriztapena egitean, analisi zuzena lortzeko aukera galtzen da.
- **Zehaztasuna**: hitz ez-estandarren kasuan, gainsorkuntza dela medio, anbiguotasuna altuagoa izaten da, bai batezbesteko analisi kopuruari baita anbiguotasun-tasari dagokionean. *Guesser* gehienetan anbiguotasuna hasieratik murrizten den arren, MORFEUSen lan hori geroko moduluek egingo dute. Bestalde, lematizazioa helburu, garrantzitsua da etiketa bakoitzeko lema bakarra ematea eta desanbiguaziorako moduluak lemen anbiguotasuna ebatziko ez duenez, testuingurua kontuan hartzen ez duen desanbiguazio-prozesu baten beharra ikusi zen. EUSLEMen lehenengo prototipoan (Ezeiza 1997), hitz ez-estandarren tratamendua gehitu zen —lan horretan desanbiguazio lokal deiturikoa—. Bertan azaltzen denez, hitz ezezagunen interpretazio kopurua 13tik 4ra jaisten da prozedura horren bidez.

Analizatzailearen zuzentasun neurriei dagokionean, emaitzak nahiko onargarriak dira, %99,5etik gorako neurriak lortzen direlarik. Dena dela, erroreen azterketaren ondorioz errore gehienak izen berezietan egiten direla ikusi da. Zehatzago esateko, hitz estandarretan egindako erroreen %75a eta aldaeretan egindakoen %80a gaizki analizatutako izen berezi eta sigletan gertatzen dira. Akats horietako batzuk analizatzaileak aplikatzeko murriztapen horiek erlaxatuz zuzen daitezkeela ikusi da. Analizatzailearen eraginkortasun eta zuzentasunaren inguruan burututako hobekuntzen berri III. kapituluan emango da.

Baina analizatzaile morfologikoaren zuzentasuna hobetzeko, anbiguotasuna handitu egiten da, eta horrekin batera analisi morfologikoan oinarriturik dauden aplikazioen zehaztasunaren galera etor daiteke. Beraz, anbiguotasuna lehenbailehen murriztea komeni da, ahalik eta errore gutxien gehituz. Hitz ez-estandarren anbiguotasunari dagokionean, eraginkortasuna hobetzeko egindako inplementazioa dela medio, MORFEUSen abiadura modu esanguratsuan handitzea lortu da, baina hitz ez-estandarren gainsorkuntza areagotu egin da, ezezagunen kasuan batez beste 19-20 interpretazio lortzen direlarik. Emaitza hauei erreparatuz, (Alegria

1995) eta (Ezeiza 1997) lanetan deskribatutako hitz ez-estandarren tratamendua birplanteatu eta, hitz hauen tipifikazioa egin ostean, sailkapen horretako hitz-multzo bakoitzerako prozedura bat diseinatu da, aurretik egindako lana ere eguneratuz. Zehaztasunaren hobekuntzarako definitutako prozesuak IV. kapituluan deskribatu eta ebaluatzen dira.

II.2.3 Hitz anitzeko unitateen tratamendua

Lematizazioa helburu izanik, gure ustetan hitz anitzeko unitateak lehenbailehen identifikatu behar dira, bere osagaiek elkarrekin osatzen baitute euren lema. Gainera, testua etiketatu nahi da, eta, askotan, unitate konplexua osatzean hitzei dagokien informazioa ere aldatu egiten da. Hau ikusirik, hitz anitzeko unitate lexikalak morfologiaren ondoren baina sintaxiaren aurretik tratatzen dira.

Hitz anitzeko unitate lexikalak mugatzeko jarraitutako irizpideak (Ezeiza 1997) lanean azaltzen dira sakonean. Laburbilduz, hitz elkartuak, lokuzioak eta kolokazio murriztuak kontsideratuko dira hitz anitzeko unitate lexikal (HAUL), eta EDBLn landuko dira gehienak. Hitz elkartuen kasuan maiztasun handienekoak landu dira soilik, elkarketa oso emankorra izanik, ezinezkoa bailitzateke datu-basean guztiak lantzea.

Horiek identifikatu eta tratatzeko HABIL programa diseinatu zen. HABIL tresna probatzeko EEBSko 3.000 unitate konplexuetatik 500 inguru aukeratu eta bakoitzari dagokion informazioa osatzeko hizkuntzalariak corpusetaz eta euren esperientziaz baliatu ziren. Saiakuntzen emaitzetatik abiatuta zenbait hobekuntzen beharra ikusi zen:

- HAULei buruzko informazio morfosintaktikoa gehiago zehaztea interesgarria da.
- HAUL gehiago landu behar dira, testuetan agertzen diren guztiak identifikatu nahi badira.
- Bestelako Hitz Anitzeko Unitate batzuk ere identifikatzea komeni da, hala nola, datak, hitz anitzeko zenbakien adierazpenak, izen berezi konposatuak, etab. MUCen (*http://www.muc.saic.com*) izendun entitateak (*Named Entities*) deritzatenak[18].

Hobekuntza hauek bideratzeko hainbat aldaketa egin dira. Hasteko, informazioa beste modu batean landu ahal izateko, datu-basearen diseinua aldatu da. Gainera, HAUL berriak corpusetatik automatikoki erauzteko tresnen azterketa egiten da eta, azkenik, bestelako unitateak identifikatzeko tresna diseinatu eta inplementatu da. Honi guztiari eskaini zaio V. kapitulua.

[18] MUC: *Message Understanding Conference.* Informazioaren Erauzketarako sistemak ebaluatzeko antolatzen den konferentzia da. Izendun entitateen identifikazioa burutu beharreko atazetako bat da.

II.2.4 Morfosintaxia

Arestian esan den bezala, analizatzaile morfologikoaren irteera sarrerako hitzen segmentazio morfologikoa da, eta morfema bakoitzari dagokion informazio morfologikoa —morfosintaktikoa askotan— erantsi zaio. Zenbait aplikaziotarako, zuzentzaile ortografikoa edo oinarrizko lematizatzailea kasu, nahikoa izango da hitz bakoitza bere osagai diren morfemetan banatu eta oinarrizko informazio morfologikoa ematea. Hala ere, lematizazio/etiketatze, sintagma-ezagutze edota esaldien mugen identifikaziorako tresnek hitz osoaren analisi morfologiko —edota morfosintaktiko— orokorragoaren beharra dute.

Azken finean, segmentazioa gainditu eta formaren barruko informazioa elaboratu behar da, hitzaren egituraren berri emateko eta horixe da analisi morfosintaktikoaren zeregina.

Analizatzaile morfosintaktikoak estaltzen dituen fenomeno linguistikoen artean hauek aipatzen dira:

- Kasu, numero eta mugatasunaren informazioa morfema batean baino gehiagotan eman daiteke. Askotan azken atzizkiaren informazioa da hitz osoari dagokiona, baina beste aukera batzuen tratamendua ere zehaztu behar da.
- Izen-elipsia sortzen da genitiboaren ondoren beste kasu bat agertzen denean. Hitz osoaren analisian lema guztien informazioa mantentzea izango da interesgarriena.
- Kategoria-eratorpena eta elkarketa. Kasu hauetan lema baten kategoria aldaketa eta bi lemen bildura gertatuko da hurrenez hurren.

Hauexek dira tratamendu morfosintaktikoaren beharra sortzen dituzten arrazoiak:

- Proiektu orokorraren asmoa testu-hitzaren azken analisia (morfosintaktikoa) ematea da, beti ere emaitza osoaren berri emanez, barruan dagoen informaziorik galdu gabe; alegia, formaren barruko informazioa antolatu eta analisia eskaini nahi du, osagaien segida gaindituz.
- Horretarako hizkuntzaren teoriari berari buruzko erabakiak hartu behar izan dira, analisia egiterakoan goratu behar den informazioa erabakitzeko, beti ere kontuan izanda analisi morfologikoa bi mailatako formalismoan oinarritzen dela. Hasiera batean behintzat *Alvey* sistema izan da nolabaiteko oinarria (Carroll 1993; Grover *et al*. 1993).
- Baina bazegoen aurretiaz hartu beharreko beste hainbat erabaki ere, alegia, lexikoien antolamendua, sarrera batzuk berez duten mugatasuna (izenordainak, izen bereziak, etab.) eta hitzaren barruko elipsia. Horiek guztiak eztabaidatu eta hartutako erabakien berri ematen da horretarako erabili diren irizpideen argitan.

Analizatzaile morfosintaktiko erabilgarria lortu nahi denez, irekia behar du izan, ez du inongo formalismo zehatzetara edo helburu bakanetara lotuta egon behar. Horretarako, baterakuntzan oinarritutako hitzaren gramatika definitu da (Aduriz *et al*. 2000; Gojenola 2000). Hitzaren gramatikak morfemetatik lortutako informazioa konbinatzen du hitz-formaren interpretazio bakoitzeko ezaugarri-egitura bat emanez emaitza gisa.

PATR formalismoaren inplementazio bat erabili da (Douglas eta Dale 1992). Inplementazio hau sinpletasun eta malgutasunagatik aukeratu da. Analizatzaile morfosintaktikoari zein hitzaren gramatikari buruzko argibide zehatzagoak Gojenolaren (2000) tesian lor daitezke.

Morfosintaxia ez da hemen aurkezten den lanerako erabili, dagoeneko inplementaturik dagoen arren, integrazio bidean dagoelako. MORFEUSen hurrengo bertsioan kapitulu honetan aipatutako gainerako moduluekin batera erabiliko da.

II.3 EUSLEM, lematizatzaile/etiketatzailea

Morfosintaktikoki analizatutako testua sarrera izanik unitate bakoitzari testuinguru horretan dagokion lema eta etiketa morfosintaktikoa esleitzea da EUSLEMen funtzioa, hau da, sarrera anbiguoa testuinguruaren arabera desanbiguatzea.

Desanbiguaziorako teknika desberdinak erabili ohi dira, VI. kapituluan ikusiko den legez, orokorrean bi multzo bereizten direlarik: teknika linguistikoak eta teknika estatistikoak. EUSLEM definitzeko bi teknika-multzoak aztertu dira eta bi desanbiguazio-modulu garatu dira modu independentean, geroago konbinatzeko asmotan.

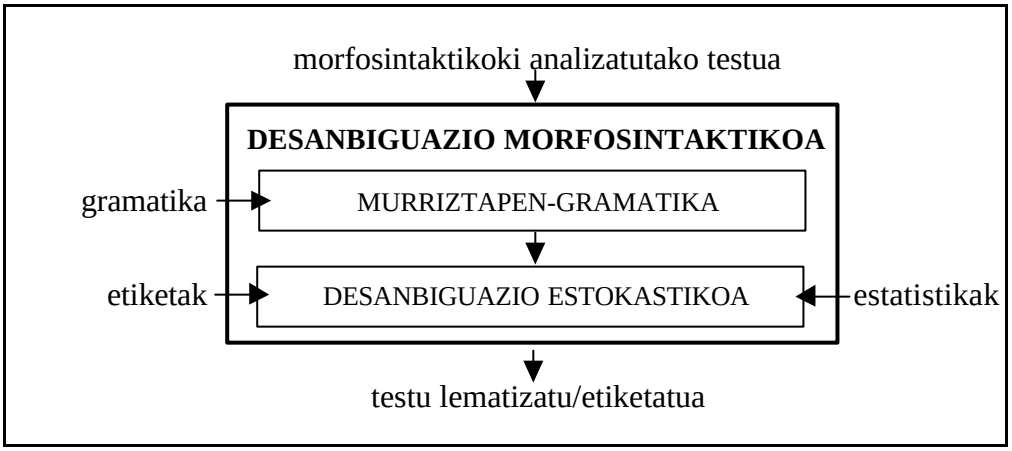
Batetik, teknika linguistikoetan oinarritutako murriztapen-gramatika jorratu da, desanbiguaziorako gramatika bat definitu delarik. Lan honen deskribapen sakona Adurizen (2000) tesi-lanean aurkezten da, baina, hala ere, modu laburrean aurkeztuko dira VI. kapituluan murriztapen-gramatikaren bidez lortutako desanbiguazioaren emaitzak.

Bestetik, teknika estatistikoak erabiltzen dituen Markov-en eredu ezkutuak erabili dira. Desanbiguazio-teknika honek corpus handi batetik automatikoki erauzitako estatistikak erabiltzen ditu hitzaren aukeren artean bat hautatzeko. Desanbiguazio estokastikoaren emaitzak (Ezeiza 1997) lanean aurkeztu ziren, baina lan honetarako eguneratu dira eta desanbiguazioari buruzko kapituluan aurkeztuko dira.

(Ezeiza 1997) lanaren ondorioetan azaldu zen bezala, desanbiguazio estokastikoaren emaitzak hobetzeko zenbait modu daude. Bertan aipatu zenez, desanbiguaziorako erregelak erabiltzeak emaitzak hobetzen ahal zituen. Hori zela eta, independenteki garatutako metodo

biak konbinatzea ildo interesgarritzat jo zen. Izan ere, euskaraz emaitza onak lortzeko tekniken konbinaketaren beharra hasieratik izan dugun hipotesia da, ikasketarako corpus gainbegiratu txikia izanez gero, behinik behin. Horixe da tesi-lan honetan jorratutako gai garrantzitsu bat eta VI. kapituluan aurkeztuko dira konbinaketaren nondik norakoak eta emaitzak.

Desanbiguatzeko erabilitako teknika edozein izanik ere, desanbiguazio-prozesuaren emaitzak automatikoki ebaluatzea komeni da. Horretarako, aldez aurretik egin beharreko lan garrantzitsu bat dago. Izan ere, baztertutako analisiak kendu beharrekoak direnentz jakiteko, analisi zuzena zein den jakin behar da. II.3.1 atalean desanbiguaziorako datuak sortu eta emaitzak ebaluatzeko beharrezkoa den eskuzko desanbiguazioa aurkezten da.



II.3 irudia.- EUSLEMen egitura.

Azkenik, etiketatzailerako erabiliko diren etiketak ere definitu behar dira. Gehienetan, kategoria edota kategoria-azpikategoria eman ohi da, baina hau aplikazioaren araberakoa izango da. Esate baterako, oinarrizko lematizazioak kategoria besterik ez du behar izango lema aukeratu ahal izateko, baina sintaxian erabili nahi denean, informazio morfosintaktiko guztia beharko da.

Horregatik, etiketa-sistema definitzerakoan analisi morfosintaktikoan oinarriturik erabilera orokorreko sistema diseinatu da, inongo aplikazioren interesetatik at, deskribapen linguistiko hutsetik abiatuta. II.3.2 atalean aurkezten da EUSLEMen erabili den etiketa-sistema.

II.3.1 Eskuzko desanbiguazioa

Arestian esan bezala, etiketatzailearen datuak sortu eta emaitzak ebaluatzeko erreferentzia-corpus bat izatea komeni da. Orokorrean, eskuz edota erdi-automatikoki markatutako corpusa erabiltzen da. Gure kasuan, aldez aurretik markatutako testurik ez zegoenez, morfologikoki analizatu ondoren eskuz desanbiguatu eta markatu behar izan zen.

Eskuzko desanbiguazioa ez da batere erraza. Orokorrean, etiketatzaileek hitz bakoitzaren kategoria gramatikala ematen dute, baina gure kasuan erabakiak ez dira kategoria-mailan geratzen. Izan ere, esan bezala, zenbait aplikaziotan kategoria besterik ez da behar izango, baina beste zenbait aplikaziotarako interesgarria izango da informazio gehiago jasotzea. Gainera, hurrengo urratsa analisi sintaktikoa izanik, analizatzaileak ematen duen informazio guztia behar da. Beraz, eskuz desanbiguatzerakoan hitz bakoitzeko informazio morfosintaktiko osoa aukeratu behar izan da, baita dagokion informazio sintaktikoa ere.

Horregatik, eskuz desanbiguatzerakoan fin jokatzea oso garrantzitsua da, emaitza horiek ondorengo urratsen oinarri izango badira. Bereziki teknika estatistikoak erabili nahi direnean rol garrantzitsua jokatuko du eskuz desanbiguaturiko corpusak, datu estatistikoak bertatik erauziko baitira. Hori dela eta, informazio morfosintaktiko guztia erabilita desanbiguatzeari ekin zaio.

Lan honetan erabili den corpusa, UZEIk EEBS proiektuan bildutako corpusetik hartutako 28.300 token inguru eta *Euskaldunon Egunkaria*ko atal desberdinetatik jasotako beste 14.800 inguru morfosintaktikoki analizatu eta eskuz markatu ziren. EEBS corpusaren ezaugarri nagusienetako bat orekatua izatea da. Ezaugarri honek eremu zehatzetatik urrundu eta aplikazioen orokortasuna ebaluatzen lagunduko du. *Euskaldunon Egunkaria*tik jasotakoak gehienbat izen berezien tratamenduaren egokitasuna egiaztatzeko balio izango du.

Eskuzko desanbiguazioa *'doble-blind'* izeneko metodologia (Voutilainen eta Järvinen 1995) jarraituz burutu da, hau da, bi pertsonek corpus bakoitza bere aldetik markatzen dute. Ondoren emaitzak automatikoki konparatzen dira eta desanbiguatzaileek irizpideak adostu egiten dituzte desadostasuna gertatu deneko kasuetan. Amaitzean adostasun maila %100etik hurbil egoten da, baina normalean ezin izaten da guztiz adostu, bereziki informazio morfosintaktikoa hain aberatsa izanik.

Gure kasuan, lehenengo hurbilpenean deskribapenaren aberastasuna dela medio %5eko desadostasuna egon zen bi desanbiguatzaileen artean. Diferentziak eztabaidatu ondoren, %1 baino gutxiago geratu zen ebazteke. Kasu horietan, estandarizazio faltaren ondorioz hizkuntzalariek perspektiba linguistiko desberdina zutelako ez zen adostasunik lortu. Hala ere, azken horietan hizkuntzalarietako baten erabakiak besterik ez dira kontuan hartu, analisi bakarra uzteko irizpide koherentea jarraitzearren.

Ondoren, corpus hau bi zatitan banatu da, arestian esan bezala. Zati handiena erreferentzia-corpusari dagokio eta EEBSko 27.000 eta *Euskaldunon Egunkaria*ko 9.000 token inguruk osatzen dute. Bigarren zatiak, egiaztapenerako corpusak, berriz, EEBSko 1.300 eta *Euskaldunon Egunkaria*ko 5.800 token inguru ditu.

II.3.2 Etiketa-sistemaren diseinua

Etiketatzaile baten diseinuan etiketa-sistemaren hautaketa oso garrantzitsua da (Aduriz *et al.* 1995), tresnaren erabilgarritasuna eta anbiguotasun-tasa erabilitako etiketa-multzoaren menpekoak baitira.

Aukeratutako etiketa-sistemak desanbiguazio-prozesua nolabait baldintzatuko du. Etiketa kopuruaren arabera anbiguotasunaren neurriak aldatuko dira. Etiketa kopurua handitzen bada, anbiguotasuna handitzen den neurrian interpretazioen arteko diferentziak etiketan islatzeko aukera ere emango da, eta etiketen araberako anbiguotasuna gutxitzen denean, berriz, interpretazioen berezitasunak gal daitezke. Horregatik da hain garrantzitsua etiketa-sistemaren hautaketa.

Baina etiketa-multzoa diseinatzerakoan anbiguotasun-tasa kontuan hartzeaz gain etiketa horiek adierazten duten informazioa ere izan behar da gogoan. Izan ere, bien arteko oreka lortzea izango da helburua. Etiketak adierazten duen informazioa konplexua bada, orduan anbiguotasun-tasa ere handia izan daiteke eta horrek desanbiguazio-prozesuaren doitasunean eragin handia izan dezake. Beraz, erdiko bidea hartzea komeni da, anbiguotasun-tasa gehiegi handitu gabe, etiketan informazio garrantzitsuena mantentzea hain zuzen ere. Aurrerago euskal testuen anbiguotasun-tasa etiketa-multzoaren arabera nola aldatzen den aztertuko da.

Euskararako etiketa-multzoa definitzerakoan honako oztopoak izan ditugu:

- tratamendu automatikorako ez zegoen etiketa-multzo exhaustiborik. Batetik, EEBS corpusa lematizatu zen, baina ez zen sistematikoki etiketatu eta, bestetik, paperezko hiztegietan kategoriak jartzen direnean ez dira sistematikoki aplikatzen.
- analizatzaile morfologikoaren irteera ezin da zuzenean erabili hitzak etiketatzeko, euskara hizkuntza eranskaria eta morfologikoki aberatsa delako. Etiketarako esanguratsua den informazioa lortzeko analisian emandako datuak tratatu behar dira elipsia, kasuen eransketa, lema konposatuak, etab.ekin lotutako arazoak ebazteko.

Dena den, etiketa-sistema "linguistikoa" diseinatu nahi delarik, analisi morfologikotik abiatuta barne-informazio garrantzitsuena etiketan islatu nahi izan da, aplikazioen emaitzen hobekuntzari erreparatu gabe. Berez, irizpide linguistikoa gailendu da beti, eta deskribapen linguistiko xeheagoa egitean anbiguotasuna handitu egin da urteetan zehar, desanbiguazio-prozesuaren konplexutasuna areagotuz.

Etiketatzaile askotan ikusi ahal denez, desanbiguatzen zailak edota maiztasun handikoak diren hitzek etiketa propioak izaten dituzte. Honen arrazoia etiketatzailearen zuzentasuna handitzea da.

> *"Kupiec divided words into equivalence classes depending on their possible POS tags. There were 129 such 'word equivalence classes'. With additional classes for the most common words, there were about 400 classes. (...) Kupiec also describes improvements to this basic model. More POS categories were added; individual equivalence classes were created for the 100 most frequent words."* (Franz 1996:16)

Euskararen kasuan ere etiketatzailearen emaitzak hobetzeko hainbat etiketa berezi defini zitezkeen, hala nola, izen berezi guztietarako etiketa bakarra edota aditz trinko eta aditz laguntzaileetarako bakarra, horiexek baitira anbiguotasun zailenetakoak (*hard ambiguities*) eta errore gehien sortzen dituztenak. Baina, gure ustez, etiketa dagokion interpretazio morfologikoarekin etorri behar da bat, eta ez hitzaren interpretazio-multzoarekin. Horregatik, ez dira hitzetarako etiketa bereziak definitu, orokortasuna eta aplikazioarekiko independentzia baitira definitu nahi izan den etiketa-sistemaren ezaugarri garrantzitsuenak.

II.3.2.1 Etiketa-sistema

Etiketa-sistema zehaztean ondoko puntuak hartu behar dira kontuan:

- elipsia, eratorpena eta konposizioa nola adierazi etiketan.
- lematizatzaile/etiketatzaileak orokorra behar du izan, etiketatzea bestelako aplikazioen aurre-pausoa izango delako, hala nola, analizatzaile sintaktikoa, indexazio automatikoa zein informazioaren berreskurapena.
- analizatzaile morfologikoak ematen duen informazioarekin bat etorri behar du.

Etiketa-multzo orokor, mailakatu eta irekia aukeratu da. Sistema lau mailatan antolaturik dago, erabiltzaileak hauta dezan programa erabiltzean parametroen bidez interesatzen zaion maila:

- Lehenengo maila kategoriari —izena, adjektiboa, aditza, etab.— dagokio. Maila honetan hamasei kategoria orokor sartzen dira eta elipsirako eta konposatuen indexaziorako etiketa bereziak erabiltzen dira. Lematizazio arrunterako oinarrizko etiketa-multzoa da. Guztira 20 etiketa dira.
- Bigarren mailan kategoria eta azpikategoria hartzen dira kontuan. Adibidez, izenetan hiru mota izango dira: izen arruntak, pertsona-izen bereziak eta leku-izen bereziak. Indexazio automatikorako interesgarria izan daiteke bigarren mailako informazioa, izen berezi eta arruntak bereizi eta azken taldekoak soilik tratatzeko. Guztira 45 etiketa.
- Hirugarrenean analizatzaileak ematen duen bestelako informazio morfologiko interesgarria gehituko da, esaterako, kasua izen eta adjektiboetan edota modu/denbora aditzetan. Bai aurreko etiketa-maila bai hauxe bera ere desanbiguazio estokastikorako datuak lortzeko zein desanbiguazio-erregelen diseinurako dira egokiak. Sintaxiari begira ere interesgarria

da hirugarren mailan desanbiguatutako sarrera izatea, sarrerako anbiguotasuna handiegia ez izateko. Ikasketa-corpusean 320-340 etiketa inguru agertzen dira, baina EEBSko 800.000 testu-hitzeko corpus batean 500 etiketa agertzen dira guztira eta *Euskaldunon Egunkaria*ko beste 400.000 testu-hitz gehituta etiketa berriak agertu dira. Horietako batzuk oso agerpen gutxi izan arren, zilegi dira eta, hortaz, kontuan hartu beharrekoak. Gai honen inguruan VI. kapituluan eztabaidatuko da.

- Azkenik, laugarrenean analizatzaile morfologikoak ematen duen informazio guztia emango da. Maila honetako etiketatzea, besteak beste, tratamendu sintaktikoaren sarrera gisa erabil daiteke. Aipatutako 800.000 hitzeko corpusean 24.000 interpretazio morfologiko desberdin agertu dira, eta beste 400.000koan beste 7.000 interpretazio berri agertu dira, baina seguruenik corpus handiago tratatuz gero, kopuru hori gaindi daiteke.

Desanbiguazioan erabiliko den etiketatze-maila erabiltzailearen eskuetan dago. Orokorrean, bigarren maila izango da erabiliko dena, baina zenbait kasutan informazio gehiago erabiltzea interesgarria izan daiteke.

Dena dela, laugarren mailan desanbiguatu nahi izanez gero, ez da teknika estokastikorik erabiliko, murriztapen-gramatikaren emaitzak baizik. Honen arrazoia, laugarren mailan erabiltzen den informazioaren aberastasuna da. Aberastasun horrek interpretazio posibleen kopurua teknika estatistikoen erabileratik kanpo uzten du, 30.000 "etiketa" edo interpretazio desberdinen estatistikak lortuta ere, emaitza horiek esanguratsuak izateko corpus itzelak beharko lirateke.

Gainera, hori lortuta ere, 30.000x30.000ko matrizeak erabiltzea —eta handiagoak, anbiguotasun-klaseak are gehiago liratekeelako— ez litzateke batere eraginkorra izango, ez erabiltzeko momentuan ezta sortu eta eguneratzen ere. Izan ere, etiketen arteko konbinaketa asko eta asko corpusetan gauzatu ez arren, posizio horiek matrizeetan adierazi beharko lirateke. Horregatik erabiliko dira lehenengo hiru mailak desanbiguazio osoa egiterakoan soilik.

Hala ere, 4. mailako emaitzak lortzeko beste modu bat ere badago, 2. edo 3. mailako desanbiguazioa burutu eta aukeratutako etiketei dagozkien analisiak berreskura daitezke, alegia. 2. mailako desanbiguazioa burutuz gero anbiguotasun handiagoa gertatuko da baina 3. mailakoan baino errore gutxiago egiten dira.

A eranskinean analizatzaile morfologikoak —eta, horrenbestez, etiketatzaileak— erabiltzen duen kategoria sistema aurkezten da, kategoria[19] eta azpikategorien etiketak eta dagokien esanahia alboan dutela.

II.4 Tresnen integrazioa: SGML

1978 urtean, *American Standard National Institute* (ANSI) erakundeak testu prozesamenduan ari ziren hainbat talde jarri zuen harremanetan, edozein motako testuak kodetzeko, egituratzeko eta elkar-trukatzeko balioko lukeen lengoaia estandar eta orokorra definitzeko helburuarekin; 1980 urterako lengoaia horren lehenengo txostenak argitaratu baziren ere, 1985 urtean elkar-lanaren emaitzaren azken bertsioa argitaratua izan zen, *International Standard for Organization* (ISO) erakundeak estandartzat onartu zuelarik: ISO 8879 edo SGML (*Standard Generalized Markup Language*) lengoaia.

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

Estandarizazioaren bideari helduz, IXA taldean hizkuntzaren tratamenduko tresnak integratzeko proposamen bat landu dugu (Artola *et al*. 2000, Aldezabal *et al*. 2002). Integrazio-proposamen honek, euskara prozesatzeko IXA taldean garaturiko —edo garatuko diren— hainbat tresnaren sarrera-irteerak estandarizatzeko eta bateratzeko bide bat eskaintzen du.

Helburu nagusia, beraz, tresna bakoitzaren sarrera eta irteera adierazteko *ad hoc* egindako formatu bat erabili gabe, formatu estandar eta orokor bat erabiltzea da. Formatu orokor eta estandar hori ezaugarri-egituretan oinarritzen da, TEIren araberako ezaugarri-egituretan hain zuzen ere (Sperberg-McQueen eta Burnard 1994)[20]. SGMLz[21] kodeturiko ezaugarri-egiturak erabiliko dira, bada, hizkuntza-tresnon emaitza (eta sarrera) den informazio linguistikoa adierazteko. Formatu estandarra erabiltzeak berarekin dakar sarrera-irteerako programeria normalizatzea eta bateratzea.

---

19 Bertan agertzen diren 16 kategoria lexikalez gain, bereizgarri eta puntuazio ikurretarako beste 4 etiketa erabiltzen dira.

20 *TEI: Text Encoding Initiative.*

21 SGML (*Standard Generalized Markup Language*) dioen lekuan berdin-berdin uler daiteke XML (*Extensible Markup Language*). Izan ere, eta kontu hauek hartu duten norabidea zein den aintzat harturik, (Aldezabal *et al.* 2002) zehazten den guztia indarrean jarri orduko —dagoeneko hala dela esan daiteke ia— SGML ordez XML izango da estandar zabalduena, eta TEI XMLratua edo XCESen jarraibideak nagusituko dira. Hala ere, uste dugu oso aldaketa gutxi eta xeheak egin beharko direla guztia egokitzeko.

Tresnen artean trukatu beharreko informazio linguistikoa konplexua da. Horregatik aukeratu dira ezaugarri-egiturak (FS) adierazpide gisa: ezaugarri-egiturak asko erabili izan dira informazio linguistikoa adierazteko, eta beren espresio-ahalmena handia da. Bestalde, ezaugarri-egitura motak definitzea ere posible da, eta, hartara, eratzen diren egituren zuzentasuna egiaztatzea automatikoki egin daiteke. Sarrera-irteerak SGMLz kodetzeak, gainera, erabili nahi den markatzea formalki deskribatzera behartzen gaitu, eta, horrela, anotazio-corpus batean markatze-arauak betetzen direla bermatzen ahal da.

SGMLz kodeturiko ezaugarri-egiturekin egin beharreko lana erraztearren, datu-mota abstraktu (DMA) multzo edo liburutegi bat eratu, diseinatu eta inplementatu da (Artola *et al*. 2002-ab). Liburutegi honi esker, FSak dituen SGML dokumentu batetik informazioa eskuratzea, edo FSD[22] jakin baten arabera eratutako FSak dituen SGML dokumentu bat sortzea lan erraza eta, batez ere, normalizatua da.

Anotazio linguistikorako oinarritzat hartu dugun eredua, anotazio linguistiko banatuarena da (Zajac *et al.* 1997; Zajac 1998). Ingelesez, *stand-off, sparse, distributed annotation* dira erabili ohi diren terminoak. Eredu honi jarraituz, analizaturiko corpusa daukan dokumentua ez da ezertarako ukitzen, anotazioak bertan idatziz —prozesu desberdinetan lortutako informazio linguistikoa txertatuz—; aitzitik, informazio linguistikoa corpusetatik aparteko dokumentuetan jasotzen da, eta corpusa eta analisiok gordetzen dituzten dokumentuen artean estekak ezartzen dira.

Esan beharra dago, proposatzen den eredua ongi egokitzen zaiola, batik bat, testu-corpus baten *batch* tratamenduari, hau da, testua tokenizatu, segmentatu, morfosintaktikoki analizatu, etab. ondoz ondo egin, eta prozesu bakoitzaren emaitzak gordez joaten den tratamenduari.

Eredu honi jarraiki, beraz, tresna bakoitzaren emaitza hainbat SGML dokumentutan banatzen dela esan liteke; lauzpabost, kasurik konplexuenean. Lauzpabost dokumentu horiek *amaraun* edo *hiperdokumentu* bat osatzen dute, elkarren arteko estekez josiak baitaude. Eredu hau egokia da oso interfaze unibertsal gisa nagusitzen ari den nabigatzailearen bitartez jorratzeko: hizkuntzalaria aise mugitu ahal izango da bertan toki batetik bestera —corpuseko hitzetik bere lematizazio edo analisietara, HAUL baten analisitik corpuseko bere agerpenetara, etab.

Amaraun edo hiperdokumentu hori osatzen duten dokumentuen artean, ondoko sailkapena egin dezakegu:

[22] FSD (*Feature Structure Definition*): ezaugarri-egitura mota konkretu baten definizio formala egiteko TEIk proposatutako dokumentua.

A. Jatorrizko testu-dokumentua.

B. Jatorrizko dokumentuan ezaguturiko testu-zatiak:

-Tokenizatzaileak ezaguturiko token bakunak, hitz bakarrekoak.

-Testuko hitz anitzeko tokenak, multitokenak (HAULak, izen nagusiak, datak, etab.).

C. Analisi-bildumak: segmentazioak, analisi morfologikoak, lematizazioak…

D. Estekak: tokenen eta dagozkien analisien artekoak, esate baterako.

Amarauneko *hariak*, berriz, ondokoen artean aurkituko ditugu:

- B dokumentuetan, A jatorrizko dokumentuko erreferentziak egongo dira, zehatz-mehatz azalduz testuko zein zatiri dagokion identifikaturiko tokena. Jatorrizko dokumentuko erreferentziak egiteko, <*p*> elementuen *id* atributuak eta elementu horien barruko karaktere-desplazamenduak (*offset*-ak) erabiliko dira.
- D dokumentuak, berez, esteka-multzoak dira: B dokumentuetako tokenen eta C analisi-bildumetako analisi-unitateen artekoak, eta multitokenen eta C analisi-bildumetako analisi-unitateen artekoak. Esteka hauek SGML elementu orok eraman dezakeen *id* atributuaren bitartez egiten dira beti.
- Multitokenak, jatorrizko testuko hainbat tokenen bildura diren neurrian, token horien arteko esteken bitartez adieraziko dira. Beraz, multitokenon kasuan, beren analisiekiko estekak D lotura-dokumentu baten eta C analisi-bildumetako analisien artekoak izango dira.

II.4 irudian lematizatzailearen irteera irudikatu nahi izan dugu. Lematizatzailearen emaitza xeheki aztertzen badugu, konturatuko gara dokumentu-amaraun batek osatzen duela, eta dokumentu-amaraun horretan zehar nabigatuz ustiatu ahal izango dugula emaitza hori.

Ikus dezagun, bada, zer dokumentuk osatzen duten amaraun hori, arestian egindako sailkapenaren arabera (A, B, C eta D) multzokatuz:

- EUSLEMen sarrera direnak, hau da, jatorrizko dokumentua (A.1) eta bertan ezaguturiko testu-zatiak (B.1 eta B.2):

  A.1) *t0001.sgm*: jatorrizko dokumentua, SGMLz.

  B.1) *t0001.w.sgm*: tokenizatzailearen emaitza, hau da, testuan bereizturiko token bakunen zerrenda. Dokumentu honetan egiten da jatorrizko dokumentuko itemekiko lotura zuzena, <*xptr*> elementuen bidez. <*xptr*> elementu horiek erabiltzen dira beste dokumentu batzuetan ere, dokumentutik at dauden itemak erreferentziatu behar baldin badira.

B.2) *t0001.mwlnk.sgm*: HAULen egitura deskribatzen duen dokumentua, hau da, testuko multitokenen zerrenda. Dokumentu hau, EUSLEMen sarrera ez ezik, irteera ere bada, EUSLEMek aldatu egin baitezake.

- Dokumentuko lematizazioen bilduma (C.1):

C.1) *t0001.lem.sgm*: jatorrizko testuko token eta multitokenei dagozkien lematizazio desberdinen multzoa.

- Dokumentuko itemen eta lematizazioen arteko loturak (D.1):

D.1) *t0001.lemlnk.sgm*: dokumentuko token eta multitokenen, eta beren lematizazioen arteko loturak, non token edo multitoken bat hainbat lematizaziorekin lotzen bada, ebatzi gabeko anbiguotasunaren seinale baita.

II.4 irudian gezi batzuk ere marraztu dira eta, gezi horiei jarraituz, aise iristen ahal da *t0001.sgm* dokumentuko *hala ere* lokuziotik bere lematizazioraino *t0001.lem.sgm* dokumentuan (L-LOT-LOK-87 identifikadorea duen ezaugarri-egitura), <*xptr*> eta <*link*> elementuen "harian" barrena.

(Ide eta Véronis 1995)-en dioten bezala, dokumentu-amaraun honek prozesu desberdinekiko independentzia eta malgutasun handia ematen digu, prozesu desberdinen emaitzak antzeko modu batean antolatu eta ustiatzeko erraztasuna.

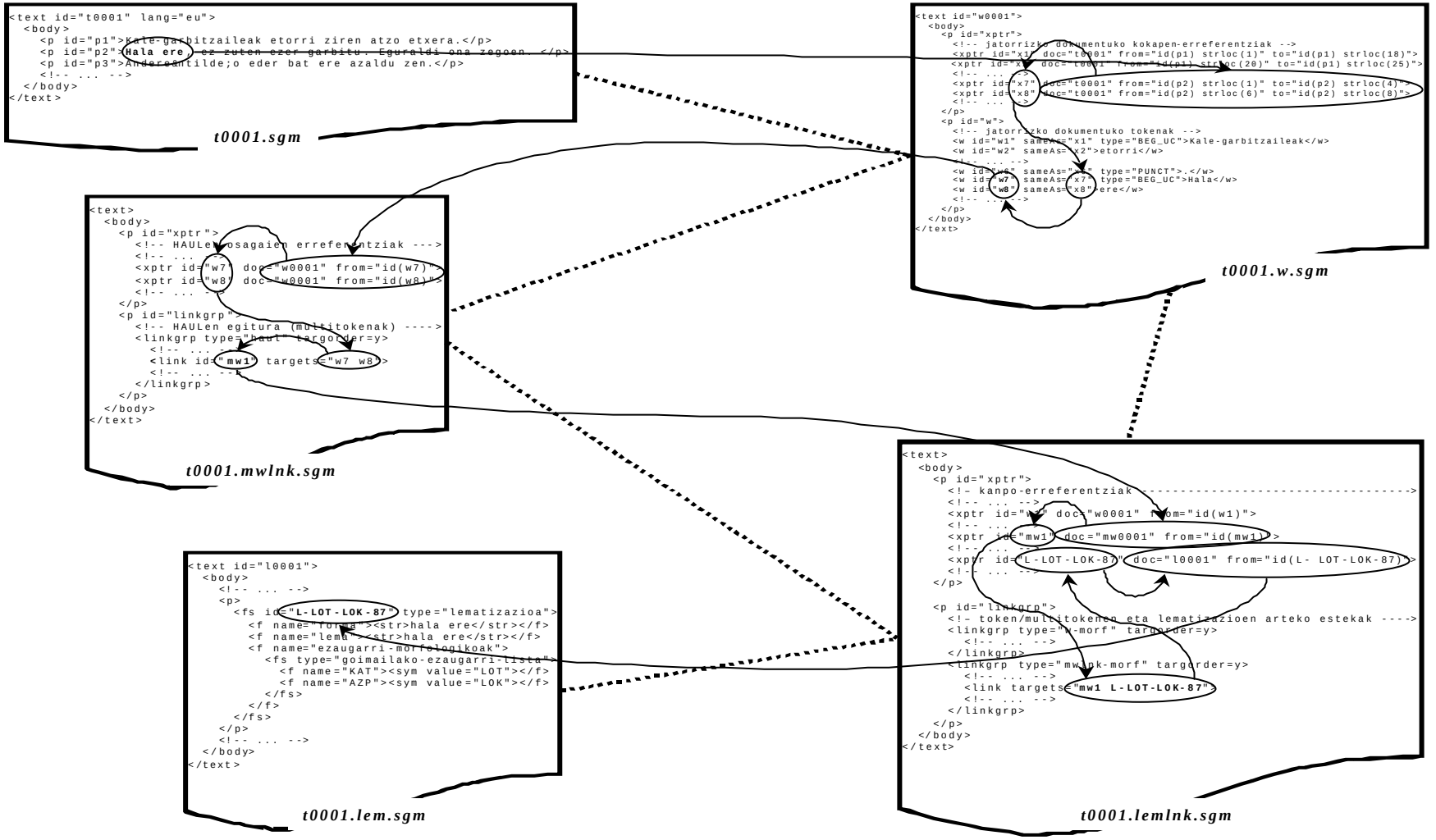
Horregatik da oso inportantea, lehenbailehen horrelako dokumentu-amaraunen gainean sarean nabigatzen dugun bezalaxe —hau da, hipertestu baten gainean bezala— nabigatzeko erraztasuna emango digun programeria sortzea eta garatzea. Halako ingurune interaktibo bat izugarrizko laguntza izango litzateke analisi linguistikoak aztertzeko, eskuzko desanbiguatze-lanak egiteko, etab. Lehen prototipo bat egina dago (Perurena 2000), eta hori bera izan liteke ingurune funtzional eta oso bati buruzko abiapuntu egokia.

Bukatzeko, aipa ditzagun zer abantaila dituen, gure iritziz, tresna arteko sarrera/irteerako informazio-fluxua gauzatzeko, SGML (edo XML) erabiltzeak.

- Testu egituratuak errepresentatzeko ongi definituriko estandar bat da, barne-prozesaketarako molde formal bat eskaintzen duena.
- Datu-trukerako ematen dituen erraztasunak ezagunak dira: DTDa emanik, DTD horrekin bat datorren dokumentu bat prozesatzea gauza erraza da.
- Testuaren analisi linguistikoan erabiltzen ditugun tresnen sarrera/irteerak formalki definitzera behartzen gaitu.
- Dokumentuen zuzentasun sintaktikoa bermatzeko ez ezik, badago eskura softwarea informazio-bilaketarako, formatu-aldaketak egiteko, iragazleak eratzeko, etab.erako. Eta

aise molda liteke informazioa molde desberdinetan, dela prozesaketarako, dela inprimatzeko, dela pantailaratzeko, dela web-ean argitaratzeko, edo, baita beste hizkuntza batzuetan emateko ere.

- Azkenik, anotazio linguistiko banatuari esker, aukera izango dugu analisi-multzo desberdinak (segmentazioak, analisi morfosintaktiko osoak, lematizazio-emaitzak, etab.) testu tokenizatu bakarrari loturik gordetzeko, analisi jakin bat fisikoki behin bakarrik gordez.



II.4 irudia.- Lematizatzailearen irteera: dokumentu-amaraunaren adibide bat.

III Analizatzaile morfologikoaren doikuntza

Lematizatzaile/etiketatzaile bati begira funtsezkoa da analizatzaile morfologiko sendo batean oinarritzea. Baina sendotasuna funtsezko ezaugarria izanik ere ez da bakarra; aitzitik, eraginkortasunarekin, zuzentasunarekin eta zehaztasunarekin lagunduta ez badator lortuko den produktuaren kalitatea eta erabilgarritasuna oso eskasa izan daiteke. Kapitulu honetan estaldura, eraginkortasuna eta zuzentasunaren hobekuntzarako burututako lana aurkeztuko da.

Analizatzaile/sortzaile morfologikoa, lematizatzaile/etiketatzailerako ez ezik, lengoaia naturalaren prozesamendurako beste aplikazio askotarako oinarrizko tresna da. Hori dela eta, sendotasunak eta eraginkortasunak eragin zuzena izango dute gainerako tresnen emaitzetan.

Lan honen abiapuntua, aurreko kapituluan adierazi den bezala, Alegriaren tesi-lanean (1995) deskribatutako euskara estandarrerako analizatzaile/sortzaile morfologikoa da. Lehenengo fase horretan tresnaren estaldura-tasa %95ekoa zen. Emaitza horren arrazoi nagusienetakoa estandarizazio-prozesua oraindik ere berria zela eta, euskalkien erabilera idazkera-mailan oraindik ere oso hedatua zegoela. Gaur egunean, hizkera-mailan euskalkien erabilera zabaldua den arren, idatzizko lan gehienetan euskara batuaren erabilera ezartzea lortu da, eta, hori dela eta, erabilera dialektalen proportzioa jaitsi da. Hala ere, ezin daiteke hitz ez-estandarren tratamendu morfologikoaren beharrik ez dagoela esan. Aurrerago ikusiko den bezala, nahiz eta testua estandartzat jo, EDBLn landu gabe dauden izen bereziak agertzea zilegi da, baita hitz berriak eta termino teknikoak ere. Horiekin batera, ez-jakintasunak, fonetikak eragindako erroreak eta arauen aldaketek eragindakoak ere kontuan hartu behar dira testuak prozesatzerakoan.

Hori dela eta, II. kapituluan esan bezala, analizatzaile morfologikoaren eraikuntzaren bigarren fasean euskara estandarrerako prozesadore morfologikoak ezagutzen duen hitz-multzoa —estaldura-tasa— handitzen duen analizatzaile sendoa garatu da.

Bi faseetan erabili diren teknikak bi mailatako morfologian (Koskenniemi 1983) daude oinarrituta, eta horri esker sistema osoa homogeneoa da, irtenbide partikularretatik aldenduz. Hiru hobekuntza burutu dira bi mailatako formalismoaren inguruan: lehenengoz erabiltzaileen lexikoen erabilera bideratu da, bigarrenik bi mailatako paradigmaren erabilpen berri bat egin da, aldaera dialektal eta gaitasun-desbideratzeak deitu ditugun forma ez-estandarren tratamendurako (Aduriz *et al*. 1994); eta azkenik hizketaren tratamenduan (Black *et al.* 1991) erabilia zen lexikorik gabeko analisia testuen analisirako izan da hedatua, morfologian oinarritutako asmatzailea (*guesser*) osatuz.

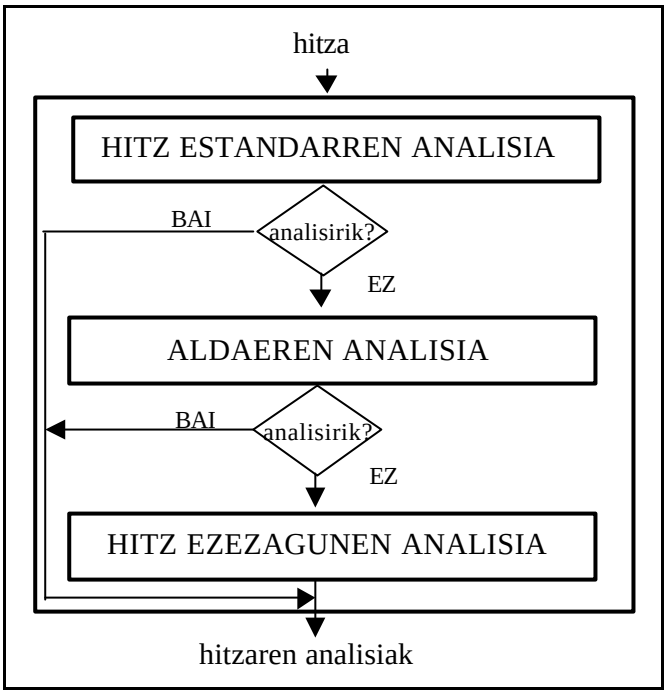
Beraz, analisi morfologikoa hiru urratsetan eta inkrementalki gauzatzen da. Lehenengoak hitz estandarrak analizatzen ditu. Modulu honetan Euskaltzaindiaren arau eta erabakiak jarraitzen dituzten hitzak tratatzen dira, EDBLn landutako lexikoan oinarrituz. Baina erabaki hauen hedapenak denbora behar izaten duenez eta hainbat erabilera dialektal oso zabaldurik daudenez, aldaera dialektal eta gaitasun-desbideratzeak tratatzeko analizatzailea izango da bigarrena. Azkenik, aurreko bi urrats hauetan analizatu gabe geratu diren hitzak tratatzeko lexiko-sarrerarik erabiltzen ez duen hirugarren analizatzaile morfologikoa erabiltzen da. Hiru analizatzaileen bidez, MORFEUSen estaldura %100ekoa izatea lortu da, zuzentasuna %99,5etik gorakoa izanik, III.1 taulan ikus daitekeen bezala.

(a)	**DT**	**AR**	**I/A**	**I/T**	**R**
estandar	%77,91	%80,72	3,81	3,27	**%99,73**
aldaerak	%1,74	%81,83	4,47	3,84	**%92,26**
ezezagunak	%2,65	%100	18,09	18,09	**%98,33**
testu-hitzak	%82,30	%81,37	4,39	3,75	**%99,52**
batez beste	%100	%66,96	4,39	3,27	**%99,61**
(b)					
estandar	%78,76	%81,13	3,82	3,29	**%99,75**
aldaerak	%0,70	%74,00	4,14	3,32	**%70,00**
ezezagunak	%3,04	%100	19,42	19,42	**%99,54**
testu-hitzak	%82,50	%81,76	4,53	3,89	**%99,51**
batez beste	%100	%67,45	4,53	3,38	**%99,60**

III.1 taula.- Anbiguotasun neurriak analizatzaile morfologikoaren irteeran.

Bestalde, analizatzailearen abiadura hobetzea ezinbestekoa da euskararen prozesamendurako gainerako tresna askoren eraginkortasuna onargarria izan dadin. Analizatzailea egoera finituko teknologia baliatuta inplementaturik dagoenez, prozesuaren abiadura hobe daiteke egoera finituko tresnen liburutegi eraginkorragoa erabilita. Horregatik,

inplementazioa egokitu da *Inxight*-en[1] garatutako *fst* liburutegia erabiltzeko eta, horrela, analizatzailearen abiadura hobetzeko.



III.1 irudia.- Analizatzaile inkrementalaren egitura.

Gainera, analizatzailearen hasierako inplementazioa erabiltzean hiru prozesu abiatu behar ziren bata bestearen atzetik sarrerako hitz guztiak tratatzeko. Baina inplementazio aldaketa dela medio, hiru analisi faseak prozesu bakar batean integratu dira.

Zuzentasuna hobetzeari begira, erroreen iturburu desberdinak aztertu behar dira. Batetik, gaur egunean gehienbat euskarri magnetikoan dauden arren, bildutako testu asko eskaneatuak izan dira, eta OCR programaren akatsak direla medio analisi zuzenak lortu ezin direla esan behar da. Errore horietako batzuek hitzak nahastera edota tokenizazio-erroretara eramaten dute MORFEUS. Tokenizazio-akatsak, testuak zuzenduz besterik ezin dira ekidin, euskara prozesatzeko OCR programa bat erabilita, esate baterako. Bestetik, analisi-prozesua modu inkrementalean egitetik datozen erroreak daude, urrats batean tratatutako hitzak ez direlako beste moduluen bidez analizatuko. Azken hauek dira atal honetan aztertuko direnak.

III.1 taulan ikusten denez, aldaeren analizatzailea da zuzentasun txikiena duena hiruen artean. Aldaeren kasuan, anbiguotasun-neurriak estandarrenetatik nahiko hurbil dauden arren, zuzentasunean galera handiak daude. Erroreak aztertuz gero %80 inguru izen berezietan egin dela ikusi da. Hitz horiek aldaera edo gaitasun-desbideratze gisa analizatu direnean, bestelako

[1] Inxight Software, Inc., Xerox taldeko enpresa bat da (www.inxight.com).

interpretazioak lortzen dira. Esate baterako, *Bruselas* (*Brusela*) analizatzean, lema gisa *bruselaza* ematen du eta *El Pais* prozesatzerakoan *heldu* eta *bahi/baia*[2] (instrumental mugagabeaz) lemak ematen ditu.

Lehenengo adibidearen kasuan, errorea konpontzeko datu-basean *Bruselas Brusela*ren aldaera gisa lantzea nahikoa da —*El, de* eta *del* hitzak erdaratako izen berezien parte gisa (*BST* etiketa) lantzea ere komeni da—, eta *Pais* bezalako hitzak ondo analizatzeko akats gehien sortzen dituzten erregelen aplikazioa murriztea nahikoa izan daitekeela ikusi da, z/s aldaketa morfema bukaeran ez onartuz, amaierako *–s* hori instrumentalaren morfemarekin ez nahasteko.

Hemen aipatutako aldaketa gehienak burutu dira, lexikoari dagozkion batzuk izan ezik. Esate baterako, *El* ez da oraindik datu-basean landu, baina beharrezkoak diren lexiko-aberasketak aztertu eta burutuko dira. Edozein kasutan, izen berezi konposatuetan agertzen diren era horretako erdal hitzak, hitz anitzeko unitateen tratamendua egiterakoan izendun entitatearen osagai gisa identifikatu eta tratatuko dira.

Baina, gainerako errore gehienak analisi inkrementala egiteagatik sortzen dira. Izan ere, aldaeren analizatzaileak interpretaziorik esleitzen dionean, ez da bestelako aukerarik aztertzen. Analisi hori lortzeko, ordea, askotan aldaeratik lema lortzeko aldaketa handiak egiten dira, zuzena ez den analisia erauziz.

Hala eta guztiz ere, ezin da aldaeren analizatzailea baztertu, lexikorik erabili gabe analizatuz gero, aipatutako kasuetan izen bereziaren analisia ziurtaturik izango zuketen arren, modulu horrek kategoria irekiak besterik erabiltzen ez dituenez, zenbait aldaera gaizki analizatuko lirateke, adibidez hainbat determinatzaileenak —*batzu*, *hoiek*.

Ondorioz, zuzentasun-neurriak hobetu nahi badira, kontuan izan behar da analizatzaileak inkrementalki aplikatzeak zenbait errore sortzen dituela. Berez, analisi estandarrean gertatzen diren errore askoren iturria hauxe da. Zenbait kasutan, izen bereziekin gehien bat, maiztasun txikiko lemak edota erabilera murritzeko morfemak erabilita interpretazio okerra lortu eta bestelako interpretaziorik ez da bilatuko. Adibide argi bat *Barak* izena da. Bere horretan agertzen denean, *bara* izen arrunta ematen du, baina *Barakek* agertzean hitz ezezagunen analizatzaileak tratatzen duenez, lema zuzena ematen da.

---

[2] *El* hitzaren kasuan, *h*-ren galera aurreikusten duen erregela aplikatu da. *Pais*-en kasuan, berriz, amaierako *s*-n *z/s* nahasketa izan dela eta *Pai* zatian *b/p* aldaketa eta bokalen arteko *h*-a galduta *bahi* lematik datorkeela eta, bestetik, *b/p* aldaketa eta amaieran berezko *a* galduta *baia* lematik datorkeela interpretatzen du.

Aipatutako akatsak lehenengo interpretazio-multzoa jasotakoan gelditzearen ondorioz gertatzen direnez, murriztapena erlaxatuz horietako batzuk ekidin daitezkeela ikusi da. Gai honen inguruan egindako hobekuntzak kapitulu honen III.2 atalean aurkezten dira.

III.1 Eraginkortasuna eta estaldura areagotzeko hobekuntzak

Analizatzaile morfologikoa euskararen oraingo eta etorkizuneko prozesamendu automatikorako oinarrizko tresna da. Analizatzailean oinarritutako tresnen artean EUSLEM lematizatzaile/etiketatzailea, GaIn Intranet-eko bilatzailea (Aizpurua *et al*. 2000) edota bertsolaritzarako laguntza-tresna (Arrieta *et al*. 2001) daude. Hori dela eta, garrantzitsua da batetik, testuko unitate guztiak prozesatzeko gai izatea —estaldura %100 izatea, alegia— eta, bestetik, erantzun azkarra emateko gai izatea, bereziki Internet-en bidezko zerbitzuak eman nahi direnean —eraginkorra izatea, hain zuzen—.

Analizatzailea bi mailatako formalismoan oinarrituta dago. Morfologia konputazionalerako bi mailatako eredua Koskenniemi-k (1983) proposatu zuen eta onarpen-maila handia lortu du, gehienbat aplikazio orokorrekoa delako, erregelen adierazgarritasunagatik eta ezagumendu linguistikoa eta programaren arteko banaketa argiagatik. Fonologia sortzailearekin duen diferentzia nagusiena maila lexikoa eta azaleko mailaren artean bitarteko egoerarik eza da. Bi mailatako ereduan hitz baten analisia egiteko azaleko formari dagozkion errepresentazio lexikal zuzenak aurkitzean datza. Alderantziz, sorkuntzak errepresentazio lexikal jakin batetik abiatuta berari dagozkion azaleko formak bilatzen ditu. Ereduaren konplexutasun konputazionalari buruzko azterketa sakona (Barton 1985) lanean egiten da eta hizkuntzaren konplexutasunak analisi edota sintesiaren abiaduran eragin esanguratsurik ez duela ondorioztatzen du.

Morfologiarako bi mailatako ereduak hizkuntza eranskari eta malgukarietan aplikatzeko egokia dela ikusi da (Antworth 1990; Sproat 1992; Oflazer eta Guzey 1994).

Bi mailatako sistemak bi osagai nagusi ditu —ikus Sproat (1992)—:

- Morfemak (lemak, aurrizkiak, artizkiak eta atzizkiak) eta morfemen arteko loturak (morfotaktika) definitzen dituen lexikoia, azpilexikoetan banaturik. Lexikoko sarrera bakoitzak morfotaktikari buruzko informazioa erazagutzen du azpilexikoen multzoa den jarraitze-klasearen bitartez. Azpilexikoak (erpinak) eta jarraitze-klaseak (arkuak) konbinatuz morfotaktika grafoa definitzen da. Morfemei dagozkien 75.000tik gora sarrera definitu dira 170 azpilexikoetan antolaturik.

- Erregela-multzoa, maila lexikoa eta azalekoaren arteko mapaketak kontrolatzen dituen erregela-multzoa. Osagai hau eraldaketa morfofonologikoak deskribatzeko ezinbestekoa da. Erregelak morfotaktikarekiko independenteak dira eta definitu diren bi mailatako erregelak 24 dira.

Analizatzailea transduktoreak erabilita dago inplementatuta. Transduktore lexikal bat (Karttunen 1994) automata finitu bat da, trantsizio bakoitzean azaleko mailako ikur bat maila lexikaleko ikur batekin mapatzen duena. Hortaz, bi mailatako morfologiaren eboluzioa kontsidera daiteke, ondoko ezaugarri hauek dituena:

- Kategoria morfologikoak forma lexikalaren parte gisa adierazten da, karaktere anitzeko ikurren bitartez (*Multichar Symbols*). Beraz, diakritikoen erabilera ekidin daiteke.
- Hitz beraren forma deklinatuak hiztegi-sarrera kanoniko berari lotzen zaizkio. Horrek azaleko eta lexikoko formen arteko distantzia areagotzen du. Esate baterako, *hobe* hitza *on* forma kanonikoaren bitartez adierazten da (*on+KONP:hobe*).
- Transduktoreen ebakidura eta konposizioa zilegi da (Kaplan eta Kay 1994). Horrela, lexikoa eta erregelak automata bakar batean integratzen dira, eta maila lexikoa eta azalekoaren arteko aldaketak bi mailatako sistemen kaskada baten bitartez adieraz daitezke. Erregelen arteko ebakidura burutzen da lehenengoz eta, ondoren, maila desberdinen konposizioa egiten da.

Inplementazio honetan analisi- zein sorkuntza-prozesuak oso azkar burutzen dira eta deskribapen morfologiko osoak espazio txikiagoan gorde daiteke, analizatzaile zein sortzailearen eraginkortasuna modu esanguratsuan hobetuz (Alegria *et al*. 2001). Inplementaziorako, esan bezala, *Inxight*-en garatutako *fst* liburutegia erabili da (Karttunen eta Bessley 1992; Karttunen 1993; Karttunen *et al*. 1996; Bessley eta Karttunen 2002).

Estaldura handitzeko, hiru transduktore eraiki dira eta ondorengo ataletan deskribatzen dira (Alegria *et al*. 2001).

III.1.1 Transduktore estandarra

Lehenengo osagaia euskara estandarraren arauak jarraitzen dituzten hitzak analizatu eta sortzeko diseinatu da. Hitz estandarren prozesamendurako transduktoreak hiru osagai dauzka:

- **FST1**: lexikoia. Automata honek maila lexikoa deskribatzen du. Maila lexikoa analisiaren emaitza eta sorkuntzaren sarrera izango da. EDBLtik datorren deskribapen hau konpilatu eta minimizatu ondoren, 1,5 milioi egoera eta 1,6 milioi arku dituen transduktorea bihurtzen da. Bere goiko aldean informazio morfologiko osoaren deskribapena ematen da.

Beheko aldea, berriz, morfemaz eta kaskadako gainerako transduktoreen aplikazioa kontrolatzeko informazio morfologiko minimoaz osaturik dago.

- **FST2**: urruneko mendekotasunak kontrolatzeko murriztapenak. Morfemen arteko dependentzia batzuk ezin daitezke jarraitze-klaseen bidez adierazi, hitzaren barruan fisikoki banatuta daudelako. Baina morfema horien arteko agerkidetzak murriztapenak bete behar ditu. Esate baterako, *bait-* eta *-lako* ezin dira batera erabili —*baitielako* gaizki erabilita dago, *baitie* ala *dielako* erabil daitezke—. Erabilera oker horiek kontrolatzeko modu erraz bat erregela morfofonologikoak erabiltzea da, gure sisteman morfemen konbinaketa batzuk debekatzea nahikoa delako. Erregela hauek maila lexikotik hurbilago aplikatzen dira, erregela-sistema desberdin bat osatuz, morfotaktika eta morfofonologia nahastu gabe. Transduktorea oso txikia da: 26 egoera eta 161 arku.
- **FST3**: erregela morfofonologikoak. Esan bezala, morfemak konbinatzean maila lexiko eta azalekoaren artean gertatzen diren aldaketa morfologiko, fonologiko eta ortografikoak deskribatzeko 24 erregela definitu dira. Erregela hauei buruzko informazio sakonagoa (Alegria *et al*. 1996) lanean aurki daiteke. Transduktorea ez da oso handia baina nahiko konplexua da. 1.300 egoera eta 19.000 arku dauzka.

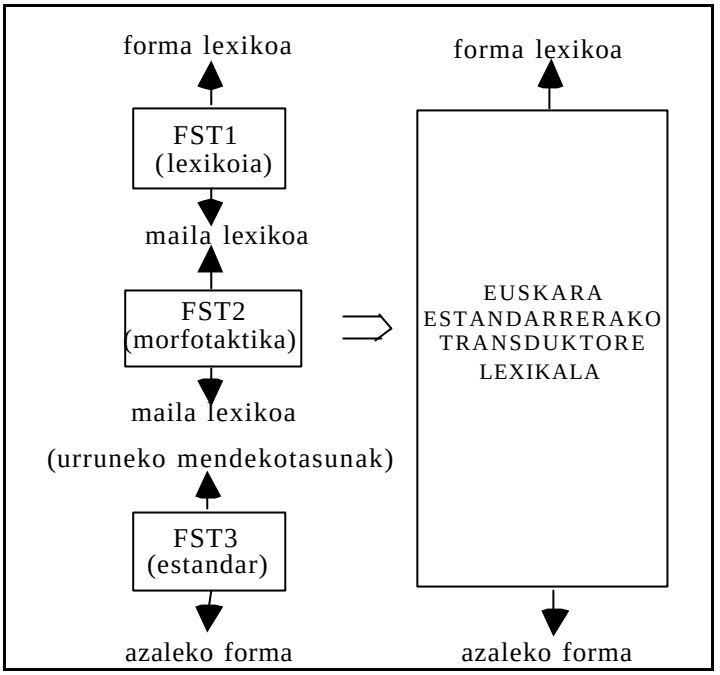
Honakoa izango litzateke *zuhaitzetik* hitzaren emaitza maila bakoitzean:

```
zuhaitz[zuhaitz][IZE_ARR]]+0[DEK_S_M]]+Etik[tik][DEK_ABL]
                          FST1
                   zuhaitz++Etik
                          FST2
                   zuhaitz++Etik
                          FST3
                    zuhaitzetik
```

Hiru transduktoreak konposatu egiten dira analizatzaile estandarra eraikitzeko —ikus III.2 irudia—. Transduktore konposatuak 3,6 milioi egoera eta 3,8 milioi arku ditu, baina 1,9 milioi egoera eta 2 milioi arkura minimizatzen da, diskoan 3,2 Mbyte erabiliz.

3 *IZE_ARR*: izen arrunta, *DEK_S_M* numero singularra, *Etik tik* atzizkia *e* epentetikoarekin, *DEK_ABL*: ablatiboa.

4 FST3-ko erregela batek *e* epentetikoaren gauzatzea kontrolatzen du (ondorengoa erregela sinplifikatua da): `E:e <=> Cons +: +: _` Honen irakurketa "*e epentetikoa gauzatzen da aurrekoa kontsonantea bada*".



III.2 irudia.- Sistema estandarrerako hiru transduktoreen kaskada.

III.1.2 Transduktore hedatua

Bigarren prozesadore morfologikoa aldaera dialektal eta gaitasun-desbideratzeak dituzten hitzak analizatu, normalizatu eta sortu egiten ditu, prozesu morfologikoaren estaldura handitzeko gehitu da. Sistema honek ere hiru osagai ditu:

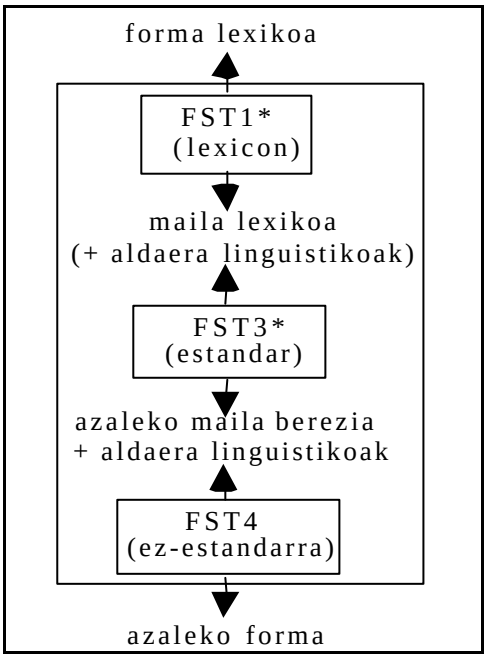
- **FST1***: lexikoia. Lexikoi estandarrari morfema berriak gehituz lortzen da. Morfema berri hauek dagokien morfema estandarrei lotzen zaizkio morfema ez-estandarrak normalizatu edota zuzendu ahal izateko. Horrela, *-tikan* sarrera berria *-tik* morfema estandarrarekin lotuta *etxetikan*, *kaletikan*,... hitz formak analizatu, normalizatu eta zuzentzeko gai izango da prozesadore morfologikoa. 1.500etik gora morfema gehitu dira lexikoi hedatu honetan. Lexikoko morfema batzuei dagokien informazio morfotaktikoan —jarraitze-klasean— gerta daitezkeen aldaketak ere gehitu dira. Lexikoi hedatuaren transduktore konpilatuak 1,6 milioi egoera eta 1,7 milioi arku ditu. Gainera, urruneko mendekotasunei dagozkien murriztapenak ezabatu egin dira, gaitasun-desbideratzeen artean murriztapen horiek ez erabiltzea gerta daitekeelako batzuetan, beraz, FST2 ez da aplikatzen.
- **FST3***: erregela morfofonologiko estandarrak aldaketa txikiarekin. Beheko mailan ez da morfema muga (+ ikurra) ezabatzen FST4n aldaketak kontrolatu ahal izateko. Beraz, maila honetako lengoaia azaleko mailakoa izango da, baina + ikurrekin aberastuta.
- **FST4**: aldaera dialektaletan gertatzen diren maiztasun handieneko aldaketak deskribatzeko erregela berriak. Erregela hauek estandarren egitura eta erabilera berdintsua dute, baina guztiak aukerazkoak dira. Esate baterako, `h:0 => V:V_V:V` erregelak maila lexikaleko

h ikurra azaleko mailan bokalen artean gal daitekeela deskribatzen du. Horrela, *bear* hitza analiza daiteke, *behar* hitz normalizatua lortuz. III.3 irudian ikus daitekeenez, erregela berri hauek azaleko mailatik hurbilago jar daitezke, gehienetan aldaketa fonetikoekin lotutakoak direlako, eta, horregatik, ez dago informazio morfologikoaren beharrik aldaketa horiek tratatu ahal izateko.

Honakoa izango litzateke *zuaitzetikan* hitzaren emaitza maila bakoitzean:

```
zuhaitz[zuhaitz][IZE_ARR]]+0[DEK_S_M]]+Etikan[tik][DEK_ABL]
                        FST1*
                  zuhaitz++Etikan
                        FST3*
                  zuhaitzetikan
                        FST4
                  zuaitzetikan
```

FST1* eta FST3* transduktoreen arteko konposizioa transduktore estandarraren antzekoa da egoera eta arku kopuruetan, baina FST4 gehitzean, egoera kopurua 12 milioietara eta arku kopurua 13,1era igotzen dira. Dena dela, 3,2 milioi egoeratara eta 3,7 milioi arkuetara minimizatzen da, diskoan 5,9 Mbyte erabiliz.

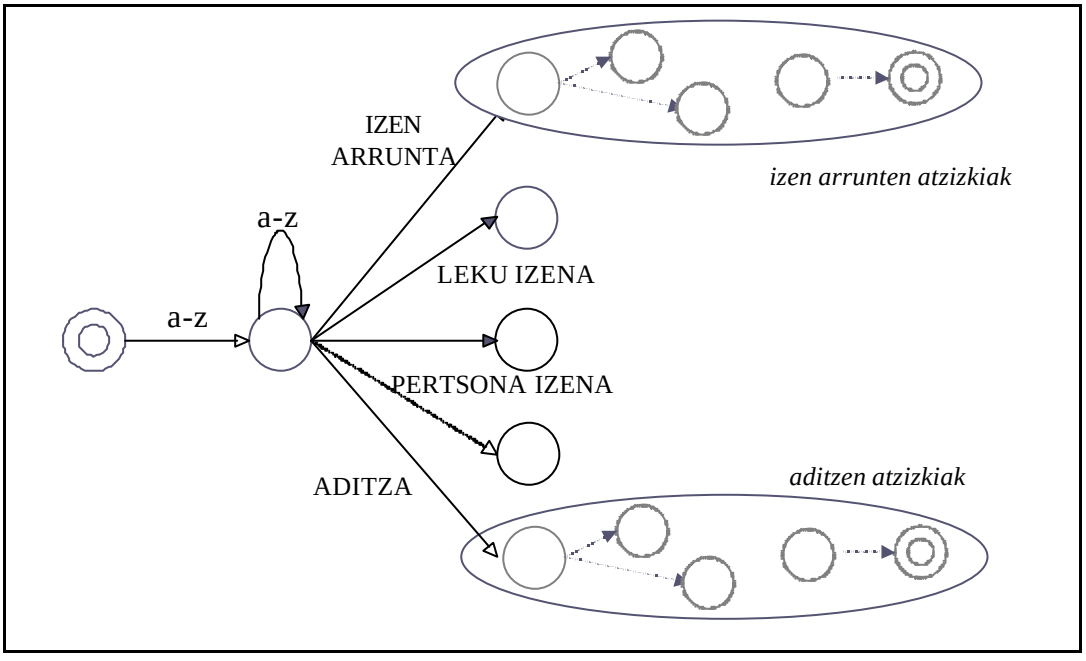


III.3 irudia.- Sistema hedaturako hiru transduktoreen kaskada.

III.1.3 Transduktore orokorra

Aurreko transduktorearekin ez da hitz ezezagunen arazoa ebazten. Horretarako transduktore orokorra diseinatu da, lexikoian lemak eduki beharra erlaxatzen duena. Hasierako bertsioan (Alegria *et al*. 1997), transduktore hau ahotsaren sintesian erabilitako teknika batean oinarritzen zen (Black *et al*. 1991), baina hurbilpen honetan transduktorea sinplifikatu egin da. Geroago, Daciuk-ek (2000) antzeko ebazpena proposatu du hitz ezezagunak tratatzeko (*guessing automaton*), baina automataren eraikuntza konplexuagoa da berak proposatutakoa.

Transduktore berria estandarrean oinarritzen da baina honako aldaketa eginda: lexikoia lema generikoak besterik ez dituzten kategoria irekietara[5] mugatzen da. Erregela estandarrak, berriz, bere horretan mantentzen dira. Beraz, erregela estandarren sistema (FST3) lexikoi txikiarekin (FST0) konposatzen da. Lema generikoak ikur alfabetikoak konbinatuz lortzen dira eta lexikoian karakterez osatutako azpilexiko zikliko baten bidez adieraz daiteke. Azpilexiko horretan hainbat murriztapen ezartzen dira letra maiuskulak eta minuskulak onartzeko kategorien arabera. III.4 irudian ikus daiteke lexikoi txikiari (FST0) dagokion grafoa.



III.4 irudia.- Lexikoi txikiaren grafo sinplifikatua

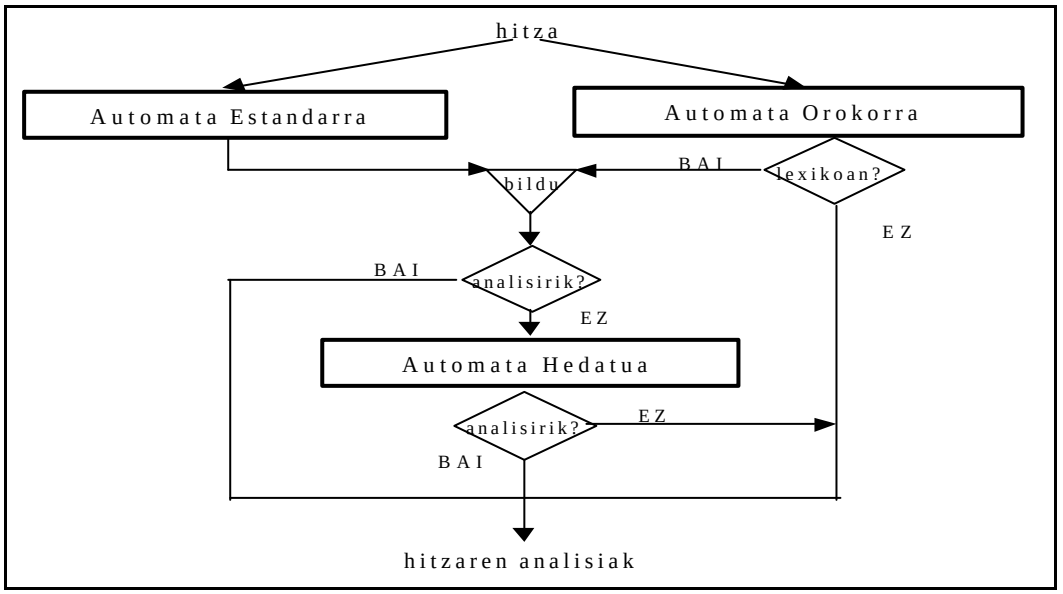
[5] Zazpi kategoria ireki daude, garrantzitsuenak honakoak izanik: izen arruntak, pertsona izenak, leku izenak, adjektiboak eta aditzak.

Transduktore konposatua oso txikia da, 8.500 egoera eta 15.000 arku inguru ditu. Emaitzako analisi bakoitzean lema posiblez gain, morfema guztiei dagokien informazio morfologiko osoa ematen da, beste asmatzaileen baino emaitza osatuagoa izanik.

Transduktore orokorra analisiaren bi faseotan erabiltzen da: analisi estandarrean eta lexikorik gabeko analisian (*guesser* moduan). Erabiltzailearen lexikoa dagoenean, sarrerako hitza transduktore estandarrean prozesatzearekin batera, erabiltzailearen lexikoan dagoen begiratu beharko litzateke. Printzipioz, erabiltzailearen lexikoan dauden sarrerak estandartzat jo beharko lirateke, baina horrek deskribapen estandarrarekin batera konpilatzea suposatuko luke. Hori ekiditeko, transduktore orokorra erabiliko da, eta honek itzulitako lema posibletako bat erabiltzailearen lexikoan aurkitzen bada, transduktore estandarraren emaitzekin batera emango dira transduktore orokorrak itzulitako interpretazioak. Bigarren erabilera transduktore estandarrak eta hedatuak emaitzarik ematen ez dutenean aktibatuko da, hitz ezezagunei interpretazioak esleitu ahal izateko.

III.1.4 Transduktoreen aplikazioa

Aurreko ataletan aurkeztutako automata guztiak prozesu bakar batean aplikatzen dira eta III.5 irudian adierazi bezala aplikatzen dira. Lehenengo urratsean hitza estandar gisa ezagutzen ahalegintzen da analizatzailea. Hitza estandar kontsideratzeko, bi aukera daude: lehenengoa, automata estandarrak ezagutzea eta, bigarrena, erabiltzailearen lexikoan egotea. Horregatik, hitza bi automata horiei pasatzen zaie eta bien emaitzak bildu egiten dira. Modu honetan analisirik lortzen ez bada, automata hedatuari ematen zaio hitza eta, honek ere emaitzarik lortzen ez badu, automata orokorraren analisiak itzultzen dira.



III.5 irudia.- Analizatzaile inkrementalaren funtzionamendua.

III.2 taulan ikus daitekeenez, token gehienak lehenengo moduluak tratatzen ditu baina testu motaren arabera banaketa aldatu egiten da. Esate baterako, 2. corpusa EEBS proiektuan jasotako testu-bilduma orekatua da eta neurri handi batean euskara batuaz idatzita dago. 4.a, berriz, bizkaieraz eta gipuzkeraz idatzitako testuak jasotzen dituen bilduma horren azpimultzo bat denez, aldaeren analizatzaileak askoz ere hitz gehiago tratatzen ditu[6] —%11,5 tratatuz, besteetan %2 denean— eta hitz ezezagunen kopurua ere beste corpusena baino handiagoa da.

	tokenak	**estandar**	**aldaera**	**ezezagun**	**beste**[7]
corpus1	1.288.257	%78,44	%0,94	%3,80	%16,82
corpus2	587.515	%74,98	%2,03	%2,92	%20,07
corpus3	148.333	%77,91	%1,01	%6,23	%14,85
corpus4	29.939	%60,54	%11,50	%7,90	%20,06

III.2 taula.- Tokenen banaketa corpus desberdinetan.

Bestalde, 1. eta 3. corpusak *Euskaldunon Egunkaria*tik jasotakoak dira eta, euskara estandarrean idatzirik egon arren, gainerakoetan baino hitz ezezagun gehiago agertzen dira, kazetaritzan agertu ohi diren gaurkotasuneko pertsona- eta leku-izenak lexikoan ez daudelako.

Izen bereziekin osatutako lexiko berezitua sortuz gero, *Euskaldunon Egunkaria*tik jasotako testuetako hitz gehienak estandar gisa tratatuko lirateke, hitz ez-estandarren proportzioa %1 eta %2 artekoa izanik.

Laburbilduz, atal honetan aurkeztutako hobekuntzak direla medio, analizatzaile/sortzaile morfologikoaren estaldura osoa lortu da —sarrerako hitz guztiak tratatzen dira—, baina zuzentasuna aztertuz gero, emaitzak ez dira baliokideak analisi-mota guztietan. III.1 taulan aurkeztu den bezala, automata hedatuak tratatzen dituen hitzetan, zuzentasun-maila estandarrak prozesatutako hitzetan baino nabarmenki okerragoak dira, batez ere egiaztapenerako corpusean.

II. kapituluan aurreratu den bezala, egiaztapenerako corpusaren zati mardulena *Euskaldunon Egunkaria*tik lortutako da. Horregatik, aldaeren proportzioa askoz ere txikiagoa da —%0, 7 egiaztapenerako corpusean eta %1,74 erreferentzia-corpusean—. Baina kasu askotan ez dira benetan aldaerak, oker analizatutako izen bereziak baizik. Zuzentasunaren

6 Testu hauek tratatzeko EDBLn landuta ez dauden maiztasun handieneko aldaera dialektalak lexiko berezitu batean landu dira zuzentasuna hobetzearren, testu hauek UZEIn garatzen ari den EEBS proiekturako automatikoki lematizatu baitira.

7 Talde honetan puntuazio ikurrak eta bestelako bereizgarri eta identifikadoreak sartu dira.

emaitzei erreparatuz gero, aldaeren analizatzaileak hirutik batean huts egiten duela ikus daiteke, beste analizatzaileek ehunetik batean baino gutxiagotan huts egiten duten bitartean.

Emaitza hauek motibatuta, zuzentasuna hobetzeko aukerak aztertu dira, beti ere, erreferentzia-corpuseko akatsetan oinarrituta, egiaztapenerako corpusa aztertu gabe. Ondorengo atalean azterketa horren ondorioz inplementatutako zuzentasuna hobetzeko proposamen zehatza aurkezten da.

III.2 Zuzentasuna hobetzeko proposamena

Kapituluaren hasieran aipatu den bezala, analisi morfologikoa modu inkrementalean egiteak zenbait akats dakartza, gehienbat izen berezien kasuan. Analisi estandarraren akatsak aztertuz gero, ikus daiteke %75 izen berezietan gertatzen direla, eta aldaeretan, aldiz, izen berezi eta siglak kontuan hartuta %80 biltzen dituztela.

Izen bereziek akatsen multzo handiena osatzen badute ere, badira testuaren prozesamendurako zuzentzen garrantzitsuak diren beste akats batzuk. Multzo honetan analisi estandarren bat eman baina erabilera handiko aldaera diren batzuk daude, *bat* determinatzailearen mugagabezko erabilerak kasu. Determinatzaile honen erabilera duela gutxi bateratu denez, oraindik orain testuetan *batetan* edo *batetako* modukoak aurkitzea ez da harritzekoa. Baina determinatzailearen analisiaren ordez izen arruntarena hartzen dute *bate* lematzat dutelarik, eta horrek bai maila morfologikoan zein sintaktiko edo semantikoan akats nabarmenak ekar lezake.

Azkenik, bestelako akatsen proportzio txiki bat datu-basearen aberasketa eskatzen dutenak dira eta gainerakoak testu-akatsak dira, erabilera okerrek eragindakoak.

Soluziobideak jorratzerakoan, akatsak tipifikatu dira. Izen bereziei dagokionean, akats guztiak ez dira mota berekoak. Badira izen arrunt edota adjektibo gisa analizatzen direnak, hala nola, *Ezkurdi*, *Gari*, *Timo*, *Saralegi* eta (*Henriette*) *Aire* estandarren artean eta *Xalba*, (*El*) *Greco*, *Diez(en)* eta *Velez* aldaeren artean. Baina badira leku izen berezi gisa analizatzen direnak baina pertsona izen berezi ere izan daitezkeenak eta alderantziz, esate baterako, *Cabanillas*, *Gerediaga*, *Olabarri* edota *Zuluaga*. Azkenik, badira aditz laguntzaile, trinko zein aditz sinple gisa analizatzen direnak ere, *Die* (*Tageszeitung*), *Da* (*Vinci*) eta *Eros* izenak adibidez.

Izen arrunt bezala analizatzen diren izen berezientzat soluzio posible bat, analizatzaile estandarraren kasuan, lemen maiztasuna kontuan hartzea litzateke, maiztasun txikiko aukera

besterik ez dagoenean, analisian aurrera jarraituz hurrengo urratsetan interpretazio zuzena lortzearren.

Horretarako maiztasun handiko hitzen zerrenda sortu da, zerrenda horretan dauden hitzak, maiuskulaz idatzirik agertu arren, izen berezi gisa kontsideratuak ez izateko. Zerrenda hori osatzeko, EEBS corpuseko 500.000 hitz inguru analizatu eta desanbiguatu ziren EUSLEM erabilita eta maiztasun handieneko 500 hitzak aukeratu ziren. Horietatik letra larriren bat zutenak eta luzera 2koa edo txikiago zutenak kendu ziren. Honela, 450 inguru hitzek osatutako zerrenda erabili da.

Izen berezi mota bat bakarra ematen dutenentzat bi soluzio posible plantea daitezke. Batetik, datu-basean izen berezi horiek bikoiztea, agerikoa delako toponimia adierazten duten izen gehienak —guztiak ez badira ere— izen edo abizen gisa erabil daitezkeelako. Baina orduan izen berezi gisa erabil daitezkeen izen arruntekin —*Zuhaitz*, *Ekaitz*, *Hodei*…— ere gauza bera egin beharko litzateke, eta orduan ez legoke hitza desanbiguatzeko modurik.

Bestetik, analisiaren emaitza zuzentzeko metodoren bat plantea daiteke. Adibidez, izen berezi konposatuak identifikatzerakoan aipatutako aukerak aurreikustea eta *Iñigo Olabarri* agertzen bada, lehenengoa pertsona izena izanik, ondoren datorrenarekin batera pertsona izen bezala tratatu, leku izenaren interpretazioa soilik izan arren. Hauxe bera izan daiteke aipatutako aditzen kasurako soluzio posible bat ere.

Analisi estandarra jaso baina aldaera direnak —*bat* determinatzaileaz gain, badira adibide gehiago ere, *kaxoi* izena esaterako— ekiditeko oso baliagarria izango da II. kapituluan aipatutako arrarotasun-eremua. Horrela, hitz baten interpretazio morfologiko guztietan arrarotasun marka agertzen bada, analisi horiek baztertu eta analisian aurrera jarrai dezake analizatzaileak. Hala ere, aldaeren analizatzaileak interpretazio ez-arrarorik ematen ez badu, estandarrak emandakoak mantenduko dira, hitz ezezagunen analizatzaileak emandako interpretazioak beti ere estandarrarenak baino arraroagoak izango direlako.

Aldaeren analizatzailearen kasuan, akatsak ekiditeko soluzio posible bat aldaketa kopurua kontrolatzea litzateke, hau da, interpretazioak lortzeko erregelen aplikazio kopuru minimo bat gainditzen bada, aukera horiek baztertu eta lexikorik gabe analiza daiteke, gainerako hitz ezezagunak bezala. IV.2 atalean azaltzen den bezala, aplikatutako erregela kopurua analisiarekin batera edizio-distantzia kalkulatzen duen funtzio baten bitartez neurtuko denez, posible da parametro baten bidez kopuru hori kontrolatu behar dela adieraztea.

Baina kopuru hori ezartzea ez da erraza. Oso kopuru txikia jarriz gero, berez aldaera diren hitzak lexikorik gabe analizatuko lirateke, neurrigabeko anbiguotasunaren igoera ekar lezakeelarik. Aitzitik, handiegia jarriz gero, ez litzateke hobekuntzarik lortuko. Beraz, hori da aztertu beharreko lehenengo datua. Bestalde, jakinik akats gehienak izen berezietan gertatzen

direla, hitzaren idazkerari erreparatuz anbiguotasuna ahalik eta neurri txikienean gehituko litzateke, IV. kapituluan aurkeztuko den desanbiguazio tipografikoan erabili direnen ildotik.

Azkenik, beste aukera bat automata probabilistikoak erabiltzea litzateke (*http://www.research.att/sw/tools/fsm*), erregela bakoitzari aldez aurretik neurtutako pisu bat esleituz probabilitate txikia duten interpretazioak baztertzearren. Baina azken aukera hau aztertzea etorkizunerako utzi da.

Beraz, analizatzailearen zuzentasuna hobetzearren zenbait akats programazio lan handirik egin gabe zuzen daitezkeela ikusi, eta, laburbilduz, honako soluziobideak inplementatu dira:

- Hitz estandar baten interpretazio guztiak arraroak badira, aldaera gisa analizatu. Horrela ere arraro ez den analisirik lortzen ez bada, zegoen bezala utzi. Bestela, estandarrak aldaerekin ordezkatu, aplikatutako erregela kopurua ondoren aipatzen den muga gainditzen ez badu.
- Hitz estandar bat letra larriz idatzita badago, izen berezi edota sigla analisirik ez badu eta maiztasun handiko hitza ez bada, hitz ezezagun gisa analizatu eta pertsona-izen berezi eta leku-izen berezien analisiak gehitu analisi estandarrei.
- Aldaera baten interpretazioak lortzeko erregela kopuruak minimo bat gainditzen badu, analisi horiek baztertu eta hitz ezezagun gisa analizatu.

Bestalde, erregela kopurua finkatzeko aldaeretan aplikatutako kopuruak aztertu dira. Aldaera dialektalak direnetatik asko datu-basean landuta daude eta gainerako gehienak erregela bat bakarra aplikatuz lortzen da analisia. Horregatik, bi erregela edo gutxiago aplikatuz lortutakoak aldaera kontsideratuko dira eta gainerakoak hitz ezezagunen analizatzaileak tratatuko ditu. Badira salbuespenak, hala ere. Izan ere ezezagunen analizatzaileak kategoria irekiak besterik ez ditu erabiltzen eta horregatik, kategoria itxiren bat ematen badu, badaezpada aldaeren analisiak hobetsiko dira.

Aldaketa hauek aplikatuta erreferentzia eta egiaztapenerako corpusetan lortutako emaitzak III.3 tauletan ikus daitezke. Emaitza hauek hasierakoekin —III.1 taula— alderatzean hainbat ondorio atera daitezke:

- Erroreen erdiak ekidin daitezke, prozesuaren konplexutasuna handitu gabe. Egiaztapenerako corpusaren kasuan gehiago ekiditen dira, %65 inguru.
- Aldaeren analizatzaileak, hobekuntzak burutu arren, oraindik ere emaitzarik okerrenak ematen ditu. Gaur egungo testuetan aldaera gutxi izan ohi da, hizkuntza estandarra erabiltzen denean behinik behin. Hala ere, analizatzaile honen beharra dago analizatzaile orokorra nahi bada. Aldaeren kasuan lortzen da hobekuntzarik handiena, %60

erreferentzia-corpusean eta %72 egiaztapenerakoan. Hau ulertzekoa da, hasierako emaitzetan errorerik gehien hitz-multzo honetan egiten baitzen.

- Egiaztapenerako corpusean nabarmenago egiten da aldaeren hobekuntza, aldaketak diseinatzeko kontuan hartu ez den arren. Corpus honetan *Euskaldunon Egunkaria*ri dagokion zatian aldaera gisa analizatzen diren hitz asko izen bereziak dira. Horregatik, analizatzaile hedatua erabiltzean, horietako asko ondo analizatzea ahalbidetzen da.
- Anbiguotasuna areagotu egiten da bi arrazoigatik: izen bereziak ondo identifikatu ahal izateko egindako aldaketengatik, eta aldaera gisa tratatzen diren hitzek lexikorik gabekoek baino analisi gutxiago dituztelako batez beste (eta aldaera gisa tratatzen ziren batzuk orain lexikorik gabe tratatzen dira). Eragozpen hau zuzentzeko IV. kapituluan azaltzen diren teknikak aplikatuko dira.

(a)	DT	AR	I/A	I/T	R
estandar	%77,87	%81,03	3,88	3,33	**%99,88**
aldaerak	%1,66	%83,39	4,65	4,04	**%96,84**
ezezagunak	%2,77	%99,80	18,25	18,21	**%98,41**
testu-hitzak	%82,30	%81,71	4,48	3,85	**%99,77**
batez beste	%100	%67,25	4,48	3,34	**%99,81**
(b)					
estandar	%78,65	%81,54	3,92	3,38	**%99,91**
aldaerak	%0,67	%72,92	4,83	3,79	**%91,67**
ezezagunak	%3,18	%100	19,46	19,46	**%99,56**
testu-hitzak	%82,50	%82,19	4,66	4,00	**%99,83**
batez beste	%100	%67,80	4,66	3,48	**%99,86**

III.3 taula.- Analizatzaile morfologiko hedatuaren emaitzak.

III.3 Ondorioak

Kapituluan zehar aurkeztutako hobekuntzak direla medio, analisi morfologikoaren eraginkortasuna eta doitasuna areagotzea lortu da, zehaztasunean askorik galdu gabe. Laburbilduz, honakoak dira lortutako helburuak:

- Eraginkortasuna hobetzea: batetik, aipatu den bezala, analizatzailearen abiadura modu esanguratsuan hobetu da, eta, bestetik, prozesu kopurua ere gutxitu da. Hala ere, testu bat analizatzen den bakoitzean analizatzaileak transduktoreak kargatu behar zituen eta horixe da kostu handiena duen prozesuaren atala. Hori dela eta, analizatzailea bezero/zerbitzari ereduari jarraituz inplementatu da, kostu hori aurreztearren.

- Hitz estandarren analisia hobetzea: izen berezi askok bestelako analisi estandarrak onartzen dituztenez, izen hauek modu egokian tratatu ahal izateko ezinbestekoa da analisi hedatua aplikatzea. Hala ere, oso zaila gertatzen da lema egokia ematea, horregatik, lema anitz sortu ohi da izen berezi hauen kasuetan. Anbiguotasuna modu kontrolatuagoan handitzeko lemak aukeratzeko teknikaren bat aplikatzea komeni da, ikasketa automatikoan oinarrituta edota aipatutako Mikheev-en *Document Centered Approach* (1999, 2000-a) hurbilpena erabiliz.
- Aldaeren analisia hobetzea: testu estandarretan orokorrean oso aldaera gutxi agertzen diren arren, batzuk oraindik orain hedapen handia dute. Analizatzaile hedatuari esker testuetako izen berezi asko aldaera gisa ez tratatzea lortu da. Hala ere, erreferentzia-corpusean *Euskaldunon Egunkaria*ko testu gehiago sartu beharko lirateke aldaeren tratamendua doitzeko eta emaitzak hobetzeko.

Hobekuntza hauen ekarpena VI. kapituluan ebaluatzen da, batetik, zuzentasunean lortutako hobekuntzak desanbiguazio morfosintaktikoa burutzerakoan zenbateraino laguntzen duen, eta, bestetik, zehaztasunean galdutakoak eragina duen.

IV Hitz ez-estandarren tratamenduaren hobekuntza

Aurreko kapituluetan ikusi denez, hitz ez-estandarren analisi morfologikoan gainsorkuntza handia gertatzen da, zehaztasuna galduz. Hitz ezezagunen kasuan arazoa bereziki larria da, zehaztasuna hitz estandarrena baino 6 aldiz txikiagoa delako.

Anbiguotasuna handitzearekin batera, analisi morfologikoan oinarriturik dauden aplikazioen zuzentasunaren galera etor daiteke. Beraz, anbiguotasuna lehenbailehen murriztea komeni da, ahalik eta errore gutxien gehituz.

Kapitulu honetan bereziki hitz ez-estandarren anbiguotasuna murrizteko diseinatutako tratamendua deskribatuko da. Tratamenduak ez du testuingurua kontuan hartzen[1], hitzaren izaerari besterik erreparatzen ez dioten desanbiguazio-prozedurak diseinatu direlarik.

Honenbestez, kapitulua modu honetan dago antolaturik: lehenengo atala hitz ez-estandarren problematika planteatzeari eskaintzen zaio. Ondoren, hitz ez-estandarren anbiguotasuna mugatzeko diseinatutako prozedurak aurkezten dira[2]. Hirugarren atalean, aldiz, aurreko kapituluan aurkeztu den analizatzaile hedatuari prozedurak aplikatzearen emaitza aztertzen da eta azkenik, zenbait ondorio atera eta etorkizunerako hainbat hobekuntza proposatzen dira.

1 Salbuespen bakarra dago diseinatutako prozeduretan, aurrerago aurkezten den izen berezien desanbiguazio, hain zuzen. Tokenaren testuingurua baino, token horren testuko agerpenak kontuan hartzen ditu.

2 Diseinua zein hasierako ebaluazioa analizatzaile morfologiko inkrementalaren gainean burutu da.

IV.1 Hitz ez-estandarren problematika

MORFEUS sarrerako testua morfosintaktikoki analizatzeko erabiltzen da. Hori lortzeko hiru modulu nagusi ditu. Lehenengo moduluak, aurreprozesua deiturikoak, sarrerako testua unitate edo tokenetan banatzen du eta analizatzaile morfologikoari bidali behar zaizkionak bereizten ditu, hau da, testu-hitzak direnak. Bigarren moduluak, berriz, testu-hitzak morfosintaktikoki tratatzen ditu. Eta, hirugarrenak, hitz anitzeko unitateen prozesamendua egiten du.

Beste hizkuntza askotarako ez da analizatzaile morfologikoaren beharrik ikusten eta hiztegi-bilaketa nahikoa izaten da oinarrizko informazio morfologikoa esleitzeko, hau da, etiketa morfosintaktikoa hiztegitik lortzen da. Hitz ezezagunen tratamendua orokorrean nahiko sinplea izaten da, baina ez beti.

Hasierako etiketatze-sistemetan amaierako *n*-gramak (luzera finkoko atzizkiak) erabili izan dira hitz ezezagunei etiketak emateko (Church 1988; Cutting *et al.* 1992). Honako ezaugarriak kontuan hartzen dira sistema gehienetan:

- Letra larriz hasten den hitza (*Capitalization*). Ezaugarri honek izen berezia dela adieraz dezake eta horretarako erabiliko da.
- Atzizkiak. Askotan ez dira morfema osoak izaten, baina kategoria erabakitzeko esanguratsuak izan daitezke. Normalean karaktere kopuru maximoa ezartzen da.
- Aurrizkiak eta hitzen amaierak. Atzizkien informazioa aberasteko erabili ohi dira.

Informazio hori baliatuta hitz bati dagozkion etiketak erabakitzeko maiztasun-neurria edota testuinguru-erregelak erabiltzen dira. Kasu askotan eskuz idatzitako erregela gutxi batzuk nahikoa izango dira hitz ezezagunen etiketak aukeratzeko.

Hala ere, lan hori ekidin daiteke metodo automatikoak aplikatuz (Brill 1995; Mikheev 1996a; Mikheev 1996b; Thede 1998) lanetan bezala. Brill-ek (1995) ikasketa-corpusetik automatikoki erauzten ditu hitz ezezagunen etiketak esleitzeko beharrezkoa den testuinguru-informazioa. Gainera, corpus hori ez da eskuz markatu behar. Modu honetan, hitz ezezagunen tratamendua ikasketa ez-gainbegiratuaren bidez burutzen da.

Mikheev-ek (1996a), aldiz, hiztegi orokorra erabiltzen du atzizki, aurrizki eta hitz-bukaeren erregelak erauzteko eta testu gordinekin egiaztatzen ditu emaitzak, erregelei pisuak esleituz. Pisu minimora iristen ez diren erregelak biltzen ditu ahal den neurrian minimora irits daitezen. Emaitzak hobetzearren, atzizki, aurrizki eta bukaeretan gerta daitezkeen aldaketa morfofonologikoak aplikatzen ditu (Mikheev 1996b). Gainera, lexikoan ez dauden laburdura eta izen bereziak identifikatzeko aipatutako *Document Centered Approach* (Mikheev 2000-a) erabiltzen du (ikus II.2.1.1 atalean).

Thede-ren (1998) kasuan ikasketa-corpusean aurrizki eta atzizkiek etiketen arabera duten banaketa estatistikoan oinarritzen da etiketa-multzoa esleitzeko, etiketa kopuru txikiena duen aurrizki-atzizkirik luzeena aukeratuz.

Teknika sofistikatuagoak ere erabil daitezke hitz ezezagunei dagozkien etiketak esleitzeko. Batzuek analisi morfologikoarekin batera egiten dute. (Cha *et al*. 1998) lanean korearrerako morfema asmatzailea aurkezten dute, morfemen patroien hiztegi bat eta morfemen konektibitatearekin lotutako arauak erabiliz. (Daciuk 1999) lanean, aldiz, polonierarako eta frantseserako probatu den hurbilpena aurkezten du, gure asmatzailearen antzera, atzizkien informazio morfologikoa egoera finituko sistema batean integratuz. Hala ere, hurbilpen honetan, hitza atzekoz aurrera tratatzen da, atzizkien informazioa ez dena lema izango delarik.

Beste batzuek, berriz, desanbiguazioarekin batera erabakitzen dute zein etiketa esleitu hitz ezezagunei (Weischedel *et al*. 1993; Daelemans *et al*. 1996; Orphanos eta Christodoulakis 1999), inguruan dituen hitzen informazioa baliatuz. Weischedel-en taldeak (Weischeldel *et al.* 1993) hitz ezezagunei etiketa esleitzeko desanbiguazio-algoritmoan kasu berezia du, aipatutako informazioaz gain, hitz ezezagunak testuinguru horretan etiketa ireki bakoitza izateko probabilitatea ere erabiliz. (Daelemans *et al.* 1996; Orphanos eta Christodoulakis 1999) lanetan erabaki-zuhaitzak erabiltzen dira atzizki, aurrizki eta hasierako letra larriari dagokion informazioa egituratzeko. Lehenengo lana ingeleserako egin da eta bigarrena, berriz, greziera modernorako.

Azkenik, bada sintaxiarekin batera tratatzen dituenik ere. (Barg eta Walther 1998) lanean, HPSG bidez lexikoan ez dauden hitzak tratatzen dira, agerpen kopurua eta bere testuinguruak kontuan izanik, lexikoan dauden beste hitzen antza duten erabaki eta hitz berriak dagokien informazioarekin batera lexikoan gehitzen dira. Hala ere, ez dute aipatzen sintaxira iritsi arte nola tratatzen diren. Hurbilpen hau lexikoa automatikoki eta etengabe aberasteko aurkezten da, lexikoa beti irekia baita.

Kasurik gehienetan, nahikoa da hitz-formari dagokion etiketa-multzoa esleitzea, ez da bestelako informazio morfologikorik esleitu beharrik, ezta lemaren beharrik ere. Aztertutako lanetatik bakarrean lortzen da lema zein gainerako morfemen informazio morfologiko guztia (Daciuk 1999). Baina zenbait hizkuntzatan, euskara barne, ezinbestekoa da informazio morfologiko xehea eta lema lortzea.

Euskararen kasuan, orokorrean hitz ez-estandarrak batez beste hitzen %4-%8 badira ere, zuzentasun-falta eta anbiguotasuna hitz estandarrena baino handiagoa da. Anbiguotasun handiaren zergatia bai aldaeren tratamendua bai lexikorik gabeko analisia gainsortzaileak izatean datza. IV.1 taulan III.1 taulako emaitzak ekartzen dira gogora.

Ikus daitekeenez, hitz ezezagunen kasuan batezbesteko interpretazio kopurua bereziki altua da, testu honetan analisi gehien dituenak 61 izanik erreferentzia-corpusean[3]. Horrela, zenbait hitzen kasuan hainbeste interpretazio posible izanik, zaila gertatzen da bai hitz hori bera, baita ingurukoak ere desanbiguatzea eta horrek, emaitzen zuzentasuna gutxitu egiten du.

(a)	**AR**	**I/A**	**I/T**	**R**
estandar	%80,72	3,81	3,27	**%99,73**
aldaerak	%81,83	4,47	3,84	**%92,26**
ezezagunak	%100	18,09	18,09	**%98,33**
testu-hitzak	%81,37	4,39	3,75	**%99,52**
batez beste	%66,96	4,39	3,27	**%99,61**
(b)				
estandar	%81,13	3,82	3,29	**%99,75**
aldaerak	%74,00	4,14	3,32	**%70,00**
ezezagunak	%100	19,42	19,42	**%99,54**
testu-hitzak	%81,76	4,53	3,89	**%99,51**
batez beste	%67,45	4,53	3,38	**%99,60**

IV.1 taula.- Anbiguotasun-neurriak analizatzaile morfologikoaren irteeran.

Proiektu hau garatzerakoan testuingururik gabeko desanbiguazio-prozesu bat erabiltzearen beharra antzeman zen, bai aldaeren tratamenduan bai lexikorik gabeko analisian sortutako zenbait analisi lehenbailehen baztertu ahal izateko.

Tratamendu honen helburua gainsorkuntzak eragindako gehiegizko analisiak testuingurua kontuan hartu gabe baztertzea da. Horrek anbiguotasuna gutxituko du baina kontu handiz egin beharko da analisi egokia baztertua gera ez dadin. Aipaturiko gainsorkuntza horren arrazoia bikoitza da:

- Aldaeren tratamendurako eransten diren atzizkiak eta erregela morfofonologikoak: elementu hauen bidez aldaerak aurreikusten dira baina zenbait kasutan orokortzea gehiegizkoa gerta daiteke eta horren ondorioz gainsorkuntza eragin. Adibidez, *kaletikan* analizatzean -*tikan* atzizkia lortzen da (-*tik* atzizkiaren aldaera) baina lema posibletzat *kale* eta *kala* lortzen dira. *kala* lortzearen arrazoia hauxe da: *a* itsatsiaren kasurako desbideratze-erregela bat dago aurreikusita, kasu honetan aplikatzea gehiegizkoa izan arren.

[3] Beste testu batzuk aztertzean 100etik gora analisi ere eman izan ditu analizatzaile orokorrak, egiaztatzeko erabiltzen den testuan bertan.

- Lexikorik gabeko analisian egiten den orokortzea: kategoria bakoitzeko analisi bat sortzen dela ziurtaturik dago baina gehienetan analisi asko sortzen da kategoria bakoitzeko. Horren arrazoi nagusia lema da, lema posible bat baino gehiago sor daitekeelako informazio morfologiko berbera duten analisientzat —ikus B eranskinean B.1 ataleko adibidea—.

Arrazoi bakoitzarengatik tratamendu berezitua burutzen da, kualitatiboki arazo desberdinen aurrean baikaude. Tratamendu hauek hasierako MORFEUSen aurreikusi baziren ere, lehen hurbilpen bat baino ez zen. Oraingoan zuzentasuna handitzearren tratamendua birplanteatua izan da desanbiguazio morfosintaktikoaren emaitzak hobetzearren. Gaiaren inguruan burututako lana hurrengo atalean aurkezten da.

IV.2 Anbiguotasuna mugatzeko tratamendua

Esan bezala, gainsorkuntza dela medio, bai aldaerek bai hitz ezezagunek hitz estandarrek baino interpretazio gehiago dituzte. Tratamendu honen helburuak bi dira:

- Lemen anbiguotasuna ahal den neurrian ebaztea, VI. kapituluan deskribatzen den desanbiguazio morfosintaktikoak ez baitu lemen artean aukeratzea helburu.
- Lehenengoarekin batera, ahal den neurrian interpretazio morfologiko batzuk baztertu.

Horretarako, sarreran aipatu den moduan, aldaera eta hitz ezezagunetarako prozedura berezituak diseinatu dira, eta ondorengo ataletan deskribatzen dira, aldi berean emaitzen ebaluazioa eginez.

IV.2.1 Aldaeren tratamendua

Aldaeren analisietan hitz baterako analisi bat baino gehiago lortzen denean desanbiguazioari ekitea erabaki dugu, bertan aukera "estandarrena" lortzearren. Hori egin ahal izateko zenbait informazio eskuratu behar da aldaeren tratamendutik: aurkitutako morfema ez-estandarrak eta aplikatutako erregela ez-estandarrak. Erregelei buruzko informazioa eskuratzeko helburuarekin ukitu batzuk egin ziren programa eta datuetan.

Baina, eraginkortasuna hobetzearren garatutako transduktoreen bidezko inplementazioan, horrelakorik ez da ematen analisiaren emaitza gisa, hori dela eta aplikatutako erregela

kopurua edizio-distantzia[4] kalkulatzen duen funtzio baten bitartez lortu da[5]. Hitz estandarra eta aldaera karakterez karaktere konparatzen dira honako funtzioa aplikatuz:

$$ed(a[i+1],b[j+1]) = \begin{cases} i = 0 & \Rightarrow j \\ j = 0 & \Rightarrow i \\ a[i+1] = b[j+1] & \Rightarrow ed(a[i],b[j]) \\ a[i] = b[j+1] \wedge a[i+1] = b[j] & \Rightarrow 1 + min \begin{pmatrix} ed(a[i-1],b[j-1]) \\ ed(a[i+1],b[j]) \\ ed(a[i],b[j+1]) \end{pmatrix} \\ a[i+1] \neq b[j+1] & \Rightarrow 1 + min \begin{pmatrix} ed(a[i-1],b[j-1]) \\ ed(a[i+1],b[j]) \\ ed(a[i],b[j+1]) \end{pmatrix} \end{cases}$$

Distantzia kalkulatzeko sarrerako hitz-forma eta bere baliokide estandarra behar dira. Hori lortzeko estandarrari dagokion informazioarekin sorkuntza morfologikoa egitea nahikoa da eta bi mailatako sistemak analisi zein sorkuntzarako balio duenez, nahikoa da analisia jasotakoan baliokide hori eskuratzea. Adibidez, *suaitxetikan* forma analizatzerakoan *zuhaitz+tik* segmentazioa emango du morfema bakoitzari dagokion informazio morfologikoz horniturik. Sortzaileak *zuhaitzetik* forma emango du eta bien arteko edizio-distantzia gisa 5 ematen du funtzioak (*z:s, h:0, z:x* aldaketak leman eta *0:a, 0:n –tik* eta *–tikan* atzizkien arteko distantzia).

Hala ere, esan bezala, *-tikan –tik* atzizkiaren aldaera gisa landuta dago datu-basean, beraz, bien arteko distantzia zerotzat hartzea erabaki dugu. Hori konpontzeko, lexikotik datozen morfema/lemak eta dagozkien aldaerak analizatzaileak ematen duen informazio morfologikoarekin batera datozenez, aldaera eta estandarraren arteko distantzia kalkulatu eta goiko emaitzari kentzea nahikoa da. Adibide berarekin jarraituz, *-tik* eta *–tikan* morfemen arteko distantzia 2 denez (*0:a, 0:n*), edizio-distantziari bi kendu eta gero kontuan hartu beharreko distantzia hirukoa izango litzateke, eta kopuru hori bat dator aplikatutako erregelen kopuruarekin (*z:s, h:0, z:x* erregelak). Orokorrean, honako formula kalkulatu behar da erregela kopurua lortzeko:

erregelak (estandar, aldaera) = *ed* (estandar, aldaera) - $\sum_{\text{landutako morfemak}}$ *ed* (estandar, aldaera)

Behin distantzia kalkulatuta, lexikotik ez datozen lemen aldaera sortu behar da, lematizazioa ere helburu delako, eta horretarako erregeletan kontuan hartzen diren aldaketak aplikatuko zaizkio lema estandarrari sarrerako hitzarekin bat etortzeko.

---

4 Funtzioa programazio dinamikoaren bitartez inplementatu da modu eraginkorrean.

5 Transduktoreen bitarteko inplementazioan, morfologia malgua erabilita, aplikatutako erregelak zeintzuk diren jakin daiteke, transduktoreen kaskadan zehar gainerako informazioarekin batera erregelei buruzko informazioa gehituz, Oflazerrek —komunikazio pertsonala— proposatu bezala, baina oraindik aldaketak burutzen ari direnez, edizio-distantzia erabiliko dugu.

Zenbait kasutan zaila da lema eta atzizkien arteko muga hori bestelako informaziorik gabe zehaztea, horregatik aldaera sortzean zenbait akats sortzen dira. Akats nabarmenenak aditzetan gertatzen dira, lema gisa partizipioa eman ohi delako. Kasu gehienetan partizipioa *-tu* morfemaz eratzen da, baina zuhur jokatu behar da *-i* edota *-n* bukaerak ere izan daitezkeelako.

Edizio-distantziaren informazioa erabilita aldaeren desanbiguazioa lortzeko erabiltzen den algoritmoa aplika daiteke. Algoritmoaren funtsa honako hiru puntu hauetan laburtzen da:

- Erabilpen ez-estandarren kopurua kalkulatzen da eta kopuru txikiena duen analisia aukeratu egiten da.
- Kopuru berdineko kasuetan erregela ez-estandarren erabilpena zigortu egiten da morfema ez-estandarren erabilpenen gainetik, azken hauek kasu partikularragoak direlako eta gainsorkuntzarako eragin txikiagoa dutelako. Beraz, aplikaturiko erregela ez-estandar gutxien duen analisia aukeratzen da. Erregela kopuru berdina dutenen artean, etiketa[6] bakoitzeko bat mantenduko da.
- Hala eta guztiz ere, aurreko bi urratsak iraganda analisi bat baino gehiago geratuko balira, horiek guztiak mantenduko dira irteeran. Hau gertatzeko probabilitatea oso txikia da, eta gure corpusetan ez da horrelakorik gertatu.

Azkenik, baztertu beharreko analisiak filtro batetik pasatzen dira. Filtro honek maiztasun handiko lema estandarrak dituzten interpretazioak mantendu egiten ditu. Izan ere, lema laburrak direnean atzizkian akatsen bat agertuz gero, hitz berri baten lema ere sortu daiteke eta azken kasuan erregela gutxiago aplikatuko dira. Filtro honetarako *Euskaldunon Egunkaria*ko 800.000 testu-hitzetatik 5 hizki baino gutxiagoko eta 100 aldiz baino gehiagotan agertutako lemak aukeratu dira.

Ondoren *batzuk* determinatzailearen aldaera den *batzu* hitzaren adibidea ikus daiteke. Erabilera ez-estandarra denez analizatzaileak aldaera eta idazketa-errore posible guztiak bilatzen ditu interpretazio posible gisa, hala nola *p:b, s:z, x:z* eta *o:u*.

```
/<batzu>/
    "batzuk"   /batzu/   1 LEMA/MORFEMA
    "pa+tsu"   /ba+tzu/  2 ERREGELA
    "pa+txo"   /ba+tzu/  3 ERREGELA
```

Prozedurak, kasu honetan, lehenengo interpretazioa aukeratuko luke, gainerakoak ez bezala, aldaera dialektala hobesten duelako idazkera erroreen aurrean.

6 Erabili beharreko etiketa-multzoa parametro gisa ematen da. Defektuz, 2. mailakoa, kategoria eta azpikategoria, erabiltzen da.

Tratamenduaren emaitzak IV.2(a) taulan ikus daitezke. Prozedurak gutxitan egiten du huts, %1,5ean 1. mailan eta %1,4 inguruan 2. eta 3. mailetan[7], baina analisietatik %43 inguru baztertzen ditu. Zuzentasunean %1,9 eta %2 bitartean galtzen den arren, aldaera gisa tratatzen diren hitzak testuko tokenetatik %2 baino gutxiago izanik, prozesu osoari gehitzen dion errorea %0,03 ingurukoa besterik ez da. Batez beste, analisi bat kentzen du aldaera bakoitzeko. Egiaztapenerako corpusari dagokionean, IV.2(b) taulan ikus daitekeenez, ez da errorerik gehitzen eta hitzeko analisi bat kentzen da.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%81,83	4,47	3,84	**%92,26**
ondoren (3. maila)	%75,36	2,97	2,48	**%90,36**
ondoren (2. maila)	%75,36	2,97	2,48	**%90,36**
ondoren (1. maila)	%75,20	2,95	2,47	**%90,21**
(b)				
aurretik	%74,00	4,14	3,32	**%70,00**
ondoren (3. maila)	%60,00	2,87	2,12	**%70,00**
ondoren (2. maila)	%60,00	2,80	2,08	**%70,00**
ondoren (1. maila)	%60,00	2,73	2,04	**%70,00**

IV.2 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

Erreferentzia-corpusean 2. eta 3. mailetan emaitza berberak lortzen dira, ez dagoelako kasu-mailan bereizi eta erregela kopuru ezberdina aplikatua duen hitzik. Egiaztapenerako corpusean, aldiz, hiru mailetan desanbiguazio-tasa desberdinak lortzen dira akatsik egin gabe. Aipatzekoa da prozedura hau diseinatzerakoan erreferentzia-corpuseko adibideetan oinarriturik egin dela eta, hortaz, egiaztapenerako corpuseko emaitzek diseinuaren egokitasuna konfirmatu besterik ez dutela egiten. Ondorengo adibideetan ikus daitezke maila desberdinetan bereizten diren kasuak (beltzez prozedurak baztertzen dituen interpretazioak):

[7] Jarraian argitzen da adibideen bitartez 2. eta 3. mailan emaitza berberak lortzearen arrazoia.

```
/<untzi>/
        ("huntz"  /untz/   IZE ARR + DEK DAT)      1 ERREGELA
        ("ontzi"  /untzi/  IZE ARR + DEK ABS)      1 LEMA/MORFEMA
        ("ontzi"  /untzi/  IZE ARR)                1 LEMA/MORFEMA
        ("untxi"  /untzi/  IZE ARR + DEK ABS)      1 ERREGELA
        ("untxi"  /untzi/  IZE ARR)                1 ERREGELA

/<dandai>/
        ("danda"  /danda/    IZE ARR + DEK DAT)    1 ERREGELA
        ("tanda"  /danda/    IZE ARR + DEK DAT)    1 ERREGELA
        ("tantai"  /dandai/  ADJ IZO + DEK ABS)    2 ERREGELA
        ("tantai"  /dandai/  ADJ IZO)              2 ERREGELA
        ("tantai"  /dandai/  IZE ARR + DEK ABS)    2 ERREGELA
        ("tantai"  /dandai/  IZE ARR)              2 ERREGELA

/<Axun>/<HAS_MAI>/
        ("azun"  /axun/  ADJ IZO + DEK ABS)        1 ERREGELA
        ("azun"  /axun/  ADJ IZO)            1 ERREGELA
        ("asun"  /axun/  IZE ARR + DEK ABS)        1 ERREGELA
        ("asun"  /axun/  IZE ARR)            1 ERREGELA
        ("Asua"  /Axu/   IZE LIB + DEK INE)        2 ERREGELA
```

Lehenengo adibidean, 3. mailan ikus daiteke izen arrunt absolutiborako eta kasurik gabeko izen arrunterako bi lema daudela (*ontzi* eta *untxi*) eta bigarrenaren interpretazioak baztertuko dira. Hala ere, *ontzi* lexikoan landuta dagoenez, besteen aurrean hobetsiko da, beraz, lehenengo interpretazioa ere baztertuko da. Bigarren adibidean, 3. mailan interpretazio guztiak mantendu beharko lirateke, etiketa bera dutenei (*danda* eta *tanda*) erregela kopuru berbera aplikatu zaielako; 2. mailan, berriz, hiru lema desberdin daude izen arruntaren interpretazioarentzat (*danda*, *tanda* eta *tantai*) eta azkenari bi erregela aplikatu zaizkionez, bere bi analisiak baztertuko dira. Hirugarren adibidean, 3. eta 2. mailetan ezin da lemarik baztertu, baina 1. mailan, izenari dagozkion bi lema daude (*asun* eta *Asua*) eta bigarrenaren interpretazioa baztertuko da. Edozein kasutan, azken adibidean interpretazio guztiak okerrak dira, beraz, ez du errorerik gehitzen.

IV.2.2 Hitz ezezagunen tratamendua

Funtzio hau betetzeko lau prozedura diseinatu dira, eta lauek helburu bera badute ere hurbilpen desberdinari jarraitzen dio bakoitzak. Lehenengoak idazkerari erreparatzen dio interpretazioak baztertzeko, maiuskulaz idatzitako hitzen lemen artean aukeratuz. Bigarrenak eratorpenaren bidez analiza daitezkeen hitzak desanbiguatzen ditu. Hirugarrenak, aldiz, corpus osoan agertzen diren izen bereziak identifikatzen ahalegintzen da. Eta, laugarrenak, azkenik, etiketa morfologikoak kontuan hartzen ditu lema posibleen artean aukeratzeko.

Prozedura hauek bata bestearen atzetik aplikatzeko diseinatu dira, batak bestearen emaitzak birfintzeko. Izen berezien identifikazio eta desanbiguazioa (3. urratsa) hautazkoa da,

baina gainerakoak beti aplikatuko dira. Ondoren, prozedurak deskribatzen dira zehatzago eta azkenik guztien konbinaketa posible batzuen emaitzak aurkezten dira.

IV.2.2.1 Desanbiguazio tipografikoa

Prozedura hau diseinatzeko, interpretazio kopuru handiena duten hitzak aztertu dira eta gehienak osorik ala hitzaren zati bat maiuskulaz idatzirik daudela ondorioztatu ahal izan da. Horren arrazoia sinplea da, analizatzaile morfologikoak izen, adjektibo eta aditz gisa interpretatzeaz gain, izen berezi eta siglen analisiak ere esleitzen dizkio halako kasuetan.

Beharbada analizatzaileak horrelako analisietako batzuk ekidin beharko lituzke, maiuskulaz lehenengo hizkia besterik ez duten siglen kasuak esaterako. Aldaketa burutzean, kendu beharreko analisietako batzuk ekidin diren arren, eman beharreko beste batzuk ez dira ematen. Horregatik, heuristiko hau erabiliko da hitz ezezagunen idazkeran oinarriturik bere interpretazio kopurua jaisteko.

Zenbait kasutan aukera horietako batzuk bazter daitezke idatzita dauden moduan oinarrituz. Esate baterako, hitza maiuskulaz idatzi eta bukaera minuskulaz dagoenean, gehienetan maiuskula hutsez dagoen zatia lemari eta minuskulaz dagoen zatia deklinabide atzizkiari dagokie. Ondorioz, maiuskulaz idatzita dagoen zatia lema gisa ez duten analisiak bazter daitezke.

Honen adibide *CARTIER-BRESSONen* hitza da, analisi morfologikoaren irteeran batezbestekotik gora interpretazio duen hitzetako bat dena —48 guztira. B eranskinean (B.1) analisi morfologiko guztiak ikus daitezke, baina laburbilduz lema posible gisa *cartier-bressonen*, *cartier-bressone*, *cartier-bressona*, *cartier-bresson* izango dira, sigla, izen arrunt, leku-izen, pertsona-izen edota adjektibo gisa, kasu bakoitzean bukaera modu desberdinean interpretatzen delarik. Idazkerari erreparatuz, lema *Cartier-Bresson* behar duela ikus daitekeenez, interpretazio kopurua 10era jaitsi daiteke, 5 sigla moduan harturik eta beste 5 izen berezi moduan hartuz —2 pertsona izen eta 3 leku izen gisa interpretatuz, B eranskinean (B.2) ikus daitekeenez—.

Sigla gisako interpretazioei dagokionean, prozedura beste era batean bideratzen da. Izan ere, analizatzaile morfologikoak maiuskulaz hasitako hitz oro sigla gisa interpretatzen du, baina logikaren arabera sigla denean, dagokion lema oso-osorik maiuskulaz idazten da. Beraz, maiuskulaz hasiera besterik ez duten hitzetatik sigla interpretazioak bazter daitezke, B eranskinean (B.3-B.4) agertzen den *Valentine* hitzaren kasuan bezala[8]. Gainerakoan, sigla

[8] Kasu honetan, gainera, esaldi erdian agertzen denez, izen berezien interpretazioak utziko dira soilik.

izan daitekeenean, aurretik azaldutako irizpide berbera erabiliko da siglen artean lema aukeratzeko, aurreko adibidean egin den bezala.

Prozedura honen emaitza nahiko ona da, desanbiguazio-tasa %45ekoa baita eta irteeran 10 interpretazio gutxiago baitaude hitzeko (ikus IV.3(a) taula). Prozedurak baztertutako analisien %0,17an huts egiten eta hitz ezezagunen zuzentasunean %1,66 galtzen badu ere, analisi morfologikoari gehitzen dion errorea %0,04koa besterik ez da.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%100	18,09	18,09	**%98,33**
ondoren	%99,79	8,23	8,22	**%96,67**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
ondoren	%100	7,17	7,17	**%99,08**

IV.3 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

Bigarren corpusean, aldiz, 12 interpretazio baztertzea lortzen du, nahiz eta prozedura lehenengoari erreparatuz diseinatu den. Desanbiguazio-tasa %63koa da, prozesu osoari gehitzen dion errorea %0,01ekoa izanik.

IV.2.2.2 Eratorpena

Eratorpena (eta konposizioa) nahiko emankorra eta irregularra da euskaraz, neologismoen sorkuntzan oso erabilera hedatua izanik. Eratorpena 100etik gora morfemen bidez (atzizkiak gehien bat) gauzatzen da. Eratorriak sortzeko erabiltzen diren kategoriak izenak (*kamioi -> kamioilari*), adjektiboak (*itsusi -> itsuskeria*) eta aditzak (*ekarri -> ekarpena*) dira, eta hauexek berak dira eratorpenak ematen dituen kategoria ohikoenak (Aduriz eta Aldezabal 1995).

Corpusetan agertutako hitz ezezagunetako batzuk (%5-%10) modu honetan sortuak dira, gehienbat testu teknikoetan, bestelako eratorriak lexikoan baitaude. Berriki sortutako hitzen identifikazio eta analisi zuzena ahalbidetzeko, eratorpen atzizki emankorrenak landu dira datu-base lexikalean.

Bestalde, horietako asko, neologismo izatetik lexikalizatuak izatera pasatzen dira denbora eta erabilpenaren ondorioz. Horregatik, askotan, forma batek bi analisi bide izango ditu: batetik, forma lexikalizatua badago, hiztegi sarrera gisa etorriko da eta, bestetik, atzizkiaren bidez sortzen den hitza izango da. Beraz, horrek morfosintaktikoki baliokideak diren analisien gainsorkuntza ekarriko du, anbiguotasun-tasa areagotuz.

Gainsorkuntza horren adibide argi bat *sendotasuna* hitza dugu. Lehenengo analisia *sendotasun* hitza lexikalizatua dagoena da eta bigarrena, berriz, *sendo* adjektiboari *-tasun* atzizkia erantsita sortua dena da:

```
/<sendotasuna>/
    "sendotasun"  IZE ARR + DEK ABS NUMS MUGM
    "sendo+tasun" ADJ IZO + ATZ IZE ARR + DEK ABS NUMS MUGM
```

Adibidea lexikoan landutako sarrera batena bada ere, hitz ezezagunen kasuan gertatzen den gainsorkuntza fenomenoa islatzen da[9]. Maila honetan desanbiguatzeko aurrera eraman den prozeduraren helburua bietatik bakarra uztea da.

```
/<Pepe>/
    "pe+pe"  ADJ IZO + ATZ IZE ARR
    "pe+pe"  IZE ARR + ATZ IZE ARR
    "Pepe"   IZE IZB
    "Pepe"   IZE LIB ...
```

Gainera, prozedura honek eratorpenerako oinarri den lemaren luzera ere mugatzen du hitzen interpretazio arraroak ekiditeko, eratorpena erregularra ez delako. Horren adibide *Pepe* hitza da. Analizatzaileak *–pe* atzizkiaren bidez analizatzen du, *pe* oinarri lema bai adjektibo bai izen gisa hartuz, kasu bietan izen bat emanez. Horiekin batera, *pepe* lema osoa harturik izen eta adjektibo gisa eta, maiuskulaz hasten denez, pertsona- eta leku-izen berezi gisa ere analizatzen da. Eratorrien oinarri lema oso laburra da eta horrelakoek oso maiztasun txikia izan ohi dutenez, eratorrien interpretazioak ezabatu egingo dira.

Azkenik, eratorpen bidez analizatzen bada, gehienetan lortutako lema izango da orokorrean beste kategorietako analisien lemarik probableena. Horregatik, gainerako analisien lemak desanbiguatzeko ere erabiliko da. Esate baterako, *Amatiñok* edo *Abarrategi* izen bereziak eratorpen bidez analiza daitezke eta euren lema eratorpen bidez lortzen dena —sorkuntza aplikatuta— izango da (*amati+ño -> amatiño*, *abarra+tegi-> abarrategi*) eta ez lortzen diren besteak (Amatiñok, Abarrateg, Abarrategia, abarrat, abarrata). Beraz, gainsorkuntza konpontzeaz gain, bestelako interpretazioetan lema egokia aukeratzen laguntzen du prozedura honek.

[9] Hitz estandarren kasuan ez da prozedura hau aplikatzen, baina prozedura aplikatzekotan, egokitu beharko litzateke hitz hauen kasuistika kontuan hartzeko.

(a)	AR	I/A	I/T	R
aurretik	%100	18,09	18,09	**%98,33**
ondoren	%100	17,45	17,45	**%98,33**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
ondoren	%100	19,29	19,29	**%99,08**

IV.4 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

Prozedura hau hitz ezezagunen %5,7ri besterik ez zaio aplikatzen. Batezbesteko analisi kopurua nahiko handia da hitz horietan, 19 interpretazio zehazki. Honen arrazoia nagusia hitzen erdia izen bereziak izatean datza. Desanbiguazio-tasa %60koa da, batere errorerik egin gabe. Irteeran hitzen batezbesteko interpretazio kopurua 8ra jaisten da. Testuko hitz ezezagun guztiak kontuan izanik, batezbesteko interpretazio kopurua ez da asko jaisten (ikus IV.4(a) taula), baina tratatzen diren hitzen desanbiguazioa modu egokian egitea ahalbidetzen duenez, prozedura hau erabiltzea interesgarritzat jo da. Egiaztapenerako testuan, aldiz, bi hitz bakarrik prozesatu dira eratorri gisa, eta horietako bat *Baraki* hitza da. Kasu horretan, lema zuzenari dagokion analisia baztertua izan da (*Barak* lemaren aurrean *Baraki* lema aukeratu delako).

Dena dela, aurrerago ikusiko den moduan, prozedura guztiak konbinatzean horrelako akatsak ekidin daitezke eta ez dute ondoriorik izango. Etorkizunari begira hitz estandarrei ere aplikatzeko aukera aztertuko da.

IV.2.2.3 Izen berezien desanbiguazioa

Prozedura hau corpus bateko izen berezien hautagai-zerrenda erauzteko diseinatu da. Etiketei erreparatuz, bi motatako izen berezi aurki daitezke, pertsona-izen bereziak eta leku-izen bereziak.

Bigarren taldeak duen berezitasunetako bat lehenengo taldeari dagokion deklinabide paradigma ere onartzea da, pertsona-multzoaren zentzua hartzen duenean. Esate baterako, *Donostia* leku izen gisa erabili da *Donostiara noa oporretan* esaldian, baina pertsona-multzo gisa *Donostiak hautatu du* esaldian. Gainera, leku izenak abizen ere izan daitezke, *Aita Donostiak idatzi zuen* esaldian bezala. Hortaz, hitz ezezagunak prozesatzerakoan, izen, adjektibo eta aditz aukerez gain, analizatzaile morfologikoak izen berezi mota bien interpretazioak ere esleituko dizkio maiuskulaz idatzita dauden hitzei.

Beraz, izen berezien identifikazioa oso interesgarria da interpretazio kopurua gutxitzeko. Baina prozedura hau diseinatzeko beste bi arrazoi nagusi ere baziren:

- Batetik, emaitzak lexikoia aberasteko erabil daitezke corpus-maiztasunetan oinarrituz.

- Bestetik, emaitzek lagun dezakete bestelako prozesuetan, hala nola, izendun entitateen identifikazioan (*Named Entities*).

Prozedura honetan dokumentuaren (edo paragrafoaren) informazioan oinarritzen gara, tokenetik haratago aztertzen duen urrats bakarra baita. Prozedurak honako urratsak ematen ditu izen bereziak identifikatzeko:

- Corpusean dauden izen berezi gisa interpretatutako instantzia guztiak bilatu.
- Maiuskulaz idatzita dagoen eta agerpen kopuru minimo bat gainditzen duen hautagai bakoitzeko, konparatu etiketa desberdinekin duen agerpen-kontaketak.
- Aukeratu izen berezi gisa bestelako etiketekin baino instantzia gehiago dituzten hautagaiak.
- Sailkatu, ahal denean behintzat, pertsona-izen berezi (IZB) edota leku-izen berezi (LIB) gisa. Horretarako bi deklinabide kasu erabili dira: genitibo leku-denborazkoa (GEL) LIB aukeratzeko eta ergatiboa (ERG) IZB dela erabakitzeko. Ergatiboa, aurretik esan den bezala, LIBetan ere ematen da, baina gehienetan IZB aukeratuta errore txikia egingo da. Ezin denean sailkatu, bi izen-motei dagozkien analisiak mantendu eta gainerakoak baztertu.

Logikoa den bezala, prozedurak ez ditu esaldi hasieran dauden instantziak bigarren urratseko kontaketa egiterakoan kontuan izango. Hitza puntu baten ostean dagoenean, ezin izango da erabaki izen berezi bati dagokion ala ez. Hala ere, behin erabaki denean zeintzuk diren identifikatutako izenak, puntu ondoren daudenak ere desanbiguatuko dira laugarren urratsari ekiterakoan. Honela, esaldi hasierako hitzak ere desanbiguatu ahal izango dira, dagokien lemaren corpuseko gainerako instantzietan oinarrituta.

Erreferentzia-corpuseko EEBS zatian (27.000 token) 28 izen berezi identifikatu dira, eta *Euskaldunon Egunkaria*ko zatian (9.000 token), askoz txikiagoa izan arren gehiago aurkitu dira, logikoa den moduan, beste 42 izanik. Guztien artean 20 lema anbiguo agertzen dira.

Zuzentasunari dagokionean, 70 izen horietatik 2 izen arruntak dira, 6 siglak eta izen ez-anbiguo batean lema okerra ematen da. Beraz, zuzenki identifikatu direnak orotara 61 izan dira. Horietatik 38 sailkatu ditu prozedurak, 5 kasutan soilik huts eginez. Akats horietatik bi erakunde izenak dira (*Naturgas* enpresa eta *Tageszeitung* egunkari alemaniarra) eta absolutibo eta ergatiboan soilik agertzen dira, hirugarrenaren kasuan (*Kramnik* izen berezia) hiru lema posibleetatik (*Kramn*, *Kramni* eta *Kramnik*) ergatiboa duena aukeratu du (*Kramni*) eta azken bi izenak leku izen bereziak dira (*Muskildi* eta *Ortzaika*). Azkenik, anbiguo zirenetatik 3 sailkapenari esker desanbiguatzen dira, *Kramnik* barne, eta besteetan ondo sailkatu ditu.

Izen berezi hauen agerpen gutxi egon arren, hitz ezezagunetan 8 interpretazio gutxiago egotea lortzen da, IV.5(a) taulan ikus daitekeenez. Desanbiguazio-tasa %48koa da eta %0,42an errorea egin arren, analisiaren prozesuari gehitzen dion errorea %0,09koa besterik ez da.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%100	18,09	18,09	**%98,33**
ondoren	%83,02	10,93	9,24	**%94,90**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
ondoren	%94,04	8,85	8,39	**%88,07**

IV.5 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

Egiaztapenerako corpusean 27 izen berezi identifikatu dira, horietatik 2 sigla eta izen arrunt bat izanik. Gainerakoetatik, erdia anbiguo dira eta 11 sailkatzen ditu, bi anbiguo zirenetatik. Sailkatzerakoan bi errore egiten ditu (*Yitzhak* lema aukeratu ordez *Yitzha* aukeratu du eta *Fein* erakunde izen gisa sailkatu ordez pertsona izen interpretazioak eman dizkio). Sailkatzeko akats horiek alde batera utziaz, gainerako errore gehienak sigletan egin ditu. Hala ere, aurrerago ikusiko den bezala, akats horietako batzuk desanbiguazio tipografikoa aurretik aplikatuta ekidin daitezke.

Emaitzak ikusirik, printzipioz izen gehiago izango duen corpus bati ere aplikatu zaio prozedura. *Euskaldunon Egunkaria*tik nazioarteko berrien 126.000 hitzeko testu-bilduma erabili da. Bertan, 400 izen inguru identifikatu dira, horietatik laurdenaren lemak anbiguo agertzen direlarik. Horren arrazoi nagusia izenen amaiera da. Kasuen %25 inguru *a* bukaera dute eta *a* hori berezkoa denentz ezin erabaki denez, gutxienez bi lema ematen zaizkio. Antzeko zerbait gertatzen da *n* bukaerarekin, baina askoz ere gutxiagotan gauzatzen da. Azkenik, sigla batzuk ere identifikatzen ditu izen berezi gisa eta, errore kontsidera daitezkeenez, hurrengo bertsioetarako prozedura hau siglen identifikazioarekin osatu eta emaitzak hobe daitezke.

Bestalde, analizatzailearen irteeran anbiguotasun-maila handia da eta oso zaila da izen berezien lemen identifikazio zuzena egitea. Horregatik, emaitzak lan handiagorik egin gabe hobe daitezke desanbiguazio tipografikoaren ostean aplikatuz, aurrerago ikusiko den bezala.

Hala ere, prozedura honen aplikazioa aukerazkoa da, hau da, MORFEUSen parametrizaturik dago, agerpen kopuru minimoa finkatu daitekeelarik. Modu honetan, erabiltzaileak corpus handi bat prozesatzerakoan izen berezien hustuketa egin ala ez erabaki dezake, eta bere tamainaren arabera agerpenen behe-muga eman.

Hautazko egitearen arrazoia honakoan datza: izen berezi berbera maizago agertzen da gai bereko testu-bilduma batean testu soil batean edota corpus orekatu batean baino, eta prozedurak bere funtzioa bete dezan, ezinbestekoa da agerpen kopuru minimo batera iristea ebidentzia estatistiko minimo bat izateko hautagai bat ematerakoan.

Prozedura honek izen berezien hautagai zerrenda eman arren, ezin ditu izen horiek leku ala pertsona izen gisa sailkatzen kasu gehienetan, hori askotan ezinezkoa izaten delako testuinguruaren informaziorik gabe edo, are gehiago, informazio semantikorik gabe. Beraz, izen berezi direla jakin arren, oraindik ere anbiguo izan daitezke, baina, gutxienez bestelako interpretazioak baztertzeko gai da.

IV.2.2.4 Informazio morfologikoa eta estatistika

Tratamendu honetan bi arazo bereizi behar dira: etiketen arteko desanbiguazioa eta lemen artekoa. Etiketa anitz uztea ez da arazoa, etiketa horiek desberdinak diren bitartean, aurrerago aurkeztuko den desanbiguazio morfosintaktikorako modulua horren desanbiguazioaz arduratuko baita, testuingurua kontuan hartuz gainera.

Lema anitz mantentzea, aldiz, arazo handiagoa da erabiltzen diren etiketatzaileak ez baitira gai lemen artean bakarra aukeratzeko. Lexikorik gabeko analisia burutzeko modua proposatu zutenean, Black-ek eta bere taldeak (Black *et al.* 1991) halako desanbiguazio lokal bat ere proposatzen zuten. Euren hurbilpenean hiru faktore hartzen ziren kontuan: lemaren luzera —zenbat eta laburrago hobe—, atzizkien maiztasuna eta aplikatutako erregelak. Hori ez dabil ondo gure corpusen gainean egindako probetan. Adibidez, *-iko* trigramaz bukatutako adjektibo mailegatuetan —eta hauek ez dira gutxi— *-ko* lemaren bukaera genitiboaren deklinabidetzat hartzen da aurreko irizpideei jaramonik egiten bazaie. Beraz, *bitaminikoaren* analisitzat *bitamini+ko+aren* gailenduko litzateke eta ez egiazkoa den *bitaminiko+aren*.

Guk definitutako prozeduran desanbiguaziorako erabilitako irizpideak hiru dira:

- Etiketa bakoitzeko analisi bat gutxienez mantentzea, desanbiguazioan etiketa morfosintaktikoaren arabera horien artean hautatuko delako.
- Lemaren bukaeraren arabera —azken trigrama da erabiltzen den unitatea— pisu bat esleitzea, lemari dagokion etiketa kontuan hartuz.
- Hitzaren luzera kontuan hartzea aipatutako hasierako proposamenaren ildotik.

Azken bi irizpideak konbinatzen dira lemaren bukaera dagokion kategoriarako ohikoa denean lema laburrena aukeratzeko, baina ez *bitaminiko* bezalako kasuetan. Trigramen pisuak kalkulatzeko garaian oso baliagarria izan da eskuz desanbiguatutako corpusa.

Prozedura honetan etiketatze-maila parametro gisa ematen da. Etiketa desberdin bakoitzeko gutxienez interpretazio morfologiko bat uzten duela ziurtatzen da, ondoren, testuinguru kontuan hartzen duen desanbiguazioak informazio gehiago baliatuta horien artean aukera dezan. Parametrorik eman ezean, kategoria eta azpikategoria (2. maila) erabiliko da.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%100	18,09	18,09	**%98,33**
ondoren (3. maila)	%100	12,56	12,56	**%96,35**
ondoren (2. maila)	%100	8,87	8,87	**%93,65**
ondoren (1. maila)	%100	8,52	8,52	**%81,88**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
ondoren (3. maila)	%100	14,35	14,35	**%94,04**
ondoren (2. maila)	%100	10,94	10,94	**%84,86**
ondoren (1. maila)	%100	10,11	10,11	**%78,44**

IV.6 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

Prozeduraren desanbiguazio-tasa %56koa, %54koa eta %33koa da 1., 2. eta 3. etiketatze-mailak erabiltzean, hurrenez hurren. Dagokien errore-tasa %1,72, %0,51 eta %0,36koa da, baina MORFEUSi gehitzen dion errorea %0,43, %0,12 eta %0,05koa da. Beraz, informazio gutxien (1. maila) erabiltzen denean, analisi gehiago baztertzen da (ikus IV.6(a) taula), baina beste metodoekin alderatuta, desanbiguazio-tasa bera lortzeko errore handiagoa sortzen da.

Aurkezten diren emaitzetarako erreferentzia-corpuseko EEBSko testuen informazioa erabili da soilik, *Euskaldunon Egunkaria*ko testuetan izen berezien agerpen gehiago izanik, izen arrunt edota adjektiboen lemen aurrean pisu handiegia hartzen dutela izen berezien amaierako trigramek, errore gehiago sortuz. Egiaztapenerako testuan gehienbat *Euskaldunon Egunkaria*ko testuak direnez, errore-tasa dezente handitzen da, bereziki izen berezi asko dagoelako. Hori dela eta, etorkizunerako trigramen neurriak modu egokian eguneratzeko metodo automatikoren bat erabiliko da, testu-mota aldatzeak desanbiguazio estatistikoaren emaitzetan eraginik izan ez dezan.

Gainera, oraindik ere hitz estandarren anbiguotasun-neurrietatik nahiko urrun daude hitz ezezagunena. Hori dela eta, metodo guztiak konbinatu behar direla ikusi da. Ondorengo atalean konbinaketa horiek eta dagozkien emaitzak aurkezten dira.

IV.2.2.5 Metodoen konbinaketa

Orain artean azaldu diren metodoek helburu bera modu desberdinean betetzen dute. Lehenengo hiruak hitz ezezagunen azpitalde zehatzei zuzenduak daude. Laugarrena, berriz, orokorragoa da, hitz ezezagun oro tratatuz. Hori dela eta, lehenik eta behin, tratamendu zehatzak aplikatuko dira eta horien ondoren geratzen diren aukerak estatistika erabilita murriztuko dira.

Konbinaketa posible gehienak probatu dira, baina ez dira guztien emaitzak hemen azalduko. Saiakuntza horien ondorioz, prozeduren hasierako anbiguotasuna murrizten den heinean emaitzak hobetzen direla ikusi da. Konbinaketa horietako batzuen ebaluazioa C eranskinean aurki daitezke —ikus C.1 atalean—.

Ondoren, metodo guztiak modurik egokienean aplikatuta lortutako emaitzak aurkezten dira. Lehenengo, desanbiguazio tipografikoa aplikatzen da, analisi asko baztertzen dituelako oso errore txikia eginez. Ondoren, eratorpenerako prozedura aplikatzen da. Jarraian, izen berezien identifikazioa gehitu da. Eratorpenerako prozeduraren atzetik nahiz aurretik jarrita ere, emaitza berberak ematen dituenez, bietako edozein lekutan jar daiteke. Azkenik, informazio morfologikoa eta estatistika erabiltzen duen heuristikoa aplikatzen da, 3. mailatik hasita emaitzak birfinduz lortuko dira MORFEUSen irteeran mantendu beharreko interpretazioak.[10] Parametroa zehaztu ezean, 2. maila erabiliko da prozedura honetan, IV.7 taulan beltzez agertzen den emaitza lortuz.

Identifikatzen diren izenak IV.2.2.3 atalean ematen diren 61 lema zuzenetatik 52 lortzen dira, desanbiguazio tipografiakoak lagunduta, akats gutxiago eginez. Horietatik 12 anbiguo geratzen dira eta sailkatzean horietako bi desanbiguatzen dira, lehen bezala *Kramnik* izenean huts eginez. Guztira 34 izen sailkatzen ditu eta aurreko atalean bezala *Kramnik*, *Muskildi*, *Naturgas* eta *Ortzaika* izenetan huts egiten du. Hala ere, oso zehatzak dira emaitzak, batez beste ia 2 interpretazio baztertu ondoren %0,7ko errorea egiten baitu. Analisi morfologikoaren irteeraren gainean %3tik gorako errorea egiten zuela kontuan hartuta, egokiena desanbiguazio tipografikoa egin ondoren aplikatzea dela ikusten da.

Egiaztapenerako corpusean, aldiz, 19 izen berezi identifikatzen ditu, 13 anbiguo izanik. Honen arrazoia agerpen kopuru txikia da. Anbiguo direnen artean *Fein-Feine*, *Flanagan-Flanagane*, *Yitzhak-Yitzha*, *Rabin-Rabi* eta *Volkswagen-Volkswage-Volkswag*. Erreferentzia-corpusean *Kramnik* lemarekin gertatu bezala, corpus honetan *Yitzha* lema aukeratu eta pertsona-izen berezi gisa sailkatzen du, baina azpikategoria asmatu arren, lema

[10] C.1 atalean informazio morfologikoa eta estatistikaren bidezko emaitzen bifinketaren arrazoiak aurkezten dira.

okerra aukeratzen du sailkatzerakoan. Guztira 9 lema sailkatzen ditu, eta aipatutako akatsaz gain, *Fein* lema ere gaizki sailkatzen du (*Sinn Fein* pertsona-izen gisa sailkatzen du), baina *Ortuzar* zuzen sailkatu eta desanbiguatzen du (*Ortuzar-Ortuzarrek* aukeren artean).

(a)	AR	I/A	I/T	R
aurretik	%100	18,09	18,09	**%98,33**
tipografikoa	%99,79	8,23	8,22	**%96,67**
eratorpena	%99,79	7,99	7,97	**%96,67**
izen bereziak	%83,75	6,97	6,00	**%95,94**
im+estatistika 3	%83,44	4,99	4,33	**%94,17**
im+estatistika 3+2	**%83,12**	**4,11**	**3,58**	**%92,92**
im+estatistika 3+2+1	%81,88	4,03	3,48	**%91,88**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
tipografikoa	%100	7,17	7,17	**%99,08**
eratorpena	%100	7,10	7,10	**%99,08**
izen bereziak	%94,95	5,01	4,81	**%92,20**
im+estatistika 3	%92,66	4,57	4,31	**%89,91**
im+estatistika 3+2	**%92,00**	**3,98**	**3,75**	**%88,07**
im+estatistika 3+2+1	%86,24	3,98	3,57	**%83,94**

IV.7 taula.- Anbiguotasun-neurriak prozedura aurretik eta ondoren.

IV.2.2.6 Emaitzen ebaluazioa

Tratamendu hau gehienbat hitz ezezagunetatik interpretazioak baztertzeko diseinatu da. Emaitzak ebaluatzeko erabilitako corpusean testu-hitzen %2,5-%3 besterik ez dira lexikorik gabe analizatu, baina nahiko adierazgarria da hain hitz gutxik interpretazioen %15-%20 biltzea.

Prozedura hauek aplikatu ondoren, lexikorik gabe analizatutako hitzek interpretazio guztien %3tik %4,5era besterik ez dituzte eta batezbestekoa 18-19 analisietatik 3,5-4,5era jaitsi da. Modu honetan MORFEUSen emaitzak erabiltzen dituzten prozesuen sarrerak anbiguotasun txikiagoa izateaz gain, erabilgarriagoa izango da, interpretazio kopurua lexikoan dauden hitzen batezbestekotik hurbilago izango delako. Guztiek batera %0,44ko errorea —%1 inguru egiaztapenerakoan— sortzen dute, baina prozesu osoari gehitzen dioten errorea %0,17koa besterik ez da —%0,47 ingurukoa egiaztapenerakoan.

Egiaztapenerako corpusaren emaitzetan ikusi ahal izan denez, emaitzak okerragoak dira, baina hau ez da harritzekoa, ez baitira testu hauek erabili hitz ez-estandarren kasuistika

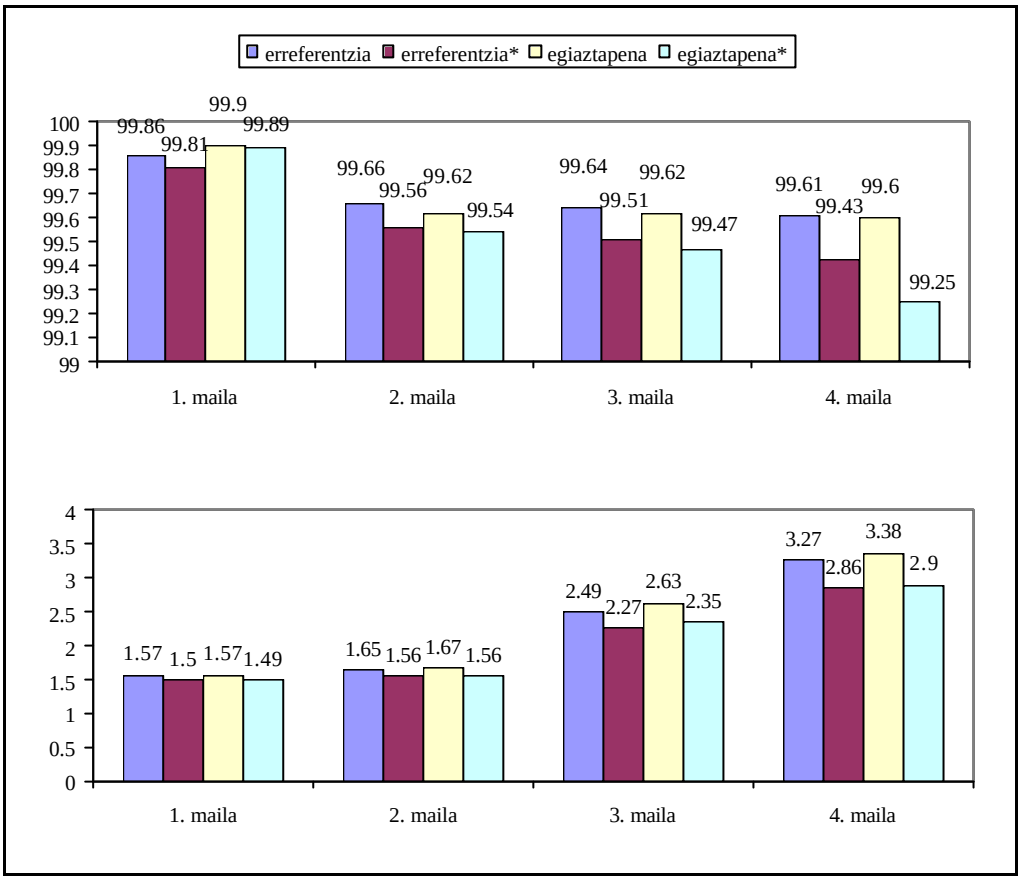
aztertu eta prozedurak diseinatzeko. Esate baterako, informazio morfologikoa eta lemaren bukaerako trigramen estatistikak erabiltzen dituen prozedurak akats asko egiten ditu. Horren arrazoiak, arestian aipatu bezala, batetik, estatistika lortzeko erabilitako testu-mota ez dela egunkarietako testu-motakoa eta, bestetik, izen berezien kopuru handia agertzen dela dira. Hain zuzen ere, horregatik aukeratu genuen corpus-mota hau, bai testuingururik gabeko desanbiguaziorako, baita testuinguruaren araberakorako ere, tratatzen zailagoa delako. Akats horiek ekiditeko modu asko daude, baina batzuk aipatzekotan, hauexek:

- Testuetako maiztasun handiko izen bereziak EDBLn lantzea, ez bakarrik euskal izen-abizenak, baita inguruko erdaretakoak. Hala ere, egunkarietan agertzen diren izen asko egoerak agindurik aldatuz doaz. Horren adibide *Barak* dugu, estatu-agintaritza utzi zuenetik egunkarietan aipatzen ez dutelako.
- Izen bereziak identifikatzeko ikasketa automatikorako teknikak erabiltzea, datu-basea aldatu gabe tratamendu egokia eman ahal izateko.
- Izen berezi konposatuak analisi morfologikoarekin batera identifikatu eta analisi egokia esleitzea. Aipatutako *Yitzhak Rabin* izen berezi konposatu gisa tratatu izan balitz, ez litzateke *Yitzhak* eta *Yitzha* lemen artean aukeratu behar izango, ezta *Rabin* eta *Rabi* artean ere. Gai honen inguruan egindako proposamen eta aplikazioak V. kapituluan ikusiko dira.

Dena dela, desanbiguazio morfosintaktikoari dagokionean, prozedura hauen aplikazioak zein eragin izan dezakeen aztertu behar da. Lehenengoz ikusi behar dena, etiketatze-maila bakoitzari dagozkion sarrerako datuak dira.

IV.1 irudian ikusten dira maila desberdinetako anbiguotasun-neurriak[11], tratamenduaren aurretik —taulan *erreferentzia* eta *egiaztapena*— eta ondoren —*erreferentzia** eta *egiaztapena**—. IV.1 irudiko datuak prozedura guztien konbinaketari dagozkio —2. mailaraino desanbiguatuz estatistikak erabilita. Ikus daitekeenez, hitz ez-estandarren tratamendua tokenen multzo txikiari aplikatzen bazaio ere, testu osoaren batezbesteko interpretazio kopurua urritzea lortzen du, prozesatutako hitzen batezbesteko analisi kopurua hitz estandarrenera hurbilduz. Gainera, zuzentasunari dagokionez, informazio morfologiko guztia kontutan hartuta analizatzaile morfologikoak eta hitz ez-estandarren tratamenduak elkarrekin egindako errore-tasa ez da %1era iristen eta 2. mailan %0,5 ingurukoa da.

11 Anbiguotasunari buruzko datuak C.2 atalean aurki daitezke.



IV.1 irudia.- Zuzentasuna eta analisi kopurua etiketa-mailaren arabera.

Batezbesteko analisi kopuruari dagokionean, IV.1 irudian ikus daiteke 4. mailan 3 analisitik jaistea lortzen dela errore txikia eginez. Egiaztapenerako corpusean, errorea handiagoa da, gehienbat izen berezien desanbiguazioan egindako akatsak direla medio.

Oraindik ere anbiguotasun-tasa handia du MORFEUSen emaitzak, baina motibazio linguistikoei jarraituz eguneratzen da datu-basea, eta kasu askotan deskribapen linguistiko zehatzagoa egitearren anbiguotasun-neurriak nabarmenki goraka egiten du, eta desanbiguazioaren emaitzek okerrera. Hala ere, gure helburu nagusia ez da izan hainbeste zenbaki onak lortzea, tresna erabilgarriak egitea baizik, eta horregatik diseinatu dira atal honetan aurkeztutako prozedurak.

VI. kapituluan tratamendu honek desanbiguazio morfosintaktikoan nola laguntzen duen sakonean azaltzen da, eta horren ondorioz ere, lematizazio/etiketatzean oinarritzen diren gainerako prozesuetan ere, hala nola, azaleko sintaxia, terminologiaren erauzketa automatikoa, etab. Dena dela, aurrerapen moduan, prozeduren lehen ebaluaziorako egindako saiakuntza batzuen emaitzak aurkezten dira jarraian.

Esate baterako, testuingurua kontuan hartzen duen murriztapen-gramatikan, hitz ezezagunen interpretazio kopurua jaisteko erregelak gehitu ziren, heuristikoetako batzuen erabilitako informazioa oinarri hartuta, kapitulu honetan azaldutako prozeduren beharra

ebaluatzeko. Gramatika hau 4 erregela-multzoetan banatu zen. IV.8 taulan analizatzaile morfologikoaren irteerari gramatika aplikatu ondorengo emaitzak azaltzen dira. Ikus daitekeenez, hitz ezezagunen batezbesteko interpretazio kopurua 4ra jaisteko %44ko errorea gehitzen du.

Emaitza hauetatik hitz ezezagunen desanbiguazioan testuinguruaren informazioak lana errazten ez duela ondoriozta daiteke, izan ere, lanaren konplexutasuna hitzaren interpretazioetan datza, hitz estandarrek jasotzen ez duten interpretazio-multzoak direlako. Interpretazio horietako batzuk baztertzeko, kontzeptualki aukera horiek guztiak dituzten hitzen portaera testuinguru bakoitzean aurreikusi eta dagokien erregela-multzoa eskuz diseinatzea oso zaila da. Horregatik lortzen dira emaitza hobeak informazioari baino formari erreparatuta.

	AR	**I/A**	**I/T**	**R**	**P**	**F**
IV.1 taulatik	%100	18,09	18,09	**%98,33**	%5,44	10,30
MG (1-2 multzoak)	%75,94	6,84	5,43	**%68,83**	%12,67	21,39
MG (1-3 multzoak)	%75,00	6,69	5,27	**%67,47**	%12,80	21,52
MG (1-4 multzoak)	%68,31	5,43	4,03	**%54,34**	%13,50	21,64

IV.8 taula.- Ezezagunen anbiguotasun-neurriak morfologikoa eta MG ondoren.

Hitz ezezagunen tratamendua egin ondoren, berriz, gramatika hori bera aplikatuta IV.9 taulako emaitzak lortzen dira. Gramatikaren 4 multzoak erabilita, errore gehien sortzen duena aukera izanik ere, IV.8ko emaitzarik onena %9an hobetzen du, nahiz eta hasierako errorea %5 altuago. Zehaztasunean ere are gehiago hobetzen da eta, horregatik, hasieran *f-score* neurrian 30eko diferentzia bazegoen ere, amaieran 33-36ko aldea dago. Hala ere, gramatika honek egindako errore-tasa nahiko altua da eta hitz ez-estandarren tratamenduaren emaitzak ikusirik hitz ezezagunetarako diseinatutako erregela horiek baztertzea erabaki da. Beraz, VI. kapituluan murriztapen-gramatikari buruzko atalean ematen diren emaitzak eta hemen aurkeztutakoak ez dira bat etorriko.

	AR	**I/A**	**I/T**	**R**	**P**	**F**
IV.7 taulatik	%83,12	4,11	3,58	**%92,92**	%25,92	40,54
MG (1-2 multzoak)	%48,01	3,01	1,96	**%81,07**	%41,27	54,64
MG (1-3 multzoak)	%46,55	2,98	1,91	**%80,44**	%41,82	55,03
MG (1-4 multzoak)	%41,63	2,69	1,70	**%77,93**	%45,73	57,64

IV.9 taula.- Ezezagunen anbiguotasun-neurriak tratamendua eta MG ondoren.

Murriztapen-gramatikaren lehenengo bi multzoen emaitzaren gainean desanbiguazio estokastikoa[12] aplikatuz gero IV.10 taulako datuak lortzen dira. Aipatzekoa da murriztapen-gramatika aplikatu ondorengo zuzentasun-tasan %12ko diferentzia izanik, amaieran %20koa dagoela. Kasu honetan ez da *f-score* neurrian hobekuntza handiagoa lortzen desanbiguazioa egin ostean 2. mailako etiketa bakarra uzten delako eta IV.8 eta IV.9 tauletako emaitzen artean batez beste 0,8 interpretazioko aldea dago 2. mailan. Beraz, desanbiguazio estokastikoak interpretazio gehiago baztertuko ditu lehenengo kasuan bigarrenean baino.

	AR	**I/A**	**I/T**	**R**	**P**	**F**
IV.8 taulatik	%75,94	6,84	5,43	**%68,83**	%12,67	21,39
ondoren	%55,54	3,15	2,20	**%47,70**	%21,72	29,85
IV.9 taulatik	%48,01	3,01	1,96	**%81,07**	%41,27	54,64
ondoren	%25,52	2,14	1,29	**%68,31**	%71,77	59,66

IV.10 taula.- Ezezagunen anbiguotasun-neurriak desanbiguazio estokastikoaren ondoren.

IV.8 eta IV.9 tauletako emaitzek hitz ez-estandarren tratamenduak ondorengo urratsetan, bereziki testuinguruaren araberako desanbiguazioan laguntzen duela eta, ezinbestekoa ez den arren, zuzentasunean ere emaitzak hobetzen dituela erakusten dute.

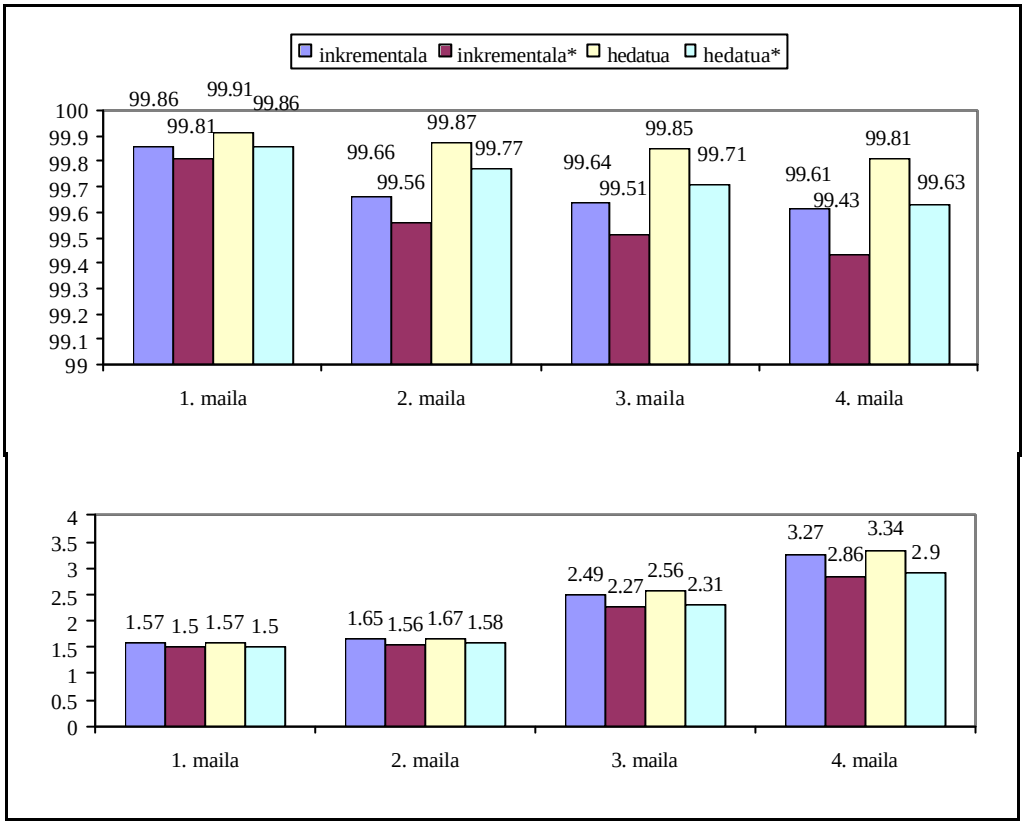
Emaitzak hauek guztiek kapitulu honetan deskribatutako tratamenduaren beharra eta dakarren hobekuntza frogatzen dute. Erakutsi den bezala, gainsorkuntza morfologikoa dela medio, hitz ezezagunak desanbiguatzen bereziki zailak dira. Baina analisiaren fasean interpretazio morfologiko posible guztiak ematea behar-beharrezkoa denez eta analisiaren atzetik datozen prozesuek hainbesteko anbiguotasuna nekez ebatz dezaketenez, testuinguruari buruzko informaziorik gabe, ahalik eta interpretazio gehien baztertzeko, prozedura hauek erabiliko dira.

IV.3 Zuzentasuna eta zehaztasunaren hobekuntza

III. kapituluan deskribatu den bezala, analizatzaile morfologikoaren zuzentasun-neurriak hobetzeko aldaketak burutu dira. Baina horren ondorioz anbiguotasuna areagotu egiten da. Atal honetan hitz bakunen tratamenduan egindako hobekuntza guztiak batera aplikatuta lortutako emaitzen azterketa egingo da.

[12] Desanbiguazio estokastikoan 2. mailako etiketa sistema erabili eta etiketa horiei dagozkien analisi morfologiko oso guztiak kontuan hartuta kalkulatu dira emaitzak.

IV.2 irudiko grafikoetan erreferentzia-corpusari dagozkion analizatzaile inkrementala eta hedatuaren emaitzak aurkezten dira. Ikusten denez, anbiguotasuna igo egiten da IV.1 taulako datuekin alderatuta, zehaztasunean galduz. Galera hori hitz estandarren eragina da gehienbat, baina oraingoz ez da tratamendu berezirik diseinatu gehitutako anbiguotasuna jaisteko.



IV.2 irudia.- Zuzentasuna eta analisi kopurua etiketa-mailaren arabera.

Hala ere, emaitza hauek aztertuz gero ikus daiteke analizatzaile hedatuaren emaitzei hitz ez-estandarren tratamendua aplikatu ondoren, analizatzaile inkrementalaren hasierako zuzentasun-maila lortzen dela, baina analisi horri hitz ez-estandarren tratamendua aplikatu ondorengo anbiguotasun-mailarekin. Horregatik, hemendik aurrerako kapituluetan analizatzaile hedatua eta hitz ez-estandarren analisiaren emaitzen gainean aplikatuko dira gainerako prozesuak.

IV.4 Etorkizunerako hobekuntzak

Kapituluan zehar hainbat hobekuntza proposatu dira etorkizunean garatzeko. Atal honetan guztiak laburbiltzen dira.

Izen berezien desanbiguazioari dagokionean, hainbat arlo geratu dira jorratu gabe. Aipatu denez, izen berezien hautagaiak baztertzeko behe-muga emaitzetan oinarrituz erabaki da.

Hala ere, corpus txikiaren gainean lan egin da eta, hortaz, corpus handiagoak tratatzerakoan, tamainaren arabera behe-muga hori ezartzea komeni da. Hau da, corpusaren tamaina handitzean, gai bereko testu gehiago egon daiteke eta, hortaz, izen berezien agerpen kopurua ere handiagoa izango da. Horregatik, oso maiz agertzen direnak soilik tratatzeko, agerpen kopuru minimo hori handitu beharko litzateke.

Hitz estandarren kasuan, analizatzaile hedatuak izen berezien analisiak gehitzen dizkie, baina ez da inongo prozedurarik aplikatu horien anbiguotasuna kontrolatzeko. Aukera desberdinak aztertzen ari gara, besteak beste, kapitulu honetan aurkeztutako heuristiko bera —edo egokitua— aplikatzea testuinguruaren araberako desanbiguazioaren aurretik.

Dena dela, bai hitz estandar bai ez-estandarren kasurako Mikheev-ek (1999, 2000a) proposatutako *Document Centered Approach* teknika erabil daiteke hitz berezien identifikazio eta sailkapenerako.

Gainera, identifikatu eta sailkatzen diren izen berezi horiek prozesuaren amaieran erabil daitezke datu-basea aberasteko modu semiautomatikoan, beti ere maiztasun handikoak badira, edota hiztegi berezitua sortu ondorengo aplikazioetan lortutako informazio hori baliatu ahal izateko.

Bestalde, informazio morfologikoa eta estatistika erabiltzen duen heuristikoa oso corpus txikitik lortutako neurrietan oinarritzen da. Komenigarria litzateke trigramen estatistikak eguneratzeko prozeduraren bat prestatzea heuristikoaren emaitzak hobetzeko. Horrela, EUSLEMek testu berriak tratatzean, horietatik informazioa erauzi eta datuak aberats daitezke.

Azkenik, emaitzen azterketa egitean, *Euskaldunon Egunkaria*ko testuetan corpus orekatuan baino errore gehiago egiten direla ikusi da eta, gainera hasierako anbiguotasuna handiagoa da. Testu hauek sakonean aztertu beharko lirateke heuristikoak hobetzeko, bereziki izen bereziei dagokienean.

V Hitz anitzeko unitateen tratamendua

Testu-sailak lematizatu eta etiketatu nahi direnean, testua osatzen duten hitzen interpretazioak behar-beharrezkoak dira. Interpretazio horiek aztertzean, hitzari dagozkion lema eta etiketa hautatuko dira. Baina prozesua ezin daiteke hitz-mailan geratu. Hitz-konbinazio askotan osagaiak elkarren menpekoak dira, eta, zenbait kasutan, osagaion azterketa eginez ezin da unitate konplexu hauen interpretazioa lortu. Hitz-konbinazio hauei hitz anitzeko unitate esaten zaie.

Hitz anitzeko unitateen tratamendua diseinatzeko garaian, bibliografiaren azterketa egitean argi ikusi genuen tresna oso garrantzitsua izan arren, aurkezten ziren aplikazio gehienetan unitate hauek ez zirela kontuan hartzen, edota maiztasun handieneko unitateen zerrendak bakarrik erabiltzen zirela. Ikuspegi linguistikoari erreparatuz gero, unitate hauek sailkatzean desadostasun nabarmenak zeuden eta ikerkuntza-gai gehienak hitz anitzeko unitateen multzo murritzetara mugatzen ziren.

Mota honetako tresna baten garrantzia neurtzeko, (Sag *et al*. 2002) lanean agertutako baieztapena aztertzea nahikoa da:

> *"As Jackendoff (1997:156) notes, the magnitude of this problem is far greater than has traditionally been realized within linguistics. He estimates that the number of multiword expressions (MWEs) in a speaker's lexicon is of the same order of magnitude as the number of single words. In fact, it seems likely that this is an underestimate, even if we only include lexicalized phrases. In WordNet 1.7 (Fellbaum 1999), for example, 41% of the entries are multiword."*

Datu hauek neurri egokia ematen badute, logika aplikatuz eta enpresen aldetik dagoen interes handia kontuan izanik, gai honetan ikertuko duenik ez litzateke falta beharko, baina, antza denez, ez da horrela (Sag *et al*. 2002):

> *"...The second key problem facing the deep processing program —the problem of multiword expressions— is underappreciated in the field at large. There is insufficient ongoing work*

> *investigating the nature of this problem or seeking computationally tractable techniques that will contribute to its solution"*

Hala ere, gaur egunean gai honen inguruko interesa pizten ari dela dirudi, hainbat proiektutan arazo hau ebazteko ikertzen ari direlarik, bai alde linguistikotik ikusita, baita alde informatikotik ikusita ere.

Ikuspegi linguistikotik, esan bezala, ikerkuntza-gai gehienak multzo zehatzetara mugaturik izan ohi dira, adjektibo-izen, aditz-preposizio[1], adierazpen idiomatikoen[2] multzoak, etab. Baina hitz anitzeko unitateak automatikoki identifikatu eta prozesatu nahi badira, ikuspegi zabalagoa behar-beharrezkoa da. Horrelako sailkapenak ere topa daitezke bibliografian (Karlsson *et al.* (*eds.*) 1995:134; Aduriz *et al.* 1996-c; Sag *et al.* 2002). Nahiko sailkapen antzekoak dira, kontuan hartzen diren hitz anitzeko unitate motei dagokionean, baina gehienetan norberak identifikatutako problematikari erreparatzen dio unitate-multzo bakoitzean sailkapena birfindu ala ez erabakitzeko. Azken finean, aipatutako sailkapen horiek prozesamendu automatikoari begira egin direnez, ikuspegi teoriko-praktikoari jarraituz garatu dira.

Azken urteotan martxan dauden proiektuetako bat aipatzekotan *Multiword Expression Project* da informazio interesgarriena bildu eta erabilgarri jartzen ari dena. Oraindik ere eztabaida fasean badaude ere, gaiak merezi duen garrantzia ematen diotela dirudi. Lan-taldeari buruzko informazio gehiago *http://lingo.stanford.edu/mwe/* helbidean lor daiteke, baita gaiari buruzko bibliografia nahiko zabalaren erreferentziak.

Alde informatikoari dagokionean, garrantzitsua da hitz anitzeko unitateak modu eraginkorrean tratatzea. Bi gauza hartu behar dira kontuan horretarako: tresna azkarrak izatea eta erabili beharreko datu kopurua neurrikoa izatea. Lehenengoari dagokionean, aurrerago sakonean aurkezten den IDAREX (*IDioms As Regular EXpressions*) Xerox-en garatutako hurbilpena da hitz anitzeko unitateen ezaugarri gehienak kontuan hartu, eta modu eraginkorrean prozesatzen duen adibide aipagarriena (Segond eta Tapanainen 1995; Segond eta Breidt 1995). Bigarrenari dagokionean, sarrerako datuak —zerrendak edo datu-baseak izanik— ezin dira etengabe handitu eta hori ekiditearren neurri estatistikoak erabili ohi dira testuetan ager daitezkeen hitz anitzeko unitateak identifikatzearren[3]. Sistema askotan erabili izan den *Xtract* tresna da horren adibide (Smadja 1993).

---

1 Ingelesezko *phrasal verb* deritzatenak barne.

2 Ingelesezko *idioms* gehienbat.

3 Erabili ohi diren neurrien azterketa (Manning eta Schütze 1999:141) liburuan aurki daiteke.

Euskarazko hitz anitzeko unitateak tratatzeko IDAREX da aztertutako hurbilpenen artean egokiena, baina erabiltzen den teknologia ez da publikoa. Gaur egun, IDAREXen oinarrian dagoen Xeroxeko *xfst*[4] tresna erabiltzeko lizentzia badugu eta etorkizunean bide hori aztertzeko asmoa dugu, baina lan honi ekin genionean, ez genuen modu horretako tresnarik eskura. Beraz, tresna publikoen eskaintza falta ikusirik, euskararen hitz anitzeko unitateen azterketari ekin genion *pattern matching* ohiko tekniketan oinarrituz, lematizatzaile/etiketatzailearen osagai funtsezkoa zela pentsatzen genuelako (eta pentsatzen dugulako). Aldi berean, tratamendu hori prozesuaren zein unetan egin behar zen ere erabaki behar izan genuen. Lan horren emaitza kapituluan zehar aurkezten da.

Lematizazioa helburu izanik, gure ustetan lehenbailehen identifikatu behar dira, hitz anitzeko unitate baten osagaiek elkarrekin osatzen baitute euren lema. Gainera, testua etiketatu nahi da, eta, askotan, unitate konplexua osatzean hitzei dagokien informazioa ere aldatu egiten da. Esate baterako, testuan *hala eta guztiz ere* aurkitzean, ezin dira hitzak independenteki interpretatu, elkarren ondoan daudenean bere funtzioa aldatzen baita.

Bestalde, hitz anitzeko unitatearen osagai batzuk banaka azterturik anbiguoak izan arren[5], unitate konplexuaren osagai gisa interpretazio bakarra izan dezakete. Hori dela eta, hitz anitzeko unitateak identifikatuz eta tratatuz zehaztasunean irabaz daitekeela kontuan hartuta, hitz anitzeko unitateen tratamenduak desanbiguazio-prozesuan lagun dezake.

Argi gera bedi tratamendu honen helburua ez dela sintaxia lantzea; ez dira unitate sintaktikoak identifikatu nahi. Lan hori formalismo sintaktikoen bitartez egingo da, baina morfologia eta sintaxiaren arteko muga oso lausoa izanik, hitz anitzeko unitateak non identifikatu erabakitzea ez da lan erraza; unitate konplexuen tratamendua oso gai irekia da eta gaiari buruzko iritzi kontrajarriak daude: batzuek sintaxi-mailan landu behar direla diote eta beste batzuek, berriz, morfologiaren aurretik edota batera. Euskararen kasuan, aipatu diren arrazoi guztiak kontuan izanik —lemaren beharra, informazio morfologikoaren aldaketa eta zehaztasunaren hobekuntza potentziala—, hitz anitzeko unitateen tratamendua morfologiaren ondoren baina desanbiguazio eta sintaxiaren aurretik egitea erabaki da.

Hitz anitzeko unitateen tratamenduaren lehen hurbilpena egin zenean (Aduriz *et al*. 1996-c; Ezeiza 1997), prozesatu beharreko unitateen azterketarako abiapuntu gisa, UZEIk EEBS proiektuan jasotako informazioa erabili zen (Urkia eta Sagarna 1991). Azterketa horren ondorioz, lokuzio eta kolokazio murriztuez gain, euskal hiztegietan sarrera merezi duten

4 *xfst* Xeroxeko programa bat da espresio erregularrak transduktore bihurtzeko eta horrela prozesatu ahal izateko (Beesley eta Karttunen, 2002).

5 B eranskineko 5. eta 6. adibideetan, *hala eta guztiz ere* hitzen analisi morfologikoa ematean, *eta* osagaia anbiguo izan arren, tratamenduaren ostean anbiguotasun hori kendu dela ikus daiteke.

bestelako unitateak ere datu-basean lantzea interesgarritzat jo zen. Beraz, EEBSko 3.000 unitate konplexuetatik maiztasun handieneko 500 inguru aukeratu ziren eta hizkuntzalariak corpusez eta euren esperientziaz baliatu ziren bakoitzari zegokion informazioa osatzeko.

Informazio horretan oinarrituta, hitz anitzeko unitate lexikalak (HAUL) identifikatu eta dagozkien interpretazioez hornituko zituen tresna (HABIL) inplementatu zen. Saiakuntza horien emaitzen ebaluaziotik honako hobekuntza hauek ondorioztatu ziren, gehienbat HAUL anbiguoen tratamenduari dagokionean (Ezeiza 1997):

- *"Zenbait kasutan osagaien artean joan daitekeen hitz kopurua muga daiteke. Honela, muga hori baino urrutiago dauden hautagaiak bazter daitezke"*. Dena dela, HAUL baten osagaien erdian zenbat hitz ager daitezkeen neurtzea oso zaila da, ebaluaziorako erabili ziren testuetan 10 hitzetik gora ere topatu direlako. Kasurik gehienetan 2-3 hitz izan ohi dira gehienez ere, baina aurrerago agertzen den (2) adibidean, esate baterako 4 hitz agertzen dira eta HAUL hori identifikatzea interesatzen zaigu.
- *"Bi osagai baino gehiago duten HAUL batzuetan ordena posibleak definituz zilegi direnak bakarrik onartuko dira"*. Adibidez, *begi bistan egon* aditzaren kasuan, zilegi da ordena horretan eta *egon begi bistan* ordenan agertzea, baina inolaz ere *bistan begi egon* edota *egon bistan begi* ordenak. Hortaz, ordena aldaketa guztiak ez dira kontuan hartuko.
- *"Nahiz eta orokorrean anbiguoak direla esan, ordena zehatz batzuetan ziurtzat jo daitezkeen HAULak topatu dira. Hortaz, ordena zilegi bakoitzak bere ziurtasun-balioa izango du"*. Adibidez, *maite izan* aditzaren osagaiek ordena alda dezakete eta erdian hitzak ager daitezke. Ordena zilegiak banan-banan emanez gero, kasu batzuk ez-anbiguo direla esan daiteke — (1) esaldian agertzen den ordena ez-anbiguoa dela esan daiteke— baina beste batzuetan, erdian edozein hitz kopuru agertzean zehazki, anbiguo izango dira, (2) esaldia kasu.

(1) "Bidaia gogorra da, baina senide eta lagunak pozik dira ***maite dituzten*** *lagunak ikusteko aukera dutelako haien aurrean."*

(2) "... Sophiek onartzen zuen serbiarrek ez ***dituztela*** *Mendebal Europako kazetariak gehiegi* ***maite****."*

Gainera, bestelako hitz anitzeko unitate batzuk ere identifikatzea komeni dela ikusi da, hala nola, datak, hitz anitzeko zenbakien adierazpenak, izen berezi konposatuak, etab., lengoaia naturalaren prozesamenduko hainbat arlotan, lematizazioan, sintaxian, informazioaren erauzketan eta berreskurapenean esaterako, horrelako unitateak identifikaturik egoteak lana erraztuko duelako.

Beraz, aipatutako hobekuntzei aurre egiteko egindako lanaren aurkezpena honela dago antolatua: V.1 atalean hitz anitzeko unitateen ezaugarriak ikuspuntu linguistikotik aztertu eta sailkatzen dira; V.2 atalean, berriz, tratamendu automatikoa diseinatzerakoan sortzen diren

arazoak aztertu eta soluziobideak planteatzen dira; V.3 atalean euskararen kasurako inplementatutako tresnaren deskribapena eta ebaluazio egiten da; eta, azkenik, V.4 atalean, etorkizunerako hobekuntzak plazaratzen dira.

V.1 Ikuspuntu linguistikoa

Arestian esan bezala, hitz anitzeko unitate edo adierazpenak ikuspuntu linguistikotik sailkatzerakoan ez dago adostasun handirik eta sailkapen bat baino gehiago aurki daiteke bibliografia aztertzean. Beraz, tratatu nahi diren hitz anitzeko unitateak gure irizpideen arabera multzokatu ditugu, beste sailkapen zehatzago batzuk gutxietsi edo baztertu gabe.

Gure helburua euskarazko hitz anitzeko unitateak aztertzea izanik, UZEIk EEBS proiektuan izandako esperientzia baliatu dugu tratatu nahi diren hitz anitzeko unitate lexikalak mugatzeko azterketa egiterakoan, hark, beharrizan lexikografikoetan oinarrituta, unitate konplexuak zehazterakoan eskuzabal jokatu baitu.

Ondoren azalduko dira lan honetan Hitz Anitzeko Unitate Lexikalak (HAUL) mugatzeko jarraitutako irizpideak. Hasteko, hitz elkartuetan sakontzen da HAULen artean multzo berezia osatzen baitute hauek. Ondoren, kolokazio lexikal eta lokuzioei atal bat eskaintzen zaie, irizpide semantikoetan oinarritutako talde honen sailkapen bat eginez. Hirugarren atalean aditz konposatuen tratamenduaren inguruko hausnarketa egiten da. Azterketa hau amaitzeko, tratatu nahi diren unitateen ezaugarri formalak deskribatu eta HAULak biltzeko diseinatu den datu-basea aurkeztuko da. Azkenik, interesgarriak diren bestelako unitate batzuk ere aurkezten dira

V.1.1 Hitz-elkarketa

Arestian esan denez, hitz elkartuek oso talde berezia osatzen dute. Izan ere, modu desberdinetan idatz daitezke eta idazkeraren arabera sailkatzen badira, batzuk hitz anitzekoak izango dira eta beste batzuk, berriz, hitz bakarrekoak. Atal honetan, idazkerari dagokion sailkapena aztertuz, talde bakoitzeko hitz-elkarketak izango duen trataera azalduko da —adibideen analisiak B eranskineko B.7 atalean ikus daitezke—.

Euskaraz hitz elkartuak lau modutara idatz daitezke nagusiki (Euskaltzaindia 1992):

- loturik (*idazmakina, plazagizon*).
- marraz bereizirik (*datu-base, begi-nini*).

- bereiz eta elkarketaren lehen osagaiak forma aldatua duelarik. *Euskal etxe*, *itsas armada*, *Filologi Saila* adibideetan, *euskal*, *itsas* eta *filologi* hitzak eratorpen eta hitz-elkarketan *euskara*, *itsaso* eta *filologia* hitzen aldaerak dira, hurrenez hurren.
- bereiz baina lehen osagaiak aldaketarik jasan gabe (*hauteskunde emaitzak*).

Lematizazioa eta etiketatzea helburu, loturik idazten diren hitz elkartuak hitz bakunen pareko dira eta halakotzat dauzkagu datu-basean. Beraz, gainerako hitz bakunen modura tratatuko ditu analizatzaile morfologikoak (ikus B.7.1).

Marraz bereizirik idazten diren hitz elkartuen artean, berriz, badira batzuk lexikalizatutzat jotzen direnak —hiztegietan daudenak— eta horiek hitz bakun gisa daude sartuak datu-basean (*begi-nini, botoi-zulo*). Hortaz, loturik idazten direnak bezala tratatuko dira (ikus B.7.2).

Beste batzuk, ostera, sorrera librekoak izanik (*mahai-hanka, hauteskunde-emaitzak*), orokorrean ez daude hiztegietan eta, beraz, ez daude datu-basean. Baina analizatzaile morfologikoa gai da horietako batzuk ezagutzeko. Bi mailatako morfologiak (Koskenniemi 1983) hala ahalbidetzen duenez, marratxoa elementu lexikal gisa tratatzen da datu-basean. Marratxo ezberdinek[6], beraz, datu-baseko gainontzeko elementuek bezalaxe, morfema- eta lexikoi-multzo bat har dezakete atzetik, hots, euren jarraitze-klasea dute, eta horrela zenbait hitz elkartu —*izen arrunt*+-+*izen arrunt* elkarketak zehazki— ezagut daitezke.

Dena dela, elkarketa hauen tratamenduak bi arazo planteatzen ditu ebazteko. Batetik, hiztegietan sarrera dutenen kasuan gainsorkuntza gertatzen da (ikus B.7.2). Hau da, datu-basean sarrera gisa sartuta daudenez, hitz bakun baten moduan ematen du emaitza —marratxoa kontuan izan gabe— eta, gainera, bi hitzak landuta daudenez, lema hiru osagaiez —lehenengo izena, marratxoa eta bigarren izena— osatuta egonik ere analizatuko du. Horrek, zehaztasunaren galera dakar. Bestetik, izen arrunten arteko elkarketak tratatzen diren arren, bestelako elkarketak ez dira aurreikusi, hala nola, *izen berezi*+-+*izen berezi* (*Bilbo-Behobia*), *adjektibo*+-+*adjektibo* (*urdin-urdin*), etab. Horrelakoak ere landu beharko lirateke, ahal den neurrian gainsorkuntza ekidinez.

Lehen osagaia aldatua duten hitz elkartuen artean ere badira lexikalizatuak eta sorrera librekoak. Lexikalizatuta daudenak bakarrik sartu dira datu-basean (*Euskal Herria*). Baina gainerako guztiak ere detektatu ahal izango dira, fonologikoki eraldatu daitezkeen hitzen multzoa aski mugatua baita:

---

6 Marratxo mota bat baino gehiago dago datu-basean, noski, elkarketa-mota desberdinak ezezik marratxoaren gainontzeko erabilerak adierazteko ere (*Iñakik-eta, William-engandik*,...).

- bukaerako *a* galtzen dutenen artean, batetik, *-ia* bukaeradun izen guztiak dira —*anaia* izan ezik— (*filologia*, *energia*, *psikologia*,...), erregela morfofonologiko sinple batez ezagutzen direnak, eta, bestetik, salbuespena osatzen duten izen gutxi batzuk (*natura*, *literatura*, *kultura*, *burdina*, *eliza* eta *hizkuntza*).
- gainontzeko kasuek nahiko multzo murritza osatzen dute (*itsas*, *erret*, *euskal*,...).

Elkarketan eraldatzen diren hitzek talde oso mugatua osatzen dute, baina hauen bidez osatzen diren hitz elkartuen taldea, berriz, zabala da oso. Horregatik, HAULen datu-basean ez lantzea erabaki da. Lehen osagaia bilaketa-prozesu erraz baten bidez aurki daiteke eta, datu-basean aukera guztiak landu gabe, hitz elkartu hauek HAUL gisa marka daitezke.

Azkenik, bereiz eta marratxorik gabe idazten diren hitz elkartuak antzemateko, datu-basean sartzea beste modurik ez dago. Hauetako batzuk marraz ere idatz daitezke eta, ondorioz, bi tratamendu desberdin izan ditzakete[7] (*hauteskunde-emaitzak* edo *hauteskunde emaitzak* —ikus B.7.2 eta B.7.4—). Sorrera libreko elkarketak detektatzerik ez dago momentuz. Mota honetakoak tratatzeari ez zaio garrantzi handiegirik eman etiketatze-prozesuan desbiderapen handirik sortzen ez dutelako, hitz bakunen modura jasotako analisiek elkarketaren interpretazioa ere islatzen dutelako. Dena den, maiztasun handiko agerkidetzak corpusetan automatikoki bilatuko dituen tresna gara daiteke, lortutako agerkidetza berri horiekin datu-basea aberasteko asmoz.

Atal honetan aurkezten diren hitz anitzeko elkarketa batzuk hurrengo atalean aurkezten den sailkapenean ere agertzen dira. Izan ere, irizpide semantikoei jarraiki, atal honetako adibideak multzo desberdinetan kokatzen dira.

V.1.2 Kolokazio lexikalak eta lokuzioak

Hitz anitzeko elkarketekin batera badira beste hitz-konbinazio batzuk unitate bakar gisa lematizatu beharrekoak. Orokorrean, unitate lexikal esatean, kolokazio[8] lexikal eta lokuzioei buruz hitz egiten da. Ondoren, hauek aztertzen dira HAUL bezala landu behar diren erabakitzearren.

---

7 Datu-basean bi lekutan agertu beharko lirateke, batetik, hitz bakun moduan eta, bestetik, HAUL moduan. Dena dela, bikoizketa ekidin daiteke informazioa esportatzerakoan marradun elkarketa horien kasuan, HAUL informazioa automatikoki sortzen bada.

8 *Kolokazio* terminoa ingeleseko *collocation* zentzuaz erabiliko dugu. Kolokazio batean terminoetako batek oinarri (*base*) funtzioa du eta besteak, berriz, kokakide (*collocate*) funtzioa izango du. *Agerkidetza* diogunean, ingeleseko *co-ocurrence* adierazi nahi dugu. Kasu honetan, osagaiak corpusetan elkarrekin askotan ager daitezke baina osagaiek ez dituzte funtzioak kolokazioetan bezala banatzen.

Kolokazio lexikalen eta lokuzioen artean bereizteko irizpide semantikoa erabili ohi da nagusiki (Heid 1994). Alegia, lokuzio baten interpretazioa nekez egin daiteke haren osagaien esanahietatik abiatuta (*adarra jo* ≠ *adarra* + *jo*). Kolokazioetan, berriz, osagaiak —edo gutxienez horietako bat— euren hitzez hitzeko ohiko zentzuan erabiltzen dira (*zarata atera*).

Halere, lokuzio eta kolokazio lexikalen arteko muga zedarritzea zaila da oso, lokuzio guztiz opako eta kolokazio irekien artean hitz-konbinazio mota ugari baitaude mailaketa edo *continuum* batean banatuak (Cowie *et al.* 1985):

- **lokuzio guztiz opakoak**: Hauexek dira lokuzioak *strictu sensu*. Historikoki, lokuzio opakoak prozesu baten azken muturra dira; hitz-konbinazioak finkatu egiten dira lehenbizi erabiliaren erabiliaz, gero izaera figuratiboa hartu, eta azkenik fosildu eta guztiz opako bihurtzen dira. Hauen artean badira lokuzio lexikal hutsak (*ahuntzaren gauerdiko eztula*), batetik, eta guk gramatikal deritzegunak (*harik eta, hala eta guztiz ere*), bestetik.
- **lokuzio figuratiboak**: Hitz-konbinazio hauek lokuziotzat hartu ohi dira nahiko osaera finkoa dutelako —nekez onartzen dute aldagairik—, baina hiztunarentzat jatorrizko erreferentzia erreala ez da lokuzio opakoetan bezain urruna (*hutsaren hurrengoa*).
- **kolokazio murriztuak**: Sasi-lokuzioak ere deitu izan zaien hauetan osagai bat bere ohiko zentzuaz erabiltzen da. Gainerakoek testuinguru zehatz honetatik kanpo aurki ez daitekeen zentzu figuratiboa dute (*eskerrak eman*).
- **kolokazio irekiak**: Hitz-konbinazio hauetako elementu bakoitza bere hitzez hitzeko zentzuan erabiltzen da (*hego haizea*).

Gorago esan bezala, lan honetan, UZEIren irizpideari jarraiki, eskuzabal jokatu da HAULa zer kontsideratu erabakitzerakoan. Lokuzio guztiak hartu dira kontuan —opakoak zein figuratiboak— baita kolokazio murriztuak ere. Kolokazio irekiak, berriz, ez dira landu, kontzeptu jakin bat adierazteko balio zutelarik, euskarazko hiztegi modernoetan sarrera merezi zutenak baino (*Euskal Herria*).

Azkenik, erdal hitzak ere landu dira (*in situ, in fraganti, a priori*), osagaien esanahia euskaraz —eta gainerako hizkuntzetan— guztiz opakoa izanik, hauek ere nolabait lokuziotzat har daitezkeelakoan.

Esaera zaharrak ere ez dira lan honetan aztergai izan, nahiz badiren horien artean gardenak ("*San Bizente hotza, neguaren bihotza*") zein idiomatikoak ("*zakur zaunkaria ez da horzkaria*"). Esaldi estereotipatuak ("*Hauxe behar genuen!*") eta similak ("*berakatz atala baino finagoa*") tratatzea ere ez da lan honen helburua. Hauen inguruan *"27.173 Atsotitzak - Refranes - Proverbes - Proverbia"* (Garate 1998) liburuan aurki daiteke informazio zabalagoa. Gainera *http://www.argia.com* web-gunera jo daiteke, bertan, Garateren liburuko

atsotitzen kontsultak buru daitezke, horietako batzuen ordainak espainolez, ingelesez eta latinez lortzeko modua ere ematen duelarik.

V.1.3 Aditz konposatuak

Orain arte mugatutako HAULetan zenbait aditz konposatu agertu dira, irizpide semantikoen arabera talde desberdinetan sailkatuak (*eskerrak eman*, *adarra jo*). Aditz konposatu hauek unitatea osatzen dutenez, euren osagaiak elkarrekin lematizatu/etiketatu behar dira.

Baina aditz konposatuen artean *izan* osagaia dutenak (*behar izan, nahi izan, ezin izan eta ahal izan*[9]) ere badira, orain arte aipatu ez diren arren. Izan ere, osagai bat *izan/ukan* lema duten formetako bat da eta aditz hauen formak, aditz trinko —"dirua behar *dut*" eta "ez *dut* dirurik" adibideetan— gisa ez ezik, laguntzaile funtzioan —"dirua ekarri *du*" adibidean— ere ager daitezkeenez, maiztasun handia dute euskal testuetan. Hori dela eta, tratatu beharreko testuetan behin eta berriro agertuko diren lemak dira eta horrek aditz hauek identifikatzerakoan arazo handiak sortaraziko ditu.

Identifikazioa egiteko, lehenengoz HAULaren osagai guztiak agertzen diren egiaztatu behar da, eta HAUL izateko murriztapen guztiak betetzen badira, osagai bakoitzeko agerpen bakarra dagoen begiratuko da. Izan ere, maiztasun handiko osagaien kasuan batez ere, osagairen baten agerpen bat baino gehiago topa daiteke esaldi berean, gainerako osagaien agerpen bakarra dagoen bitartean. Kasu hauetan, hautagaien artean bakarra aukeratu behar da.

Arazo hauek sortarazten dituzten HAUL garrantzitsuenak aditz konposatuak dira, izan ere, unitate hauetako askotan gutxienez osagaietariko baten maiztasuna oso handia da. Orokorrean, bilaketa-esparrua[10] murriztuz agerpen kopurua gutxitu ahal izango da. Hala ere, adibideetan ikusiko denez, argi dago zenbait kasutan, esaldiaren luzera mugatu arren, bilaketan arazoak sor daitezkeela. Hau ulertzeko EEBS corpusetik lortutako honako adibideak aztertzea nahikoa da[11].

(1) ***Ezin nuen*** *sinistu ere, behin eta berriro begiratzen* ***nuen*** *eta* ***ezin nuen*** *sinistu ere.*

(2) *... bi kainabera luzeen mutur-puntak* ***ezin nituen*** *aurkitu, eginahalak egin ondoren* ***ezin*** *nezakeen deus ikusi ...*

(3) *Ohizko egun arrunta* ***zenez****, don Jenaroren klase partikularrera joan* ***behar izan nuen*** *beti bezala.*

9 Aditz hauek beste aditz lexikal batekin doazenean —*etorri nahi dut*—, aditz modaltzat har daitezke, ingelesezko bere ordainen antzera. Hala ere, oraingoz aditz konposatutzat hartzen dira.

10 Bilaketa-esparruaren muga, esaldi bukaerako puntuaz gain (.), galde-ikurrak (?), harridura-ikurrak (!), puntu eta koma (;) eta bi puntuak (:) definitzen dute.

11 Adibideetan beltzez idatzitako hitzak aztertu nahi den HAULaren osagai-hautagaiak dira.

*(4) Oso urduri nengoen klasean eta Fisika-ko ariketetatik **bat ere** ezin izan nuen ongi **egin**.*

Lehenengo adibidea *ezin izan* HAULarena da. Kasu honetan bitan agertzen da HAUL bera, baina bigarren osagaiaren 3 instantzia agertzen dira. Hortaz, lehenengo osagai bakoitzeko bigarren osagai bakarra hautatu beharko litzateke, kasu honetan hurbilen dagoena, hain zuzen.

Bigarren adibidean, aldiz, azken osagaiaren instantzia bakarra (*nituen*) dagoen bitartean, lehenengoa (*ezin*) bitan agertzen da[12]. Kasu honetan ere, hurbilen dagoena aukeratu beharko litzateke.

Hirugarren adibidean are aukera handiagoa dago. Adibide honetan, lehenengo osagaiaren instantzia bakarrerako bigarrenaren hiru instantzia daude eta arestian aipatutako irizpideen arabera *behar* hitzaren ondoan dagoen *izan* instantzia hurbilena aukeratuko litzateke.

Lehenengo hiru adibideetan HAUL bakarraren hautagaiek sortzen dituzte arazoak esaldian[13], baina askotan HAUL bat baino gehiago aurki daiteke esaldi berean eta haien arteko interferentziak lehenengo hiru adibideetan ikusitako arazoari gehituko zaizkio. Laugarren adibidearekin erakutsi nahi dena HAULen arteko interferentzia da[14]. Kasu honetan, *bat ere* eta *bat egin* HAULen arteko interferentzia dago eta *bat* hitza bietako bakarraren osagaia izan daiteke, kasu honetan *bat ere* HAULarena.

Hauek adibide gutxi batzuk besterik ez dira, baina testu errealetan agertzen diren arazoen konplexutasunaren isla dira. Askotan pertsona batentzat begi bistakoa dena modu automatikoan gauzatzea ez da nabaria izaten.

Hala ere, eta arazoak arazo, gainerako HAULekin batera landu dira aditz konposatuak, eta tresnaren deskribapena egitean azalduko da maiztasun handiko osagaiak dituzten HAULek sortarazten dituzten arazoak nola ebatzi diren.

V.1.4 HAULen ezaugarriak

Arestian esan bezala, ezaugarriak lortzeko corpusetako HAULen agerpenak aztertu dira. Batzuetan, HAULaren agerpen guztietan bere osagaiek unitate bat osatzen dutela ziurta daiteke (*hala eta guztiz ere*). Testuingurua kontuan hartu gabe HAUL dela esan daitekeenean

12 Orain arte landutako HAULen artean *ezin izan* lemaren bigarren osagai posibleak *izan* eta *ukan* daude. HAULaren bigarren agerpena identifikatu ahal izateko **ezan* —ezin *dezake* egin— eta **edin* —ezin *daiteke* etorri— lemak ere onartu behar dira bigarren osagaitzat.

13 Lehenengo adibidean, *behin eta berriro* HAULak ez du interferentziarik sortzen *ezin izan* HAULaren agerpenekin.

14 Esaldi berean ezin izan HAULaren bigarren osagaia ere aukeratu beharko da.

unitate ziur gisa definitzen da. Beste batzuetan, berriz, testuingurua ezagutu gabe ezin da hitz anitzeko diren esan. Azken multzo honetan *bat etorri* dago; HAUL honen agerpenetan testuingururik gabe ezin daiteke erabaki unitate lexikal bakarra osatzen duen ala bi lexia diren. Adibidez, "mutil *bat etorri* da" esaldian *bat* eta *etorri* hitzak ez dira unitate lexikal bakarra, baina "guztiak *bat etorri* ziren" perpausean, berriz, lexia bat bakarra osatzen dute.

Bestalde, HAUL batzuen osagaiak ordena berean kokatzen dira agerpen guztietan, eta beste HAUL batzuen osagaiak ordena desberdinean ager daitezke. Aurrenekoaren adibide garbia *hala eta guztiz ere* da. HAUL honen osagaiak beti ordena berean agertzen dira. Beste kasu batzuetan berriz ordena alda dezakete; adibidez, *adarra jo* HAULa: "niri *adarra jo* nahian?" esaldian lemaren ordena berean agertzen dira osagaiak, baina "ez *jo* niri *adarra*, gero!" esaldian kokalekuak aldatzen dituzte.

Azkenik, unitateen osagai batzuk forma flexionatu desberdinetan agertzen dira eta beste batzuk, berriz, forma bakarrean. Adibidez, *bat etorri* HAULaren kasuan *bat* beti forma horretan agertuko da, inoiz ez da agertuko *batean etorri* edo antzekorik, baina *etorri*k, berriz, lemaren edozein forma onartzen du ("*bat dator*", "*bat etorri* ziren", "*bat etortzea* espero dut", etab.).

Fenomeno hauek ikusirik, HAUL bakoitzak datu-basean behar duen informazioa eta ezaugarriak definitzeko orduan hiru irizpide erabiliko dira: batetik, HAUL osoari dagozkion ezaugarriak, bestetik, agerkidetzarekin zerikusia dutenak, eta azkenik, osagai bakoitzaren portaerari dagozkionak.

1. **HAUL osoari dagozkion ezaugarriak**: talde honetan aztertuko den ezaugarri nagusia ziurtasuna da. Baina gainera, unitate osoaren interpretazio morfologikoa definitu behar da HAUL bat identifikatu denean, dagokion analisi morfologikoa eman ahal izateko.
   - Ziurtasuna: unitate bat ziurra dela esango da bere osagaien bestelako analisiak baztertzen dituenean, alegia, bestelako interpretaziorik ez duenean, *ez ezik* adibidez. Bestela, anbiguoa dela esango da; esate baterako, *bat etorri* HAULaren kasuan, testuingurua aztertzen ez denez, ezin jakin daiteke HAULa denentz. Horregatik, gainerako interpretazioak mantentzeaz gain, HAULaren analisia gehitu beharko zaio interpretazio posibletzat, anbiguotasuna gehitu arren.
   - Informazio morfologikoa: HAUL bat identifikatzen denean unitateari dagokion analisia eman behar da. Horretarako, HAULen datu-basean unitateari dagokion informazio morfologikoa landuko da, orokorrean kategoria eta azpikategoria zehaztuz. Analisi hori osagai bakoitzari edo HAUL osoari eman dakioke. Gure aukera osagai bakoitzari informazio osoa eta ordena-zenbaki bat ematea izan da, linealtasuna mantentzearren. Baina zenbait aplikaziotarako osagaiak biltzea interesgarri gerta daitekeenez, bigarren aukera ere aurreikusi da.

2. **Agerkidetzari dagozkionak**: talde honetan aztertuko diren ezaugarriak jarraitasuna eta ordena dira.
   - Jarraitasuna: unitate baten osagai pare bat jarraian doala esango da unitatea osatzen duten osagaiak beti bata bestearen ondoren agertzen direnean, haien artean beste hitzik onartzen ez delarik (*hala ere*). Aldiz, hitzik tarteka badaiteke, unitatea eten daitekeela esango da; adibidez, aditz konposatuek zenbait partikula onartzen dituzte osagaien erdian ("hori *behar* omen *du*").
   - Ordena: unitate baten osagaiek ordena finkoa edo aldakorra izan dezakete. Hortaz, agerpen guztietan ordena bera duten unitateak izango dira batetik (*hain zuzen ere*, *noizean behin*), eta agerpen desberdinetan ordena alda dezaketenak bestetik (*adarra jo*). Bigarren multzo honetako adibide nabarienak aditz konposatuak dira. Unitate hauek orokorrean HAULaren leman adierazten den ordena bera duten arren, aginduzko perpausetan, esate baterako, alderantziz agertzeko joera dute; horrela, "*lo egin* dut" esaten da baiezkoetan, baina "*egizu lo*!" aginduzkoetan.

3. **Osagaiei dagozkienak**: multzo honetan, batetik, osagaiek flexiorik onartzen ote duten aztertuko da eta, bestetik, osagaiak HAULari informazio morfologikoa eransten ahal dion ikusiko.
   - Flexioa: Osagai batzuk forma bakarrean agertzen dira eta beste batzuk flexio bat baino gehiago onartzen dute HAULaren parte direnean. Horregatik, HAULen osagaien flexioari murriztapenak ezartzen zaizkio, ezinezkoak diren formak lehenbailehen baztertzeko asmoz. Forma bakarra onartzen dutenen kasuan, murriztapen argia ezartzen zaio unitatearen osagaiari, bilaketa egiterakoan hitz-formari erreparatzea nahikoa izango da flexio-murriztapenak betetzen dituenentz erabakitzeko.

     Forma bat baino gehiago onartzen dutenen kasuan, berriz, murriztapena ezartzea ez da hain lan erraza. Talde honetan bi kasu bereiziko dira: batetik, edozein flexio onartzen dutenena eta bestetik, flexio-multzo murriztua onartzen dutenena.

     Lehenengo kasuan, osagaiaren lema duen edozein hitz onartuko da unitatearen osagai posibletzat, hau da, ez zaio inongo murriztapenik ezartzen osagaiaren flexioari. Murriztapen ezak tratamendua erraztuko du, hitzaren lema HAULaren osagaiarenarekin bat datorren egiaztatzea nahikoa izango delako. Hala ere, errazago gerta daiteke esaldi berean lema bera duen hitz bat baino gehiago topatzea, hortaz, HAULaren osagaia horietako zein den erabakitzea zailduko du.

     Flexio-multzo murriztua onartzen dutenek, berriz, definizio eta tratamendu konplexuagoa izango dute, datu gehiago egiaztatu behar izango baitira. Dena dela, egiaztapen horiek ziurtasun handiagoa emango dute HAULaren osagaia den erabakitzeko momentuan eta gainera, zailago gertatuko da esaldi berean hitz-forma bera behin baino gehiagotan topatzea.

- Informazio morfologikoa: HAULa flexioren batean ager daitekeenean, osagaietako batek emango dio, kategoria eta azpikategoria izan ezik, gainerako informazio morfologikoa. Hau ez da kasu guztietan behar izango, baina lemaren edozein flexio onartzen denean, flexio horren informazio morfologikoa HAUL osoarena ere izan ohi da. Horregatik, kasu horietan HAULaren informazio morfologikoa osatzeko osagai baten informazioa erabiliko da. Esate baterako, *bat egin* HAULaren agerpenetan *egin* osagaiak emango dio aditz konposatuari gainerako informazioa, hala nola, aspektua, erlazio atzizkiak, etab.

Lehen hurbilpenean HAUL etenen osagaien arteko distantziaren mugarik ez zen ezarri. Hau da, HAUL baten osagaien artean edozein hitz kopuru onartzen zen, beti ere bilaketa-esparrura mugatuta. Ordenari dagokionean, bi osagairen arteko posizio aldaketak zehazten ziren soilik, eta hau bi osagai besterik ez dituzten HAULetarako nahikoa bazen ere, osagai gehiago dituztenean konbinazio posibleak gehiago ere izan daitezke.

Gainera, corpus azterketan sakonduta, HAULen informazioa are gehiago zehaztu daitekeela ikusi da, eta horrela bilaketaren emaitzak hobetu daitezkeelakoan gaude, bereziki HAUL anbiguoen kasuan. Esate baterako, ziurtasunari buruzko informazioa HAULaren ordena/jarraitasun konbinazio guztietarako berbera izan beharrean, kasu bakoitzean lantzen bada, lehen hurbilpenean anbiguo ziren HAUL batzuen agerpenak ziurtzat eman daitezke. Modu horretan, emaitzaren zehaztasuna hobetzen da, osagaien interpretazioak baztertu eta irakurketa bakarra utziko delako.

Esate baterako, *aditzera eman* HAULaren kasuan "*aditzera emango* zuen" bezalako agerpenetan ziurtzat eman daiteke. Baina "ez zuen *eman* berri zehatzik *aditzera*" edo "sarrera *eman* zenidanez hitzaldia *aditzera* joan nintzen" modukoetan ordena aldatu eta tartean hitz bat baino gehiago txertatzea dagoenez, zailago gertatzen da HAUL ote den erabakitzea. Ikus daitekeenez, lehenengoan HAULa da eta bigarrenean ez, biek tartean hitz kopuru berbera duten arren.

Beraz, jarraitasun eta ordena posible bakoitzeko ziurtasun balioa eta osagaien flexio murriztapenaren adierazpena emango da tratamenduaren emaitza ahalik eta zehatzena izan dadin. Ordena eta jarraitasunari buruzko informazioa modu egokian emateko adierazpideari gauzatze-eskema esango zaio.

Ordena adierazteko osagai bakoitzak HAULaren leman duen posizioa erabiliko da eta jarraitasunerako, berriz, karaktere berezi batzuk. Osagaiak bata bestearen ondoan agertzen badira, posizioak zuzenean idatziko dira. Adibidez, *hala ere* HAULaren gauzatze-eskeman *12* jarriko da, osagaiak jarraian eta leman agertzen diren ordena berean gauzatzen direlako testuetan. Osagaien artean zero hitz edo hitz bat onartzen bada, galde-ikurra (?) tartekatuko da

osagaiak adierazten dituzten zifren artean. Adibidez, *negar egin*en ordena posibleetako bat *21* da, baina tartean zero hitz edo hitz bat onartzen duenez, *2?1* espresioaren bidez adieraziko da. Azkenik, zero hitz edo gehiago onartzen direnean izartxoa (*) tartekatuko da.

Gainera, arestian esan bezala, gauzatze-eskema bakoitzari lotuta, osagai bakoitzaren flexio murriztapenen adierazpen bat ematen da. Adierazpen hori zehazteko honakoari jarraitzen zaio:

- Forma finkoan agertzen denean, murriztapena '-' izango da.
- Flexio posible guztiak onartzen direnean, murriztapena '+' izango da.
- Flexio-multzo bat besterik ez badu onartzen, orduan, ezaugarri-balio bikoteen bidez adieraziko da, adierazpena osatzeko *and*, *or* eta *not* eragile logikoak erabili ahal izango direlarik. Adibidez, absolutibo mugagabean edo partitibo mugagabean ager badaiteke, orduan honako adierazpena esleituko zaio:

  ((KAS = ABS) *and* (MUG = MG)) *or* ((KAS = PAR) *and* (MUG = MG))

Ondorengo adibideen bidez gauzatze-eskemen esanahia azalduko da. Osagaiak jarraian eta ordenan agertzen direnean, gauzatze-eskema bat nahikoa izango da HAULaren portaera deskribatzeko:

(hala eta guztiz ere, 1234, +,(-, -, -, -))

Hauxe da kasurik sinpleena eta *hala eta guztiz ere* HAULaren osagaiak beti ordenan eta jarraian agertzen direla (*1234*), ziurra dela (+) eta bere osagaiek bestelako flexiorik onartzen ez dutela (-) adierazten du.

Horrelakoetarako lehen hurbilpenean ematen zen informazioarekin nahikoa zen, baina ondorengo adibideek HAULaren informazioa zehazteko gauzatze-eskemek eskaintzen dituzten baliabide guztiak erakusten dituzte:

```
(begi bistan egon, 123, +,
     (((KAS = ABS) and (MUG = MG)) or ((KAS = GEN) and (NUM = PL))), -, +))
     "begien bistan zegoen honela bukatuko genuela"

(begi bistan egon, 312, +,
     (((KAS = ABS) and (MUG = MG)) or ((KAS = GEN) and (NUM = PL))), -, +))
     "ez dago begi bistan"

(begi bistan egon, 3?12, +,
     ((KAS = ABS) and (MUG = MG)), -, +))
     "ez dago horren begi bistan, ez dezazula pentsa"
```

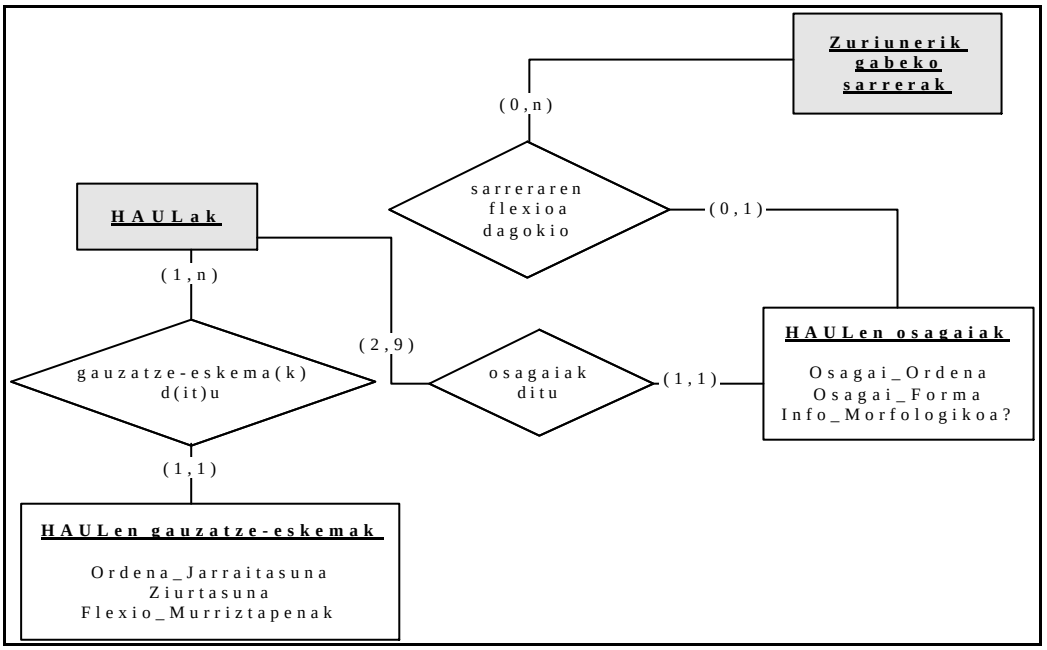
Adibideotan, *begi bistan egon* HAULa kasu guztietan ziurra den arren, osagaiak jarraian agertzen direnean (*123* eta *312*) *begi*k absolutibo mugagabea eta genitibo plurala onartzen dituen bitartean, hitz bat tartekatzen denean (*3?12*) absolutibo mugagabea besterik ez du onartzen. Gainerako osagaiei dagokienean, kasu guztietan *bistan* hitz-forma (-) agertu behar da eta *egon* aditzaren edozein forma (+) onartzen da.

Azkenik, *aditzera eman* HAULa lehen hurbilpenean anbiguo zen bitartean, gauzatze-eskemak erabilita anbiguoak diren kasuak muga daitezke. Horrela, agerpena *12* gauzatze-eskemarekin bat datorrenean ziurtzat emango da:

```
(aditzera eman, 12, +,(-, +))
     "aditzera emango zuen"

(aditzera eman, 2*1, -,(-, +))
     "ez zuen eman horrelako astakeriarik aditzera"
```

Ezaugarri horiek guztiak kontuan izanik, HAULen informazioa datu-basean gordetzeko diseinua V.1 irudiko entitate-erlazio eskemaren araberakoa da. Datu-basearen diseinuari buruz sakontzeko (Agirre *et al*. 1994-a; Aldezabal *et al*. 1999-b) barne-txostenak azter daitezke.



V.1 irudia.- HAULen datu-basearen entitate-erlazio eskema.

V.1.5 Bestelako unitateak

Esan bezala, HAUL eta sintaxiaren arteko muga ez da oso argia. Baina lan honetarako definitu behar izan da. Ondoren hortxe dauden zenbait hitz anitzeko egitura aurkeztuko dira eta mugaren zein aldetan jarri aztertuko da.

Lan batzuetan, zenbait egitura sintaktiko ere (*zenbat eta ...-ago ... [orduan eta] ...-ago*) HAUL bezala lematizatu ohi dira eta lehen hurbilpenean halaxe tratatu nahi izan genituen. Izan ere, talde honetan sartzen diren egiturak maila honetan tratatzea oso interesgarria da, batetik, bere osagaien esanahia loturik baitago, eta, bestetik, azterketa sintaktikoa egiterakoan

perpaus-mota batzuk jadanik identifikatuak geratuko bailirateke; adibidez, *zenbat eta ...-ago* detektatzen bada, analisi sintaktikoarekin hasi aurretik konparaziozko perpausa markatua egongo litzateke. Baina beste lan batzuetan egitura horiek sintaxi mailan ere tratatzen dira. Lan honen helburua ez da bi maila hauen arteko muga definitzea, baizik eta, bietako batean egitura konplexuak tratatzea.

Egitura hauek automatikoki tratatzeko modurik zuzenena eta errazena gramatika baten bidezkoa da. Egitura hauen patroiak definitu eta gramatikaren erregeletara ia zuzenean itzul daitezke; baina gainerako HAULekin batera gramatika berean sartu nahi izango bagenitu, ordea, erabat ulergaitza eta trata ezina bihurtuko litzateke; izan ere, egituraren erdian bestelako HAULak ager daitezke eta horrek gramatika korapilatu egingo luke. Horregatik, mugan dauden egitura eten hauek prozesutik at utzi dira, baina badira egitura jarrai batzuk, izaera sintaktikoa izan dezaketen arren, identifikaziorako horrenbeste arazorik sortzen ez dutenak.

Egitura hauen artean zenbakien adierazpenak, datak eta izen berezi konposatuak aurkitzen dira. Edozein aplikaziotarako unitate hauek identifikatzea eta tratatzea komeni den arren, berez, unitate hauek gehienbat informazioaren erauzketarako sistemetan tratatzen dira, *named entity task* edo izendun entitateen ezagupenerako atazan (Chinchor 1997).

Zenbaki eta daten kasuan, ezaugarri nagusienetako bat zera da, oso egitura finkoak jarraitzen dituztela eta hiztegi itxia darabiltela; horregatik, modu errazean identifika daitezke. Izen berezien kasuan, ordea, ez da horrelakorik gertatzen.

Izen bereziek hiztegi oso irekia erabiltzen dute eta tratamendu orokorra diseinatu nahi bada, ezin daiteke euskal izenetara mugatu. Askotan, izenen zerrendak (*gazeteer*) erabiltzen dira izen bereziak identifikatzeko. Gehienetan, zerrendaren muga dela medio, tratamendu osagarriak erabili behar izaten dira zerrendetan agertzen ez diren izen berriak tratatu ahal izateko. Horrela, izen berezi asko prozesatu ahal izango dira, orokorrean data eta zenbakien tratamenduak baino zehaztasun eta estaldura txikiagoa lortu arren.

V.2 Tratamendu automatikoa

Tratamenduari dagokionean, hitz anitzeko unitateek eskatzen duten prozesamenduari erreparatuz bi multzotan sailkatu ditugu. Batetik, ondo definitutako identifikazioa eta tratamendu zehatza dutenak izango ditugu, hau da, bai unitate itxiak, datu-basean sarrera izango dutenak, baita hiztegi itxia erabilita egitura finkoa dutenak ere. Bestetik, gainerakoak unitate irekiak izango ditugu, tratamendu orokorragoa eskatzen dutenak.

Lehenengo taldean, datu-basean sarrera izango duten lokuzioak eta kolokazio murriztuak daude. Hauen kasuan, osagai guztien edo batzuen esanahia edo interpretazioa galdu eta berria sortzen dute hitz anitzeko unitatea osatzen dutenean. Hori dela eta, lexikalizatuak izaten dira eta sarrera merezi dute datu-base lexikalean. Prozesamenduari dagokionean, unitate osoari eta osagaiei buruz aipatutako ezaugarriak kontuan izanda, guztiak identifikatu eta tratatuko dituen programa diseinatu eta inplementatu da, aurrerago ikusiko den bezala.

Gainera, datak eta zenbakien adierazpenak ere lehenengo multzo honetan kokatu ditugu. Hitz-elkarketa eta kolokazio irekiek bezain multzo irekia osatzen ez duten arren, ez da komeni dagokien informazio guztia datu-basean lantzea. Horretarako, zenbaki posible guztiak kalkulatu eta landu beharko lirateke, hartara benetan testuetan agertuko diren zenbakiak baino askoz gehiago sortuz[15].

Hala ere, zenbakiek oso egitura finkoa dute eta hiztegi murriztua darabilte. Horregatik, formari bakarrik erreparatuta eta zenbaki eta hilabeteen lexikoi txiki bat erabilita guztiak egoki identifika daitezke[16].

Ondoren, identifikatu den unitatea tratatzeko, nahikoa da esleituko zaion kategoria eta azpikategoria erabakitzea, gainerako informazio morfologikoa azken osagaiak emango baitio, horixe izango baita flexioaren berri emango duena. Hortaz, data eta zenbakien identifikazioa aldez aurretik eginez gero, gainerako hitz anitzeko unitate itxiekin batera trata daitezke.

Bigarren taldeari dagokionean, osagaien esanahi eta interpretazioetatik abiatuta unitate osoarena lor daiteke. Multzo oso zabala eta erabat irekia osatzen dutenez, tratamendu orokorragoa behar dute, datu-basean unitate hauei dagokien informazioa landu gabe aplika daitekeena.

Datu-basean ez lantzearen arrazoia nagusiak honako hauek dira:

- datu-basea etengabe aberastu beharko litzateke.
- datu-basea inoiz ezin izango litzateke osotzat hartu.
- tratamenduaren eraginkortasunean galduko litzateke, datu kopurua handituz, programak memoria eta denbora gehiago beharko luke aukera guztiak aztertu eta testua tratatzeko.

Beraz, hauek identifikatu eta osagaien informaziotik automatikoki hitz anitzeko unitatearen informazioa lortzeko beste prozesu bat diseinatu eta inplementatu beharko da sistema sendoa izango bada.

---

15 Gainera, ezinezkoa da zenbakien konbinaketa posible guztiak lantzea eta nonbait muga jarri beharko litzateke.

16 Guztiak ez badira ere, gehien gehienak behintzat identifika daitezke. Izan ere, zenbakiei eta datei dagokienean, idazkeran aldaera asko erabili ohi da oraindik orain.

Hitz-elkarketa eta kolokazio irekien tratamendua planteatzean, identifikazio-fasea izango da prozesuaren alde konplexuena. Izan ere, multzoa hain zabala izanik, elkarren ondoan agertzen diren hitz konbinaketa posible guztien artean tratamendua merezi dutenak aukeratu behar dira soilik. Hortaz, identifikaziorako prozesuak testuko agerkidetza guztiak eta osagaien informazio morfologikoa kontuan izanik, maiztasun handikoak bakarrik aukeratu beharko lituzke.

Hau egin ostean, hitz anitzeko unitateari dagokion informazio morfologikoa nola erauzi erabaki eta burutu beharko da. Elkarketen kasuan, tratamendua informazio morfologikoa aukeratzea errazagoa izango da, gehienetan kategoria bereko osagaiak izango dituztelako, baina bestelako unitateekin ez da hain berehalakoa. Hori dela eta, kolokazio ireki batzuk datu-basean lantzea erabaki da, kontzeptu jakin bat adierazteko balio dutelarik, euskal hiztegietan sarrera dutenez, tratamendu errazagoa aplikatzearren.

Azkenik, hitz anitzeko unitate irekiekin amaitzeko, izen berezi konposatuen identifikazioa ere burutu behar da. Izen berezien kasuan identifikazioa idazkeran oinarrituta burutu behar da, osagai guztiak maiuskulaz hasita idatzi beharko baitira[17]. Data eta zenbakien antzera, informazio morfologikorik gabe egin daiteke identifikazioa, baina kasu honetan ez da lexikorik erabili behar. Horrek identifikazioa are konplexuago egiten du, idazkerari erreparatuz gero, testuaren osagai batzuk beti maiuskulaz hasita —edo guztiz maiuskulaz— idazten direlako, tituluak kasu. Hala ere, konplexutasuna ez da identifikazio-fasean amaitzen. Tratatu ahal izateko, lortutako izen bereziak sailkatu egin behar dira, pertsona-izen, leku-izen edota erakunde diren jakiteko. Horretarako, ikasketa automatikoan oinarritutako teknikak zein erregela-sistemak erabili ohi dira, aurrerago ikusiko den bezala.

Oraingoz, unitate itxiak tratatzeko tresnak garatu dira eta izen berezien prozesamendua burutzeko lehen urratsak eman dira, V.3 atalean azaltzen den bezala. Baina euskararako burututako lana aurkeztu aurretik, hitz anitzeko unitateak tratatzeko hainbat hurbilpen aurkeztuko dira ondorengo atalean.

V.2.1 Hitz anitzeko unitateen tratamendurako hurbilpenak

Lengoaia naturalaren prozesamenduan hitz anitzeko unitateen tratamendua ikerkuntza-lerro irekia da. Orain artean proposamen gutxi egin dira arlo honetan eta bibliografia aztertzerakoan hitz anitzeko unitate itxiei buruzko lan gehienak hiztegigintzari lotutakoak

[17] Badira salbuespenak ere. Esate baterako, zenbait abizenek minuskulaz idatzitako osagaiak dituzte (*Lopez de Ipiña, de la Calle, etab.*).

direla ondorioztatu dugu. Lan hauetan, kolokazio eta adierazpen idiomatikoak zein sarreratan kokatu behar diren aztertzen da, eta, beharrezkoa dela erakutsi arren, ez da tratamendu automatikorako proposamen gehiegirik egiten (Krishnamurthy 1996).

V.2.1.1 Hitz anitzeko unitate itxiak

Hitz anitzeko unitate lexikal edo itxien tratamendurako bi muturrak hauexek dira: zerrenda batean biltzea batetik, eta sintaxiaren arazotzat hartzea bestetik. Lehenengo multzoan bi aukera aurki daitezke: morfologiaren aurretik edo etiketatzearen ondoren tratatzea. Lehenengo aukeraren adibide gisa UPCko taldearen gaztelerarako analizatzaile morfosintaktikoa (Carmona *et al*. 1998) eta Finlandiako taldeak (Karlsson *et al*. (*eds.*) 1995) lanean ingeleserako deskribatzen direnak aipa daitezke.

Hurbilpen hauetan espresio konplexuak lehenbailehen identifikatzearen beharra ikusi da. Karlsson-ek (Karlsson *et al*. (*eds.*) 1995) halaxe azaltze du, morfologiaren aurretik tratatzeak bere ustez morfologia eta sintaxiarekin hasi aurretik HAULak identifikatzeak morfologiaren emaitza zehatzagoa lor daitekeelako.

UPCko taldearen analizatzaile morfosintaktikoak (MACO+) datak, laburdurak, izen bereziak, zenbait konposatu (*'sin embargo'*, *'no obstante'*, ...), zenbakizko adierazpenak eta puntuazio ikurrak tratatzen ditu tokenizazioan. Modulu hauetan laburdura, izen berezi —pertsona, toponimo, marka, enpresa, etab.en izenak—, hitz anitzeko konposatu, etab.en zerrendak erabiltzen dira, nahiz eta, izen berezien kasuan behinik behin, prozesua zerrendetan bilatzea baino konplexuagoa izan, ikasketa automatikoan oinarritutako teknikak eta erregelak erabiliz.

Finlandiako taldearen kasuan, sintagma finkoen (*'fixed syntagms'*) zerrenda hutsak erabiltzen dituzte tokenizazioaren ostean maiztasun handiena dutenak biltzeko. Multzo nagusia sintagma preposizionalek osatzen dute, baina zenbait konposatu ere tratatzen dira. Heikkilä-k (Karlsson *et al*. (*eds.*) 1995:133-137) aipatzen duenez, COBUILD hiztegiko 5000 konposatu landu eta testuetatik lortutako beste 600 inguru gehitu dituzte, gehienak izenak direlarik.

Etiketatzearen ondoren tratatzen dutenen adibidea CLAWS4 sistema (Leech *et al*. 1994) da. Sistema honetan hitz anitzeko espresioak etiketatzearen beharra ikusten den arren, arazoa sakonean aztertu gabe irtenbide erraza ematen zaio eta horretarako sistema honek IDIOMTAG programa erabiltzen du. Aplikazio honek hitz anitzeko espresioen zerrenda erabiltzen du testuan bilaketa egiteko. Zerrenda horretan espresio konplexuan parte hartzen duten hitzak lantzen dira, dagokien etiketarekin nahi izanez gero. Programak espresioa aurkitzen duenean hitzetariko bati edo guztiei etiketa aldatzen die. Honela, desanbiguazioan

sortzen diren zenbait arazoen ebazpidea lortzen da. Hurbilpen hau Finlandiako taldearen antzekoa da, baina testu desanbiguatuaren gainean lan egiten du.

Beste muturrean, formalismo sintaktikoak erabiltzen dituztenak daude. Sistema hauek proposatzen dituztenek hitz anitzeko espresioen tratamendua arazo sintaktikotzat hartzen dute. Baterakuntzan oinarritutako sistemak erabili izan dira, baina gai hori lan honetatik kanpo geratzen denez, ez dugu horretan sakondu.

HAULak sintaxiaren aurretik identifikatzearen beharra onartuz, baina tratamenduan zerrendetan adieraz daitezkeenak baino xehetasun handiagoak aurreikusiz, zenbait proposamen interesgarri ere aurkeztu dira. Horien artean, gure ustetan bibliografian aurkitutakoen artean hitz anitzeko unitateen multzo zabalena tratatzen duena, orokorrena eta interesgarriena *Xerox*eko taldearen proposamena da, egoera finituko teknologian oinarritutakoa (Segond eta Tapanainen 1995; Segond eta Breidt 1995; Breidt *et al*. 1996).

Xeroxen garatutako sisteman, hitz anitzeko espresioak definitzeko egoera finituko teknologia erabiliz erregela lokalak idaztea proposatzen da. Orokorrean hitz anitzeko espresioak detektatzerakoan kontuan izan behar da forma desberdinetan ager daitezkeela testuetan. Segond eta Tapanainenek (1995) euren txostenean hiru aldaketa-mota bereizten dituzte:

- Aldaketa morfologikoak. Espresio batean lema beraren forma desberdinak onar daitezke, adibidez, absolutibo singular eta partitiboa (*adarra/adarrik jo*).
- Aldaketa lexikalak. Espresio batean hitz bat baino gehiago onar daiteke (*ezin izan/*edun*).
- Egitura-aldaketa. Espresio batzuek egitura-aldaketaren bat onar dezakete, adibidez, posizio-aldaketa (*ezin izan -> izan ezin*).

Ezaugarri hauek deskribatzeko bi maila erabiltzen ditu: azaleko maila eta maila lexikoa. Bi maila hauek analizatzaile morfologikoan azaldu ditugunak dira, baina analizatzailean bai azaleko mailan bai lexikoan karaktereak bereizten ziren. Kasu honetan, analizatzaile morfologikoan ez bezala, azaleko mailan hitz-formak izango ditugu eta lexikoan lemak.

Hitzak banaka deskribatzerakoan aurretik esan dugunaz gain informazio gehiago eman dezakegu. Formalismo honek oinarrizko lau deskribapen eskaintzen ditu hitzari dagokion informazioa adierazteko[18]:

- :azaleko_forma (adib. *:bat*)
- :azaleko_forma aldagai_morfologikoa: (adib. *:egin ADI:*)

[18] Deskribapenean bi puntuak (:) lexiko-mailakoa (*IZE:*) edo azaleko mailakoa (*:ohi*) den adierazteko erabiltzen dira.

- lema aldagai_morfologikoa: (adib. *ezin IZE:*)
- hitz-klasearen_aldagaia (adib. *ADI*)

Lehenengo bien bitartez hitz-formak deskriba ditzakegu. Hirugarrena lemak deskribatzeko erabiltzen da. Eta azkenekoa klase bateko hitz guztiak onar daitezkeenean. Adibidez, *ezin izan* aditz konposatuaren osagaien artean partikulak bakarrik onartu nahi baditugu, honela deskribatuko genuke:

ezin IZE: (PRT) izan ADI:

Adierazpen erregularren bidez modu errazean deskriba daitezke hitz anitzeko unitate lexikalak. Aldaketa morfologiko zein lexikoaren kasuan erregela bakarrean eman daitezke posibilitate guztiak. Egitura-aldaketaren kasuan, berriz, egitura posible bakoitzeko erregela bat eman beharko da.

Behin erregelak definiturik, konpilatu egin beharko dira egoera finituko transduktore bat lortzeko. Horretarako, sistemaren osagai bat lexikoa izango da. Konpilazio-prozesuan erregeletan definitu ditugun aldagaien instantzia guztiak lexikotik lortuko dira eta balio horiekin transduktorea sortuko du sistema honek.

Hau da, arestian azaldu dugun adibideko erregela izanik, PRT etiketa duten hitz forma guztiak (*omen, ote, ohi*...) lortuko lituzke sistemak, erregela honela aldatuz:

ezin IZE: ([:omen | :ote | :ohi | ...]) izan ADI: [19]

Horrelako erregela bat konpilatzean, transduktoreak partikula bakoitzeko erregela bat izango du. Kasu honetan ez dira asko, baina partikula bat izan beharrean adberbio bat izan balitz, sistemak lexikoan dituen adberbio guztiak sartu beharko lituzke eta erregela-multzoa handituz joango litzateke.

Gainera gogoratu *ezin izan* HAULaren osagaiek ordena alda dezaketela. Hortaz, aukera hori adierazteko beste erregela bat idatzi beharko genuke:

izan ADI: (EDOZER)* ezin IZE: [20]

Horrelako erregela batetik abiatuta konpilatzaileak lortuko lukeen transduktorearen tamaina izugarria izango litzateke. Izan ere, *izan* eta *ezin* lemen artean nahi adina token egon daitezke, esaldi-bukaerako markak alde batera utzirik, jakina. Dena dela, Xeroxen garatutako

19 Mako ([ ]) artean aukerak multzokatzen dira, eta marra bertikala (|) aukerak bereizteko erabiltzen da.

20 Erregela honetan erabili dugun EDOZER aldagaiak edozein hitz onartuko luke. Izartxoak, berriz, EDOZER behin edo gehiagotan ager daitekeela esan nahi du, baina posible da behin ere ez agertzea.

sistemak transduktoreak optimizatzen ditu, erregela kopuru handia izanda ere, aplikazioa azkarra izan dadin.

Laburbilduz, espresioak definitzeko modu honek aukera dezente ematen digu: lehenengo, hitz-forma eta lemak modu sinplean adierazten dira; gainera, egitura sintaktikoak definitzen errazak dira; eta azkenik, bai osagaiei bai egituretan ager daitezkeen hitzei murriztapen zehatzak ezar diezazkiekegu. Inplementazioari dagokionean, erregela kopurua handitu arren, aplikazio azkarra garatu da. Beraz, formalismo honen bidez V.1 atalean azaldu ditugun ezaugarri gehienak modu errazean adieraz daitezke eta testuetako HAULak bilatzen ez da denbora asko erabiltzen.

Baina aztertu ditugun hurbilpen gehienetan bezala, osagaien arteko orden aldaketa ez da erraz adierazten. Kasu horietan erregelak bikoiztu egin beharko genituzke osagai trukagarri pare bakoitzeko eta gure ustez, honek idatzi beharreko erregela kopurua izugarri handituko luke. Izan ere, orain artean aztertu ditugun HAULen multzo handi batek fenomeno hori aurkezten du eta ordena aldaketa gertatzen denean osagaien erdian edozein hitz joan liteke. Ordena aldaketari buruz kapitulu osoan zehar hitz egingo dugu, ezaugarri hau inplementazioa erabakitzerakoan garrantzitsuena delako, baita arazo gehien sortarazi duena ere.

Tresna honen azken hurbilpenean (Breidt *et al*. 1996), no labaiteko irtenbidea eman diote; ordena posible guztiak adierazteko makroak definitu ahal dira. Makro horietan osagaiek har ditzaketen posizioak adierazten dira. Horrela, ordena posible guztiak erregela batean idatzi eta dagokien espresioei automatikoki aplikatuko zaie konpilazio-prozesuan, eskuz egin beharreko lana benetan murriztuz.

Hurbilpen honek oso tratamendu sofistikatua egiten du, eta literaturan aztertutako sistemen artean estaldura handiena du kasuistikari dagokionean. Gainera programa azkarra eraiki dute. Baina programa hori ez da publikoa eta horregatik HAULak identifikatzeko hurbilpen baten beharrak hortxe darrai. Gaur egun tresna horiek erabiltzeko aukera dugunez, etorkizunean hitz anitzeko unitate guztiak egoera finituko teknologia honekin tratatzeko aukera aztertuko dugu. Bitartean, gure beharrei aurre egiteko V.3 atalean aurkezten den tresna propioa diseinatu eta inplementatu dugu.

V.2.1.2 Hitz anitzeko unitate irekiak

Hitz anitzeko unitate irekietan bi multzo bereizi ditugu arestian, batetik, hitz-elkarketa eta kolokazioak eta, bestetik, izendun entitateak. Lehenengo multzoari dagokionean, esan bezala, neurri estatistikoen bitartez interesgarriak diren agerkidetzak erauzi eta tratatzea da hartu beharreko bidea.

LNPko komunitatean onarpen handia izan duen hurbilpena aurkeztuko da ondoren, Smadja-k (1993) landutako Xtract sistema, alegia. Beste hurbilpen bat ere aipatu nahiko genuke, (Silva *et al*. 1999) lanean aurkezten den LocalMax algoritmoa, hain zuzen ere. Oraingoz hainbesteko hedapena izan ez duen arren, neurri desberdinekin burututako saiakuntzetan (Dias *et al*. 2000-b) algoritmoa eurek proposatutako *mutual expectation* neurriarekin batera aplikatuz, nahiko emaitza onak lortzen dituzte.

Sistema hauetan erabiltzen diren tekniken inguruan sakontzekotan (Schone eta Jurafsky 2001-a) lana ere kontsulta daiteke. On-line hiztegien aberasketa automatikoa egin ahal izateko teknika automatikoak aztertzea eta hitz anitzeko unitate berrien sarrerak aukeratzeko teknika egokirik ote dagoen erabakitzea da lan honen helburu nagusia[21]. Bertan, neurri bakoitzaren aldeko eta aurkakoak aztertzen dira, saiakuntza konparatiboen emaitzak ere aurkezten direlarik. Izendun entitateei dagokienean, proposamen ugari egin da, gehien bat MUC konferentziari lotutako sistemak. Ondoren, multzo bakoitzari atal bana eskaintzen zaio.

V.2.1.2.1 Kolokazioen erauzketarako sistema: Xtract

Aipatu bezala, onarpen handia izan duen sistema bat Xtract da, Smadjak (1993) garatutakoa. Sistema honek kolokazioak erauzten ditu automatikoki corpusetatik. Sistema garatu eta probatzeko 10 milioi hitzeko corpusa erabili du. Smadjaren lanean identifikatu nahi diren kolokazio lexikalak bi mota hauetakoak dira:

- Kolokazio lexikal malguak: modifikatzaile-modifikatu erlazioaz lotutako bi hitzek osatzen dituzte. Hitzek flexioak onar ditzakete, ordena alda daiteke eta hainbat hitzez banaturik ager daitezke.
- Kolokazio konposatuak: oso modu finkoan agertzen diren bi hitz edo gehiagok osatzen dituzte. Ez duten flexiorik onartzen, ordena finkoa da eta jarraian agertzen dira beti.

Xtractek hiru fase dauzka. Lehenengoan informazio estatistikoan oinarriturik bi hitzeko hautagaien zerrenda lortzen da. Bigarrenean, berriz, bi hitz baino gehiagoz osatutakoak lortzen dira. Eta, azkenik, hirugarrenak kolokazioen egitura sintaktikoa lortzea du helburu.

Lehenengo pausoan hitz bikoteak (hitzak eta etiketak kontuan izanik) lortu eta dagokien informazio estatistikoa lortzen da. Bikoteak lortzeko ±5 hitzeko leihoa erabiltzen du, hots, hitzen artean gehienez 5 hitzeko distantzia egon daiteke, eta ez dira hitz guztiak kontuan hartzen, klase irekietakoak baino ez dira erabiltzen.

[21] Teknika egoki esatean, paperezko hiztegietan erabiltzen diren irizpide berberak mantenduta sarrera berriak proposatuko dituen teknika esan nahi da lan horretan. Nahiz eta on-line dauden hiztegiek ez duten paperezkoek duten espazio-arazorik, ez dute hiztegia neurririk gabe handitu nahi.

Informazio estatistikoa lortzeko elkarrekiko informazioa edo *mutual information* (Church eta Hanks 1990) formula erabil daiteke:

$$MI(a,b) = \log_2 \left(\frac{P(a,b)}{P(a)P(b)} \right)$$

P(a) eta *P(b)* osagai bakoitzaren corpuseko agerpen-probabilitatea diren, eta *P(a,b)* jarraian edo gertu agertzeko duten probabilitatea den. (Manning eta Schütze 1999) lanean, kolokazioei eskainitako kapituluan, neurri honen erabilerari buruzko kritika egiten da, *mutual information* hautagaiak aukeratzeko baino baztertzeko neurri egokiagoa delako. Dena dela, maiztasun txikiko hautagaiak aldez aurretik baztertuz gero emaitza onak lor daitezke. (Dunning 1993) lanean ere maiztasun txikiko agerkidetzen kasuan elkarrekiko informazioak gainestimazioa egiten duela aipatzen da.

Neurri horren ordez, egokiagoak izan daitezkeen beste batzuk ere aipatzen dira, hala nola, Pearson-en c^2 hipotesi-testa, test horretan oinarritutako f^2 (Gale 1991), *Log-Likelihood ratio* (Dunning 1993), Dice-ren koefizientea (Dice 1945) eta *normalized expectation* kontzeptuan oinarritutako *mutual expectation* (Dias *et al*. 1999-b). Neurri hauen guztien konparaketa, *mutual information* neurriarekin batera, (Dias *et al*. 2000-b) lanean aurki daiteke LocalMax tresnan aplikatuta.

Dena dela, *mutual information* neurriaren murriztapenak, aipatutakoez gain, handiak dira: batetik, bi hitzei dagokien neurria da, eta ondorengo urratsetan bi hitz baino luzeagoak ere tratatu nahi dira; bestetik, ez ditu benetako kolokazioak identifikatzen, elkarren ondoan oso maiz agertzen diren hitz bikoteak baizik —ikus (Smadja 1993:150)—. Hori dela eta, hautagaiak baztertzen joateko hiru baldintza erabiltzen ditu, distribuzioaren parametroak, batez besteko maiztasun eta probabilitateak banakakoekin alderatuz eta hitzen arteko distantziak kontuan izanda[22].

Bigarren urratsean, bigrametatik n-grametara luzatzen saiatzen da sistema. Lehenengo urratseko bigramak oinarritzat hartuz, osagaien inguruko hitzen (eta etiketen) agerpenak aztertzen ditu. Kasu honetan, maiztasunetan oinarritzen da soilik hitz anitzeko unitate bat aukeratzeko, ez du analisi sakonagorik egiten.

Hirugarren eta azken urratsean informazio sintaktikoa gehitzen die kolokazioei. Emaitza hau lexikografoentzat da bereziki interesgarria, kolokazioak egitura sintaktikoaren arabera

[22] Smadjak *strength*, *spread* eta *distance* izendatzen ditu hiru baldintzetan kontuan hartzen diren neurriak. Beste lan batzuetan *z-score* izendatzen dute neurri-multzo hau.

aztertu ahal izateko. Horretarako, etiketatutako testuan azaleko sintaxia aplikatzen du kolokazioen egitura sintaktikoa lortu ahal izateko.

Sistema honek ondo funtzionatzeko corpus handia behar du, maiztasun txikiko kolokazioen kasuan ez dituelako emaitza onak ematen[23]. Gainera, corpus orekatuen gainean lan egitea komeni da, lortutako emaitzak nahiko orokorrak eta domeinuaren independente izateko.

V.2.1.2.2 Izendun entitateen tratamendurako tresnak

Aipatu den bezala, MUCen izendun entitateen artean hainbat mota bereizten dira (Chinchor 1997), horien artean nagusienak honako hiruak dira: daten adierazpenak, zenbakien adierazpenak eta izenen adierazpenak. Tratamendua diseinatzerakoan bi arazo ebatzi behar dira: lehenengo, entitateak mugatu behar dira eta, ondoren, mota desberdinetan sailkatu.

Entitateak mugatzeari dagokionean, data eta zenbakien kasuan hasiera eta amaiera identifikatzea nahiko erraza da, idazkera oso zehatza dutelako, baina izenen kasuan ez da berehalakoa. Izan ere, izen bereziak maiuskulaz idatzitako sekuentziak aztertuz mugatu ohi dira, baina izenburuetan, esate baterako, hitz guztiak maiuskulaz idatzita agertzea oso ohikoa da, eta, gainera, izen berezi bat baino gehiago jarraian agertuz gero, bakoitza non hasi eta non amaitzen den erabaki behar da. Adibidez, *Xabierrek Mikel etxera lagundu zuen* esaldian, *Xabier* izena ergatiboan dagoenez, izena bertan amaitzen dela jakin daiteke, baina *Xabier Mikelekin joan zen etxera* esaldian zaila da izen bat (*Xabier Mikel*) ala bi (*Xabier* eta *Mikel*) bereizi behar diren jakitea.

Behin entitateak mugatu direnean, sailkapenari ekin behar zaio eta, horretarako entitatearen informazio morfologikoaz gain, testuinguruko informazio morfosintaktiko, semantiko eta pragmatikoa erabiltzea komeni da.

Aztertu diren hurbilpen gehienetan data eta zenbakien tratamendua modu errazean egin daitekeela adierazten da, orokorrean egunkarietako testuetan nahiko egitura finkoan agertzen direlako. Hortaz, adierazpenen egituren patroiak edo gramatika-erregelak definitu eta horien bitartez identifika daitezke.

Izen berezien adierazpenak, aldiz, ez dira hain errazak tratatzen. Batetik, izen bereziek klase irekia osatzen dutelako eta, bestetik, ingelesa ez den beste hizkuntza gehienetarako baliabideak oso urriak direlako.

---

23 Aipatutako (Dunning 1993) lanean hau ere komentatzen da: *"...and z-scores substantially overestimate the significance of rare events"*.

Sistema gehienek McDonald-ek (1996) esandakoari jarraituz, izen bereziak sailkatzeko barne- eta kanpo-ebidentzietan oinarritzen dira. Barne-ebidentziak hitz edo hitz-sekuentziaren ezaugarriak dira, esate baterako, osagai bat edo gehiago izen berezia izatea. Kanpo-ebidentziak testuinguruari buruzko informazioa da; izenaren alboan *andere*, *ibai*, *kooperatiba* edo antzeko hitz-gakoak (*trigger words*) agertzeak izen hori pertsona-izen, leku-izen edota erakunde diren erabakitzen lagunduko dute hurrenez hurren.

Informazio hori erabiliz testuan agertzen diren izenak identifikatu eta sailkatzen dira, baina sistema batetik bestera aurretik burutu beharreko prozesamendua asko aldatzen da. Sistema sinpleenak tokenizazio hutsetik abiatzen diren bitartean, sistema konplexuenetan analisi morfologiko, desanbiguazio, azaleko sintaxia eta semantika ere erabiltzen dira.

Kontuan hartu beharreko beste ezaugarri garrantzitsu bat sailkapenerako erabiltzen den metodoa da. Hiru multzotan banatu ditugu aztertutako sistemak:

1. Metodo estatistikoak erabiltzen dituztenak:
   - Markov-en eredu ezkutuak edo *Hidden Markov Models* (HMM): entitate-mota bakoitzeko egoera bat eta gainerakoetarako NOT-A-NAME egoera erabiltzen dira (Bikel *et al*. 1997-1999).
   - Entropia maximoa (*Maximum Entropy*): desanbiguazioan emaitza oso onekin erabilitako teknika entitateak sailkatzeko ere erabili da. Kasu honetan entitate hasieran, erdian, amaieran edota hitz bakarrekoa den erabakiko da (Borthwick *et al*. 1998-ab). Hizkuntz eredua aplikatzean, hitzen sekuentziaren entropia maximizatzea du helburu. Desanbiguazioari buruzko kapituluan azaltzen da hizkuntza eredua eraikitzeko ematzen diren pausoak (ikus VI.1.2).
   - Erabaki-zuhaitzak (*Decision Trees*): zuhaitza definitzerakoan entitateen hasiera eta bukaerari dagozkion ezaugarriak modelizatzen da hitzaren eta bere ingurukoen informazioa baliatuta (Cowie 1995) eta (Sekine *et al*. 1998).
   - *Boosting*: hainbat ezaugarri eta bere testuinguruak erauzita, erabaki-zerrendak eratzen dira (*decision lists*) eta, ondoren, *boosting* teknika erabilita, ikasketa-prozesua burutzen da (Collins eta Singer 1999). Ikasketa gainbegiratua burutzeko arruntena den EM (*Expectation Maximization*) algoritmoa erabiltzen da batetik, eta *DL-CoTrain* algoritmoa bestetik. Emaitza onenak azken honekin lortzen dira.
   - *Voted Perceptron*: hizkuntz ereduaren probabilitateak sare neuronalen bitartez hurbiltzen dira (Collins eta Duffy 2002; Collins 2002-b).
2. Metodo linguistiko hutsak erabiltzen dituztenak:
   - Egitura sintaktikoan oinarrituta: orokorrean esaten denaren aurka, (Black *et al*. 1998) lanean entitateen erauzketa/sailkapena zein informazio erauzketaren gainerako atazetan

testuaren analisi sakona egin behar dela diote, eta, horrela, oso tresna sendoa eta emaitza onekoa lortzen dute.

- WordNet-eko informazioan oinarrituta: (Magnini *et al*. 2002) lanean WordNet baliabideak eskaintzen duen informazioa erabiltzea proposatzen dute, sailkapena egiteko entitatea edota bere inguruko hitz-gakoak hierarkia semantikoan kokatuz.
- WordNet eta erregelak erabilita: (Arévalo 2002) lanean MICE tresna deskribatzen da. Testuaren analisi morfologikoa eta azaleko analisi sintaktikoa (izen sintagmen identifikazioa) burutu ondoren, entitateek jarraitzen dituzten patroiak deskribatzeko gramatika erabiltzen dute. Patroi horietan hitz-gakoak (*trigger words*) agertzen dira eta hitz horiek WordNeten dituzten etiketa semantikoak entitateen sailkapenerako laguntza gisa erabiliko dira.

3. Erregelak eta estatistika erabiltzen dituztenak:
   - Erregelak eta entropia maximoa: (Mikheev *et al*. 1998-1999) eta (Uchimoto *et al*. 2000) lanetan testuinguru-erregelak erabiltzen dituzten hautagai-zerrendak osatzeko eta horiek egoki sailkatzen dituzte entropia maximoko hizkuntz eredu probabilistikoan oinarrituta.
   - Erregelak eta konfiantza-neurria (*plausibility*): (Cucchiarelli *et al*. 1998-ab) lanetan ere testuaren analisi sintaktikoa egiten da eta, ondoren, entitateak identifikatzeko erregelak aplikatu eta konfiantza-neurriaren bitartez erabakitzen da zeintzuk diren ondo sailkatutako entitateak.

Euskararen tratamenduari begira interesgarrienak direlakoan, atal honetan sakonean aztertuko diren hurbilpenak metodo linguistiko hutsak erabiltzen dituztenak eta metodo konbinatuak erabiltzen dituztenak izango dira. Mikheev-en taldeak proposatutakoa, aurreprozesu minimoa eginez, emaitzarik onenetakoak lortzen dituelako aukeratu da sakonean aztertzeko, eta Magnini-ren taldearena, prozesaketa sakona eskatzen duen arren, etorkizunari begira euskarazko izen bereziak sailkatzeko kontuan hartu beharreko hurbilpena delako. Gainerakoen inguruko datu batzuk aztertzen dira, interesgarriak izan daitezkeelakoan.

Metodo estatistiko hutsak erabiltzen dituztenetatik (Bikel *et al*. 1999) aukeratu dugu atal honetan aurkezteko, bere aplikazioa ez delako testu idatzi arruntetara mugatzen. Lan horretan Markov-en eredu ezkutuen egokitzapena aurkezten da. Ereduaren klaseak MUCen definitzen diren entitate-motak dira, eta entitateak ez diren unitateetarako NOT-A-NAME klasea gehitzen du. Gainera, esaldi hasiera eta esaldi amaiera ere erabiliko ditu. Emaitza politak lortzen ditu (%96 zuzentasunean eta %93 zehaztasunean) MUC-6 konferentziako testuetan oinarriturik[24]. Dena dela, ikasketarako corpusaren tamaina oso handia da, 650.000 hitz, eta tamaina 60.000

[24] Konferentziaren edizio horretan emaitzarik onenak, eskuzko anotatzaileen parekoak, %96ko zuzentasuna eta %97ko zehaztasuna izan ziren.

hitzetara laburtuz gero, emaitzak modu esanguratsuan okertzen dira. Hala ere, emaitza nahiko onak lortzen dituzte OCR programatik datozen testuekin, baita hizketan oinarritutako sarrera tratatzen dutenean ere.

Metodo linguistikoetan oinarritutako bi hurbilpen aipatuko dira jarraian. Batetik, (Black *et al*. 1998) taldeak garatutakoa, FACILE informazio erauzketarako sisteman (Ciravegna *et al*. 1999) erabiltzen dena, eta, bestetik, WordNeteko informazioa baliatzen duen sistema (Magnini *et al*. 2002).

FACILE informazio erauzketarako sistemak (Ciravegna *et al*. 1999) testuaren analisi sakona egiten du, teknika estatistikoetan oinarritu gabe, sistema sendoa osatu dutelarik. Lehen urratsa izendun entitateak ezagutzea da. Ezagutzailea erregeletan oinarritzen da eta ez du inongo ikasketa teknikarik erabiltzen. Sistema lau hizkuntza hauetarako dago inplementatuta: ingelesa, italiera, gaztelera eta alemanera.

Identifikazioa hasi aurretik, prozesaketa sakona burutzen da: tokenizazioa, analisi morfologikoa, desanbiguazioa eta ezaugarri semantikoen esleipena. Erregelak informazio horretan guztian oinarritzen dira izendun entitateak identifikatzeko, informazioa ezaugarri-bektoreetan antolatuz. Erregelak patroi-ezkontze (*pattern matching*) moduko aplikazioa egiteko eragiketak dituzte, baina irakurgarriagoak dira.

Abiadura ezinbesteko ezaugarria behar zuenez, Prolog ez erabiltzea erabaki eta inplementazio propioa garatu zuten. MUCen aurkeztutako sistemaren emaitzak ez dira oso onak, bereziki erakundeei dagokienean, baina garapen-fasean egonik, hobekuntzak burutzeko arazo handirik ez dutela aipatzen dute. Sistema sendoa da, baina erregelak eskuz idatzi behar izateak beti dakar sistema mantentzeko eta domeinuz aldatzeko arazo anitz.

Testuaren analisian oinarritutako beste hurbilpena, Magniniren taldeak berriki aurkeztutakoa da (Magnini *et al*. 2002). Lan honen motibazioetako bat *gazetteer*-ak lortu, eguneratu eta neurri egokian mantentzearen zailtasuna ekiditea da, hizkuntz aldaketa edota domeinu aldaketa egin nahi denean bereziki. Hori gainditzeko, kanpo-ebidentziarako erabiltzen diren gakoetan (*trigger words*) zentratzea proposatzen dute, horretarako WordNetek eskaintzen duen informazioa baliatuz.

Barne-ebidentzia gisa hitzen instantziez gain (*New York*), hitzen klaseen informazioa ere erabil daiteke (*time unit#1*). Kanpo-ebidentziarako interesgarriak diren klaseetako gakoak erauzi dituzte. Adibidez, *person#1 synset*-etik 6775 gako erabiltzen dituzte (*astronomer, physicistm Norwegian, professor*...). Hitz-klase eta instantzia hauek erregelak diseinatzeko erabiliko dira.

Identifikazioa hasi aurretik eman beharreko urratsak honakoak dira: tokenizazioa, desanbiguazioa (etiketatzaile estatistiko bat erabiliaz) eta WordNetetik lortutako hitz anitzeko unitateen tratamendua (5.000 inguru erauzi dituzte automatikoki eta patroi-ezkontzearen bitartez tratatzen dira). Identifikazioa burutzeko bi fase erabiltzen dituzte:

- Oinarrizko erregelen aplikazioa: 200 erregela inguru aplikatzen dira testuko izendun entitateak identifikatu eta sailkatzeko.
- Konposaketa-erregelen aplikazioa: goi-mailako erregelak aplikatuko dira aurreko urratsean sortutako etiketen anbiguotasuna ebazteko —entitate bat bestearen barruan agertzen denean, esate baterako— eta korreferentziaren ebazpena burutzeko.

Emaitzak nahiko onak dira, kategoria batean izan ezik (neurriak edo *measures*). Emaitzak jarraian aurkeztuko den Mikheev-enekin konparagarriak direla diote, *gazetteer*-ik ez baitute erabiltzen. Hala ere, hitzen instantzien artean, esate baterako, erregeletan erabiltzen diren 2.173 leku-izen daude eta ez du zehazten horietako zenbat erabili diren emaitzak lortzeko. Mikheev-ek. aldiz, 200 leku-izen erabilita, lan honetan emandako emaitzen parekoak lortzen ditu.

Hala ere, WordNeten (edo *EuroWordNet*-en behintzat) hainbat hizkuntzetako sarrerak loturik daudenez, hemen aurkeztutako hurbilpenak etorkizun hurbilean ekarpen handiak eman ditzakeela uste dugu. Izatez, (Arévalo 2002) lanean aurkezten den bezala, dagoeneko gaztelera eta katalanerako WordNeten informazioa baliatzearen bidea erabiltzen ari dira HERMES proiektuaren barruan, proiektuan landuko diren beste hizkuntzetara hedatzeko asmotan (euskara barne).

Azkenik, metodo konbinatuei dagokienean, bi hurbilpen aurkeztuko dira jarraian: batetik, *Language Technology Group* (LTG) taldeak garatutakoa (Mikheev *et al*. 1998), eta, bestetik, Erromako Unibertsitatean italiarrerako aurkeztutakoa (Cucchiarelli *et al*. 1998-ab)

Esan bezala, aurreprozesu minimoa erabili arren, bibliografian erreferentzia garrantzitsua den hurbilpena MUC-7 konferentzian aurkeztutako LTG taldeak garatutakoa da (Mikheev *et al*. 1998). Tokenizazio hutsetik abiatuta konferentzian aurkeztutako sistemen artean emaitzarik onenak lortu ditu. Sistemak bost urratsetan tratatzen ditu izendun entitateak:

1. Erregela ziurrak (*sure-fire rules*): maiuskulaz idatzitako hitz edo hitz-sekuentzia bat aurkitzean, testuinguruak dudarik gabe sailkatzeko aukera ematen badu, izen berezia sailkatzen du. Esate baterako, titulu bat (*Mr.*, *Dr.*, *Sen.*,...), kargu bat (*president*, *director*...) edota erakunde baten izendatzailerik (*Ltd.*, *Inc.*,...) agertzen bada, izen mota erabakitzeko moduan da sistema, beti ere, erregela batzuei jarraituz. Hala ere, pauso honetan hartutako erabakiak behin-behinekoak dira, eta hurrengo urratsetan alda daitezke.

2. Ezkontza partziala 1 (*partial match 1*): urrats honetan orain arte identifikatutako izenen ezkontza partzialak egiten ditu eredu probabilistiko baten bitartez. Horretarako bi tresna erabiltzen ditu. Lehenengoan, aurreko pausoan aurkitutako entitateen osagaien konbinaketa posible guztiak sortzen ditu, beti ere osagaien ordena errespetatuz. Adibide gisa hauxe ematen du: *Lockheed Martin Production* erakundearen kasuan *Lockheed Martin*, *Lockheed Production*, *Martin Production*, *Lockheed* eta *Martin* sortzen ditu erakunde posible bezala. Dena dela, horietako batzuk, *Martin* kasu, beste entitate mota batekoak izan daitezke. Bigarren tresnak entropia maximoan oinarrituriko eredu bat erabiltzen du ezkontza partzialak sailkatu behar diren ala ez erabakitzeko. Horretarako, entitateak esaldian duen posizioa, minuskulaz idatzita agertzen den ala ez eta testuinguruaren bestelako ezaugarriak erabiltzen ditu.
3. Erregela erlaxatzea (*rule relaxation*): lehenengo urratseko erregelen antzekoak aplikatzen dira, baina oraingoan testuinguruari buruzko baldintzak erlaxatu egiten dituzte eta dagoeneko markatuta dauden entitateen informazioa ere baliatzen du. Urrats honetan zerrendetan agertzen diren erakunde eta leku-izenak testuinguruari erreparatu gabe markatzen dira.
4. Ezkontza partziala 2: dagoeneko sistemaren baliabide guztiak erabili dira (izenen gramatikak, izenen zerrendak,...). Bigarren urratsean egiten den gauza bera errepikatzen da eta eredu probabilistikoari jarraituz, gera daitezkeen ezkontza partzialak sailkatzen saiatzen da.
5. Tituluen esleipena (*title assignment*): egunkarietan askotan tituluak maiuskulaz idatzita datozenez, ez dute informazio handiegirik ematen bertan ager daitezkeen izenen inguruan. Horregatik, azken urratsean aurretik identifikatutako entitateak tituluan agertzen badira, markatu egiten dira.

Sistemaren emaitzak V.1 taulan agertzen dira. Kontuan izan behar den datua konferentziarako eskuz etiketatu zuten hizkuntzalarien emaitzak dira, %96-98ko zuzentasuna (*recall*) eta %98 zehaztasunean (*precision*), alegia. Beraz, ez da espero MUCeko sistemek hori baino hobeto egitea.

Emaitza hauek lortzeko, esan bezala, izenen zerrendak erabili dituzte (*gazetteer*), kasu honetan 4.900 herri-izen, 30.000 erakunde eta 10.000 pertsona-izen (abizenik ez). Baina, esan bezala, izen berezien klasea irekia da, hortaz, etengabe eguneratzen joan beharko litzateke *gazetteer* horiek.

Entitate-mota	Recall	Precision
Erakundeak	%91	%95
Pertsona-izenak	%95	%97
Leku-izenak	%95	%93
Batez beste	%92	%95

V.1 taula.- LTG sistemaren emaitzak.

Zerrenda hauen garrantzia aztertzeko, Mikheev-en taldeak hainbat saiakuntza egin dituzte (Mikheev *et al*. 1999). Batere zerrendarik gabe egindako saiakuntzan erakunde eta pertsona-izenen identifikazioa nahiko ondo burutzen du sistemak (zuzentasunean %88 eta zehaztasunean %90 inguru) baina leku-izenetan nahiko kaskar (%46 eta %59). Hori konpontzearren, *www.yahoo.com/Regional* web-gunetik 200 inguru herrialde eta kontinenteen izen, eta eguzki sistemako 8 planeten izenak atera eta saiakuntza errepikatu dute. Emaitzak dezente hobetzen dira, leku-izenen kasuan %85eko zuzentasuna eta %90eko zehaztasuna lortuz. Honela, erakunde eta pertsona-izenen emaitzekin parekatzen da[25].

Emaitza horiek MUC-7 konferentzian aurkeztutako zenbait sistema sofistikatuagoenak baino hobeak dira, nahiz eta zerrenda oso laburrak erabili. Hala ere, beste hizkuntzetan horrelakorik egin daitekeen frogatu beharko litzateke, ingeleserako baliabide anitz baitago web-ean.

(Cucchiarelli *et al*. 1998-ab) lanetan, ingeleserako ez diren sistemak diseinatzerakoan *gazetteer*-en faltak eragiten dituen oztopoak ikusirik, zerrenda horiek automatikoki eta era ez-gainbegiratuan osatzen joateko modua ematen du. Esan bezala, sistema hau erregeletan oinarritzen da, gainerakoak bezala, barne- eta kanpo-ebidentzietan oinarrituz. Baina italiera tratatzen duenez, eta ingelesak baino morfologia aberatsagoa duenez, tokenizaziotik abiatu ordez, analisi morfologiko, desanbiguazio eta (ez hain) azaleko sintaxia (Basili *et al*. 1994) aplikatu ondoren bilatzen ditu izendun entitateak. Aurkitutako izenak sailkatzeko izen berezi ezagunen sintagmetan oinarrituz, ezezagunak sailkatu eta *gazetteer* moduko izen berezien sintagmen datu-basea aberasten dute gainbegiratu gabe.

[25] Beste saiakuntza batzuk ere burutu dituzte, *gazetteer* luzeagoak erabilita. Emaitzak (Mikheev *et al.* 1999) lanean aurki daitezke.

V.3 Euskararen tratamendua

Aurreko ataletan gure ustez tratamendua merezi duten hitz anitzeko unitateak zeintzuk diren eta horretarako proposatu diren hurbilpenak aurkeztu dira. Atal honetan, euskararen kasuan hartutako erabakiak laburbildu eta lan hori burutzeko beharrezkoak diren tresnen aurkezpena egiten da.

Hitz anitzeko unitate ireki eta itxien arteko bereizketa egin dugu, batez ere, ikuspuntu informatikotik prozesamendu-mota desberdina eskatzen dutelako, baina alde linguistikotik ere azterketa-maila desberdina jaso izan dutela ahaztu gabe. Unitate itxien kasuan, orokorrean zerrenda edo datu-baseetan bil daitezke, prozesatzean landutako informazioaren bilaketa egiten delarik. Gure kasuan, datu-basean landu diren hitz anitzeko unitate itxiak (lexikalak deitu ditugunak) honako hauek dira: lokuzioak, opako zein figuratiboak, kolokazio murriztuak eta kontzeptu jakinak adierazten dituzten kolokazio ireki eta hitz-elkarketak. Unitate irekietan kasu ezberdinak azter daitezke:

- Kolokazio irekiak eta hitz-elkarketak: hiztegi-sarrerarik ez duten hitz anitzeko hauek, teknika estatistikoen bitartez identifika daitezke. Hala ere, Xtract edo LocalMax moduko tresna batek emaitza egokiak lortzeko corpus itzela behar du, neurri estatistikoak esanguratsuak izango badira, behintzat. Horregatik, hauen tratamendua lan honetatik kanpo geratu da eta etorkizunerako ikerlerro gisa planteatu da.
- Data eta zenbakien adierazpenak: esan bezala, oso egitura finkoa duten adierazpenak dira, hiztegi oso txikia erabiltzen dutenak. Bilaketa egiteko inongo arazorik planteatzen ez dutenez, eta tratamendu morfologikoa berehalakoa denez, hauek tratatuko dira ondorengo ataletan azalduko den moduan.
- Izen bereziak: multzo irekia osatzen dute eta bere identifikaziorako tresna egokia garatzeak eskatzen duen esfortzuagatik (diseinu, inplementazio eta probez gain, metodo gainbegiraturen bat aplikatzeko beharrezkoa den eskuzko desanbiguazioa), lan honen esparrutik kanpo geratu da. Dena dela, dagoeneko taldean erregela eta metodo estatistikoen bidezko hurbilpena jarraitzen ari dira, izen berezien identifikazio eta sailkapenerako tresna garatzean, eta epe laburrean lehen prototipoa martxan izatea espero da.

Ondoren, V.3.1 atalean, hitz anitzeko unitate itxien tratamendua burutzeko diseinatu eta inplementatu den HABIL tresna (*H*itz *A*nitzeko unitateen *BIL*aketarako tresna) aurkezten da. Jarraian, V.3.2 atalean, izendun entitateen identifikaziorako garatu diren eta garatzen ari diren tresnen aurkezpena egiten da. Eta, azkenik, tratamenduaren ebaluazioa egiten da V.3.3 atalean.

V.3.1 HABIL

Hitz anitzeko unitate lexikalen (HAUL) tratamendua bideratzeko aztertutako hurbilpenak arrazoi desberdinengatik baztertu behar izan ditugu. Kasu gehienetan HAUL jarraiak besterik ez dira tratatzen, edo ez-jarraiak tratatzen direnean murriztapen handiak ezartzen zaizkie osagaien artean ager daitezkeen hitzei, bai kopuruari bai hitzen kategoriari dagokionean. Lan honetan murriztapenak erlaxatu dira, gure ustetan HAULaren testuingurua zabala delako.

Bestalde, HAULen osagaien ordena aldaketari dagokionean, hizkuntza gehienetan horrelako ezaugarriak dituzten HAULak izan arren, tratamenduaren konplexutasuna dela medio, gehienetan sintaxian tratatzeko uzten dira. Erabiliko den tresnak, berriz, ordena aldaketaren fenomenoaren hedapena ikusirik, ezaugarri hau duten HAULak ere tratatu behar ditu.

Osagaiei dagokionean, hurbilpen askotan forma finkoak bakarrik hartzen dira kontuan. Izan ere, hitz anitzeko espresioen tratamendua analisi morfologikorik gabe egin ohi da, eta forma finkoak besterik ez badira kontuan hartzen, tratamendua benetan errazten da. Gure kasuan, aldiz, forma finkoak zein flexioa onartzen dutenak tratatzeko gai den tresna behar da, nahiz eta horrek tratamendu konplexua eskatu.

Laburbilduz, hitz anitzeko unitate itxien tratamenduen azterketatik ondorio hauek atera daitezke: gaur egun proposatu diren sistemetan ez dira euskarak eta beste hizkuntza batzuek dituzten ezaugarri guztiak aurreikusten edota ezaugarri guztiak aurreikusten direnean tresnak ez dira publikoak. Hori dela eta, tratamendu propioa diseinatu eta inplementatzea erabaki da.

Tratatu beharreko HAULen ezaugarrietatik abiatuta HABIL inplementatu da. Bere ezaugarri garrantzitsuenak honakoak dira:

- HAUL ez-jarraiak tratatzen dira.
- Ordena alda dezaketen osagaiak dituzten HAULak tratatzen dira.
- Forma finkoak zein flexioa onartzen dituztenak tratatzen dira.
- Informazio morfologikoa ere lortzen da.

Tresna hau diseinatzerakoan argi eta garbi desberdintzen diren bi urrats behar zirela erabaki zen: unitateen bilaketa batetik, eta prozesamendu morfologikoa bestetik. Bilaketak berak du zailtasunik handiena, bereziki datu-basearen osagaietan, V.3.1.1 atalean, bilaketa-prozesuaren diseinuari eskainitako atalean ikusiko den bezala. Data eta zenbakien kasuan, morfologian erabilitako tresnaz egoera finituko definizio sinpleagoa inplementatu da bi arrazoi nagusi hauengatik: lehenengoa, adierazpenaren osagai guztiak jarraian agertzen dira, eta oso egitura finko eta hiztegi zehatza erabiltzen dute; bigarrena, zenbaki eta data guztien adierazpenak ezin dira datu-basean landu, mugatzen ez badira, behinik behin. V.3.1.2 atalean deskribatzen da tresna honen diseinua. Izen berezien tratamenduari ere atal bat

eskaintzen zaie, V.3.1.3 atala, lan honen esparrutik kanpo geratu arren dagoeneko lantzen ari den gaia delako.

Bestalde, testuetan HAULak topatzen direnean zer egin ere erabaki da: HAUL ziurren kasuan beste interpretazioak baztertuko dira, eta anbiguoen kasuan interpretazio berria erantsiko da. Hala ere, eta inplementazioa egina dagoen arren, ez dira lan honen ebaluazioan HAUL anbiguoak erabiliko, ez baitago argi prozesuaren zein urratsetan trata daitezkeen —printzipioz desanbiguazioaren ondoren baina sintaxia baino lehen egin beharko litzateke logikari jarraituz gero— eta datuen lanketa finagoa eskatzen dutelako. Etorkizun hurbilean, integrazio honi heltzea espero dugu, ebaluazio sakona egin eta behin betiko EUSLEMen integratzeko HAUL anbiguoen tratamendua.

Prozesaketa morfologikoari dagokionez, kasu askotan kategoria eta azpikategoria ematean datza, datu-basean landuta dagoen informazioa HAULari esleitzean, alegia. Beste kasu batzuetan, hitz anitzeko unitateak flexioa onartzen duenean zehazki, informazio hori ematen dion osagaiak dituen analisi morfologikoak aztertu eta landutako murriztapenak betetzen dituzten analisiak sortu beharko dira datu-baseko kategoria eta azpikategoria kontuan hartuta, patroi-ezkontza eginez. Data eta zenbakiek flexioa onartzen dutenez, bilaketarako tresnak identifikatutako adierazpenei kategoria eta azpikategoriaz gain, flexioari buruzko informazioa duten analisiak esleitu beharko zaizkie, azken osagaiaren analisietan oinarrituta. Prozesu hau V.2 irudian beltzez ematen den algoritmoaren atalean burutzen da.

V.3.1.1 HAULen bilaketa

Testuetako HAULen detekzioak gure helburuetarako balio izango badu, oinarrizko lana bilaketa-estrategia eraginkorra diseinatzea izango da. Bilaketa egiteko hartu behar izan diren erabakiak ondorengo lerroetan aztertuko dira.

Lehenik eta behin, bilaketaren esparrua definitu behar da, HAULaren osagaiak bilatzeko mugak ezarriz. Orokorrean, HAUL baten osagaiak esaldi baten barruan agertuko direla esango da. Baina testu erreal batean esaldia noraino iristen den erabaki behar da. Zabal jokatu nahi izanez gero, esaldia *puntutik puntura doan hitz-sekuentzia* izango da. Hala ere, definizio orokorregia da hau. Badira beste puntuazio-ikur batzuk esaldiaren muga adieraz dezaketenak eta horiek ere kontuan hartuz gero, bilaketa-esparru laburragoa erabiltzea ahalbidetzen da.

Honela, HAULak bilatzeko esparruaren muga hauexek erabiliko dira HABILen: esaldi bukaerako puntua (**.**), galde-ikurra (**?**), harridura-ikurra (**!**), puntu eta koma (**;**) eta bi puntuak (**:**). Erabaki honek arrazoi argia du, ez baita posible ikur hauen ondoren HAULaren osagairen bat egotea eta puntua bakarrik erabiliko bagenu, beste zeinuak erdian direla osagaiak identifikatu eta HAUL okerrak markatzeko arriskua egongo litzateke.

Bilaketan jarraitzen diren pausoak V.2 irudiko algoritmoan zehaztu dira. Bertan ikus daitekeenez, HAULak detektatu ahal izateko, oraintxe aipatutako mugen arteko hitz-sekuentziak irakurtzen dira, dagozkien analisi eta guzti, eta hitzak edota bere lemak HAULen baten osagai izan daitezkeen aztertzen da. Orduan, HAULen osagai guztiak aurkitu ez badira ez da inongo murriztapenik egiaztatuko. Erabaki hau eraginkortasuna bermatzeko hartu da. Izan ere, alferrikako lana da osagai baten flexio-murriztapenak konprobatzea gainerako elementuak ez badira esaldian topatu.

Behin osagai guztiak identifikatu diren ziurtatzean, HAULaren definizioan ezarritako murriztapen guztiak betetzen direnentz begiratuko da. Lehendabizi ordenari dagozkion murriztapenak egiaztatuko dira, eta hauek betetzen badira, flexioari dagozkionak begiratuko dira. Informazio morfologikoa egiaztatzeak lan handiagoa emango du gehienetan eta ordenari buruzko murriztapenak, aldiz, eragiketa azkar batez alderatuko dira. Horregatik eragiketarik azkarrena lehenik egingo da, ordena mantentzen ez bada eragiketa garestienak ekiditeko.

Murriztapen guztiak betetzen badira, HAUL horren osagaien agerpen bakarra dagoen begiratuko da. Izan ere, maiztasun handiko osagaien kasuan batez ere, agerpen bat baino gehiago topa daiteke, gainerako osagaien agerpen bakarra dagoen bitartean. Kasu hauetan, hautagaien artean bakarra onartuko da. Bestalde, antzeko arazoa sortzen da HAULen artean ere. Kasu honetan, HAUL bat baino gehiagoren osagai izan daitezkeen hitzak aurki ditzakegu bilaketa-esparruan.

Interferentzia hauek sortarazten dituzten HAUL garrantzitsuenak aditz konposatuak dira, multzo honi eskainitako atalean (V.1.3) azaldu den bezala. Horrelakoetan ere aukeraren bat egin beharko da.

Beraz, tresna automatikoa sendoa izan dadin irizpide batzuk erabaki dira, baina erabat orokorrak, zuzenak edota nahikoa ez izatea gerta liteke. Etorkizunean corpus azterketaren bidez irizpide hauek neurtu, findu eta lantzea izango da helburuetako bat. Oraingoz, honako irizpideak definitu dira:

- HAULaren osagai baten instantzia guztien artean hurbilen dagoena aukeratu. Irizpide honek logikoa dirudi. Elementuak, hurbilago izanik, erlazionaturik egoteko probabilitate handiagoa izango dute bata bestetik aldendurik egonda baino. Zenbait kasutan irizpide honek huts egin dezake, baina proportzio oso txikia izango denez irizpide orokortzat hartu da.
- HAULen artean ziurrak aukeratu. Interferentziak ziur eta ez-ziurren artean direnean, erraz ikus daiteke ziurra lehenetsi behar dela, kasu horretan osagaiek interpretazio bakarra dutelako.

- HAUL ziurren artean luzeena aukeratu. Honela, esaldi batean *hala eta guztiz ere* agertzen denean bi HAUL izango dira: *hala eta guztiz* eta *hala eta guztiz ere*. Lehenengoa bigarrenaren zatia denez eta biak ziurrak izanik, *hala eta guztiz* aukera baztertuko da. Kasu hau kenduta, ez da bestelako adibiderik aurkitu, baina baliteke antzeko kasu gehiago topatzea etorkizunean eta horregatik irizpide orokor gisa sartu nahi izan da.
- HAUL ziurrik ez badago, osagaien agerpenetan HAUL guztien analisia gehituko da. Honela anbiguotasuna handitzen den arren, ziurtasun-ezak beste aukerarik ez du ematen.

Lehen esan bezala, irizpide hauek ez dira behin betikoak, emaitzen arabera aldatu ahal izango dira, baina orain arteko probetan oso emaitza ona eman dute.

```
irakurri(esaldia);
/*** esaldian agertutako hitzak ea osagaien tauletan dauden konprobatu */
/*** batetik forma osoa eta bestetik bere lema konprobatu beharko dira */
for (hitza in esaldia) do
  if (osagaia_da(hitza)) then
       markatu_osagaia();
  endif;
done;
/*** osagai guztiak markatuak dituzten HAULak aztertuko dira */
for (haul in HAUL_guztiak) do
  if (markatua(haul) && murriztapenak_bete(haul)) do
       /*** murriztapenak betetzen badira, gordeko da HAULa */
       erantsi(zerrenda, haul);
  endif;
done;
/*** zerrendan dauden HAULak bakarrik tratatuko dira */
/*** HAULen interferentziak ebatzi */
ebatzi_interferentziak(zerrenda);
/*** data eta zenbakiak aldez aurretik identifikatu dira */
/*** zerrenda2 aldagaian gordeta egongo dira hasieratik */
erantsi_HAUL_informazioa(esaldia, zerrenda, zerrenda2);
inprimatu(esaldia);
```

V.2 irudia.- HAULen bilaketaren algoritmo nagusia.

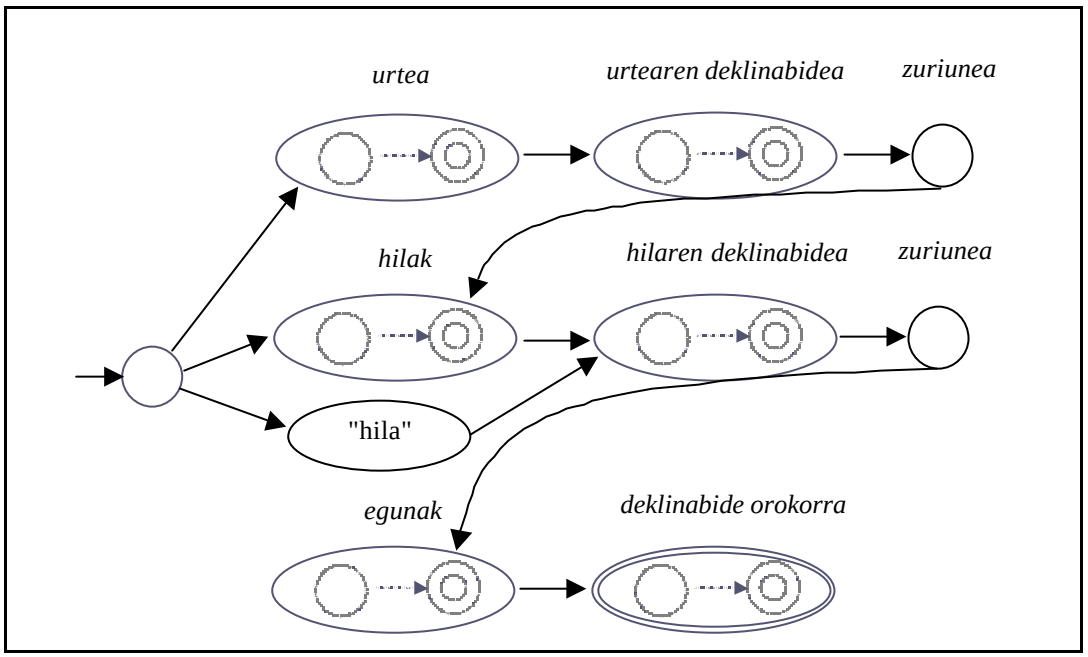
V.3.2 Izendun entitateen tratamendua

Izendun entitateen tratamendua diseinatzerakoan, beste hurbilpen batzuen ildotik, bi ataza desberdin bereizi ditugu: batetik, identifikatzen eta tratatzen errazagoak diren data eta zenbakien tratamendua burutu da, eta, bestetik, izen berezien tratamendua prestatzen ari gara. Bigarren hau lan honen esparrutik at geratu den arren, garapen-fasea amaitzear dagoenez, taldean egindako lanari ere atal bat eskaini zaio.

V.3.2.1 Data eta zenbakien bilaketa

HAULen identifikaziorako kasu askotan analisi morfologikoa behar izaten da. Data eta zenbakien kasuan, euren egitura eta hiztegia edukita, testu tokenizatuaren gainean bila daitezke adierazpenak. Hori dela eta, entitate hauen bilaketarako, lexikoi sistema murriztua eraiki da kasuan kasuko egiturak gauzatzeko lexikoien jarraitze-klaseak definituz.

Lexikoi bakoitzean zenbaki edo datei dagozkien sarrerak gehitu eta adierazpenaren osagai bakoitzak onartuko dituen flexioak kontrolatzeko, EDBLn dituzten jarraitze-klaseak baliatu ditugu. Horrela, daten lexikoien antolakuntza V.3 irudikoa da.



V.3 irudia.- Daten ezagutzailearen lexikoien antolakuntza.

Daten ezagutzaileak modu honetako adierazpenak identifikatu ditu:

- `2001eko urtarrilaren 20an`
- `2000. urteko otsailaren 15a`
- `hilak 30`
- `otsailak 23`
- `hilaren 15erako`

Baina gainsortzailea denez, ez ditu urruneko mendekotasunak egiaztatzen eta honakoak ere identifika ditzake:

- `otsailak 31n`
- `hilaren 30`

Dena dela, daten idazkera ez-zuzenak eta hilen aldaerak ere ezagutu eta modu egokian tratatzea komeni da, gainerako aldaera eta gaitasun-desbideratzeen moduan, etiketatzaile

sendoa garatuko bada. Bestela, gainsorkuntza ekidin nahi izanez gero, murriztapen-erregela batzuk idaztea baino ez litzateke egin beharko.

Zenbakien ezagutzailea antzera antolatu da, baina konplexuagoa da. Lehenengo, zenbakiak hasteko aukera gehiago daude, bilioiko, milioiko, milako, ehuneko, hamarreko eta unitateekin has daiteke zenbakia, eta ezin ahaz daitezke *ehuneko hogei* moduko adierazpenak. Kontuan hartu den beste ezaugarria *bat* eta *bi* zenbakien erabilera izan da. *Bat*en kasuan, gainerakoak ez bezala, atzetik doa, hortaz, *bat milioi* debekatu behar da. *Bi*ren kasuan, aukera biak kontuan hartu behar dira, baina *bi bilioi bi* debekatuz. Gainera, zenbaki baten bi osagaien artean *eta* ager daiteke —*ehun eta bost*— eta zenbaki luzeetan komak ere tartekatu ohi dira —*sei milioi, zortziehun mila, bostehun eta berrogeita sei*—. Azkenik, lexikoietan atzera ere egiteko aukera aurreikusi da; azken adibidean, esaterako, *sei* zenbakia ezagutu ondoren *milioi*en lexikoira joan behar da, ondoren milakoetatik pasa gabe, ehunekoetara pasako da *zortziehun* ezagutzeko, eta atzera *mila*kora, hortik ehunekora *bostehun* ezagutzeko, hamarrekoetara *berrogeita* ezagutzera eta unitateetara *sei* ezagutzera adierazpen osoa tratatu arte.

Zenbakien idazkera zaharra ere onartu nahi izan da, *hogeitabat* edo *berrogei ta zazpi* modukoak onartuz[26].

Esan bezala, identifikatutako data eta zenbakiak HABILi ematen zaizkio tratamendu morfologikoa egin dezan. Kasu bietan azken osagaiak emango baitio unitate osoari flexioari buruzko informazioa.

Ikasketarako zein egiaztapenerako testuetan ez dira data eta zenbaki asko agertzen, zatirik handiena EEBS corpus orekatuaren lagina denez, testu literario anitz dituelako. Hala ere, ondoren ematen diren adibideak aurkeztu nahi genituzke. Adibidez, zenbakien ezagutzaileak modu honetako adierazpenak identifikatu ditu:

- `ehun eta hogeitasei (aldaera)`
- `hogeita hemeretzi`
- `4.500 milioi`
- `milioi bat`
- `180 mila`

[26] Zenbakien idazkeraren inguruko azken erabakiak hartu aurretik, hainbat proposamen eta gomendio zabaldu ziren eta horien erabilera oraindik ere oso hedatua dago. Hori dela eta, azken erabakien hedapena erabatekoa ez denez, tartean erabilitako zenbakien idazkera modu egokian identifikatzea interesgarri jo genuen.

V.3.2.2 Izen berezien bilaketa eta sailkapena

Arestian aipatu bezala, izen berezien bilaketa eta sailkapena, tesi-lan honen esparrutik at geratzen den arren, garapen-fasean dago. Bilaketarako egoera finituko transduktoreak erabiltzen dira, maiuskulaz hasitako hitzen sekuentziak identifikatu ahal izateko.

Izen bereziak bilatzeko eta sailkatzeko sistemak hiru osagai ditu:

- egitura morfosintaktikoan oinarritutako entitateen identifikatzailea
- heuristiko batean oinarritutako lehengo sailkatzailea
- ikasketa automatikoan oinarritutako bigarren sailkatzailea

Entitateen identifikatzaileak testua etiketatzen du EUSLEMen bitartez, eta, ondoren, egoera finituko teknologian oinarritutako azaleko analisi sintaktikoa burutzen du. Analisi hori, etiketa morfosintaktikoetan eta maiuskulen agerpenetan oinarrituta, entitateak izan daitezkeen izen sintagmak mugatzen ditu.

Entitate osoari dagokion deklinabide informazioa azken elementuak daramanez gero, hitzen flexioaren tratamendu egokia funtsezkoa da testuan bata bestearen atzetik jarraian agertzen diren entitate independenteak bereizteko.

Dagoeneko garatuta dagoen lehenengo sailkatzailea informazio iturri desberdinetan oinarritzen da:

- sintagmako osagaien etiketak: EUSLEMek bi etiketa bereizten ditu —2. mailatik aurrera—, leku-izenak eta pertsona-izenak. Informazio hori oso lagungarria da, zenbait kasutan desanbiguatzaileak akatsak egiten baditu ere. Dena den, batzuetan nahasgarria da interpretazio egokia EDBLn ez dagoelako.
- Entitatearen deklinabide kasua: leku-izenak inesiboan agertu ohi dira maizago eta pertsona-izenak, berriz, ergatiboan.
- pertsona- eta leku-izenen zerrendak.
- entitate baten ondoren ager daitezkeen izen arrunten zerrenda (*lehendakari, zuzendari, hiri, ibai, elkarte, S.L., ...*)

Informazio iturri desberdin hauek konbinatzen dira, algoritmoak lehen hurbilpen gisa entitate bakoitzari kategoria bat esleitzeko.

Prozesu hau 20.000 hitzeko corpus bati aplikatuko zaio eta, ondoren, eskuz errebisatuko da. Entitateak identifikatuta eta sailkatuta dituen corpus hau ikasketa automatikoan oinarritutako sailkatzaileak. Dena dela, lehenengo sailkatzaileak emaitza nahiko onak brtzen baditu, bi sailkatzaileak konbinatzea ere azter daiteke, tresnaren doitasuna hobetzearren.

V.3.3 Ebaluazioa

Datu-basean une honetan landuta dauden HAULen kopuruak honakoak dira:

- beti ziurrak direnak 900 inguru dira, horietatik 153 siglen adieraziak (*Pertsona Fisikoen Errentaren gaineko Zerga*) eta 188 latinezko adierazpenak (*a priori*) dira; gainerakoetan, lokuzio, adberbio, hitz-elkarketa eta zenbait esamoldez gain, aditz gutxi batzuk daude. Batez beste gauzatze-eskema bana dute (1,02).
- batzuetan bakarrik ziurrak direnak 300 inguru dira, batez beste 3,6 gauzatze-eskema dituzte (gehien agertzen den eskema-multzoa *12, 21, 1*2* eta *2*1*)[27] eta horien erdietan anbiguoak dira (*1*2* eta *2*1*).
- azkenik, beti anbiguoak direnak 85 inguru besterik ez dira eta batez beste 1,2 gauzatze-eskema dituzte.

Aipatutako aditz konposatuen arazoa irudikatzeko nahikoa da gauzatze-eskemen kopuruak aztertzea. Aditzak ez diren 940 HAULek 950 gauzatze-eskema dituzte, eta horietatik 70 besterik ez dira anbiguo. Aditz konposatuek, aldiz, 330 inguru izanik, 1200 gauzatze-eskema inguru biltzen dituzte, horietatik erdiak anbiguoak izanik. Aditz konposatuen atalean aipatutako *izan* eta *egin* osagaia dituzten aditzak landutakoen erdia inguru dira (%55), maiztasun handiko lemak biak ere.

Datu hauekin egiaztatzen da gauzatze-eskemak gehitzearen ideia egokia zela, HABILen aurreko inplementazioan aditz gehien-gehienak anbiguo gisa tratatu behar baziren ere, modu honetan, aditzen agerpen batzuk behintzat HAUL gisa trata daitezkeelako.

Aipatzekoa da, bestalde, HAUL ziur gehienek gauzatze-eskema bakarra dutela, hau da, bere osagai guztiak jarraian eta ordena zehatzean agertzen direla. Baina HAUL hauen erdiek flexioa onartzen dutenez, anbiguotasun morfologikoa ez da beti ebazten, nahiz eta tratamenduak neurri batean behintzat anbiguotasuna murrizten laguntzen duen.

Hitz anitzeko unitateen tratamendua ebaluatzeko 650 HAUL ziur inguru erabili dira — aipatutako ziurretatik latinezko adierazpenak eta siglen adieraziak kenduta, gainerako guztiak[28]—, beti ere jarraian agertzen diren gauzatze-eskemak soilik kontuan hartuta. Horietatik 149ren 386 agerpen besterik ez daude erreferentzia-corpusean eta 45en 87 agerpen

[27] 12 gauzatze-eskemari jarraituta: ***lo egin*** *du*;
21 gauzatze-eskemari jarraituta: ***egin lo*** *behingoan, ume!*;
1*2 gauzatze-eskemari jarraituta: ***lo*** *gutxi* ***egin*** *duzu gaur*;
2*1 gauzatze-eskemari jarraituta: ***egiten*** *al duzu inoiz 8 ordu jarraian* ***lo****?*.

[28] Berez, ez dira latinezko adierazpenen eta siglen adierazien agerpenak corpusean aurkitu eta, ondorioz, ez dira ebaluazioa egiterakoan kontuan hartu.

egiaztapenerako corpusean. Guztira 167 HAUL desberdinen 473 agerpen aurkitu dira erabilitako bi corpusak kontuan izanda, hau da, landutakoen erdia.

Data eta zenbakiei dagokienean, nahiko gutxi agertzen dira erreferentzia-corpusean, 14 data eta 12 zenbaki besterik ez. Proportzioan gehiago daude egiaztapenerako corpusean, 18 data eta 9 zenbaki, baina gogoan izan behar da lehenengo corpusak 28.000 testu-hitz inguru dituela, gehienak EEBS corpusetik jasotakoak, eta bigarrenak 8.000 inguru, horietatik gehienak *Euskaldunon Egunkaria*tik jasotakoak. Logikoa da, beraz, bigarrenean data gehiago agertzea. Gainerakoan, HAUL kopuruari dagokionean, antzeko proportzioan agertzen dira bata zein bestean.

Maiztasun handienetik txikienera, 10 agerpen baino gehiago dituzten hitz anitzeko unitateak hauek dira[29]:

- `Euskal Herria (Herria, Herriak, Herriaren,...) (23)`
- `gaur egun (egun, egungo) (22)`
- `hala ere (20)`
- `batez ere (19)`
- `nahiz eta (17)`
- `hitz egin (egin, egingo, egiteko,...) (16)`
- `Estatu Batuak (Batuekiko, Batuetako,...) (13)`
- `hain zuzen ere (12)`
- `izan ere (11)`

HAUL ziurretan, esan bezala, landutakoen erdiek flexioa onartzen dute. Ikasketa corpusean, agerpenen %22 inguru HAUL flexiodun agertzen dira, data eta zenbakiak kontuan izan gabe. Egiaztapenerako corpusean, aldiz, %50 inguru dira flexiodunak.

Maiztasun handiko flexiodunen adibideak arestian aipatutako *Euskal Herria*, *hitz egin* eta *Estatu Batuak* ditugu. Lehenengo adibidean, flexio guztiak ez dira onartzen, singularrean agertzen direnak baino ez baitira mantendu behar; bigarrenean, berriz, *egin* aditzarekin lotutako aditz-interpretazio guztiak mantendu behar dira; hirugarrenean, aldiz, pluralak bakarrik onar daitezke. Hori dela eta, anbiguotasun-maila gutxitzeko aukera ematen dute horrelako HAULek ere, nahiz eta erabat ez desanbiguatu.

Beraz, maiztasun handiko HAUL ziurrak identifikatzeak anbiguotasuna murrizten du hein batean. Horregatik, aurreikusitako HAUL guztiak erabili ez diren arren, maiztasunaren

[29] *hala ere* eta *izan ere* estatistikoki ziurrak kontsideratu ditugu. Kontradibiderik asma daitekeen arren (*hau honela egin daiteke eta hala ere*), ez dugu horrelakorik aurkitu aztertutako testuetan. Hiztunak horrelako esaldi baten aurrean ziurrenik aukera hori ekidin eta gauza bera adierazteko beste modu bat bilatzen du nahasmenik ez sortzearren. Hori izan daiteke adibiderik ez aurkituaren arrazoia.

arabera aukeratutako HAUL gutxi batzuk erabiltzeak gure hasierako hipotesia baieztatzeko modua eman digu, alegia, HAULen tratamenduak etiketatzen laguntzen duela, VI. kapituluan ikusiko den bezala.

	AR	**I/A**	**I/T**	**R**
estandar	%81,03	3,86	3,31	**%99,88**
aldaerak	%76,74	3,03	2,55	**%94,68**
ezezagunak	%82,49	4,08	3,54	**%93,03**
testu-hitzak	%80,99	3,85	3,31	**%99,54**
batez beste	%66,66	3,85	2,90	**%99,63**

V.2 taula.- Analizatzaile morfologiko hedatuaren emaitzak.

Erreferentzia-corpusaren gainean lortutako emaitzak aurkezten dira jarraian. V.2 taulako emaitzak IV.19 taulan aurkeztutako berberak dira, hemen gogoratzearren errepikatu direnak. Emaitza haien gainean —analizatzaile hobetua eta hitz ez-estandarren tratamendua aplikatzearen emaitzak kontuan izanda— hitz anitzeko unitateen tratamendua aplikatu da eta V.3 taulan agertzen diren anbiguotasun-neurriak lortzen dira[30].

	AR	**I/A**	**I/T**	**R**
estandar	%79,56	3,85	3,27	**%99,88**
aldaerak	%76,69	2,99	2,53	**%94,93**
ezezagunak	%82,49	4,08	3,54	**%93,03**
testu-hitzak	%79,60	3,84	3,26	**%99,54**
batez beste	%66,51	3,84	2,86	**%99,63**

V.3 taula.- HABILen emaitzak.

Taulan ikusten ahal da hitz estandarren anbiguotasun-tasa %1,5ean gutxitzea lortu dela, zehaztasunean irabaziz, eta garrantzitsuak diren osagaiak desanbiguatzen lagunduz, zehazki esanda, loturazko osagaiak (*hala ere, izan ere, nahiz eta, batez ere, hain zuzen ere*, etab.). Aldaerei dagokienean, zuzentasun-tasa aldatzen da, baita anbiguotasun-neurriak ere, hitzen proportzioa aldatu delako. Esan bezala, zenbakien adierazpen batzuetan aldaerak agertzen dira, *ehun eta hogeitasei* adibidean zehaztu den bezala, eta lehen aldaera gisa zenbatzen zena orain estandarren artean sartu da. Batezbesteko zuzentasuna ez da aldatzen, errorerik ez delako gehitu, baina zehaztasunean irabazi egiten da, gutxi bada ere, anbiguotasuna jaistea lortzen delako.

[30] Erreferentzia- eta egiaztapen-corpusetan lortutako emaitza osoak C eranskinean ematen dira, C.3 atalean.

Ikusteke dago aditzen gauzatze-eskema ez-anbiguoen lanketak ekarriko lukeen onura, baina emaitza hauetan oinarrituta, hobekuntza esanguratsua lor daitekeela uste dugu.

V.4 Etorkizunerako hobekuntzak

Etorkizunari begira, hainbat hobekuntza plantea daitezke lan honen jarraipen gisa. Lehenengoa, garapen-maila aurreratuan dagoela-eta, izen berezien identifikazio eta sailkapena integratzea litzateke.

Datu-basean landutako hitz anitzeko unitateei dagokienean, lan honetan erabili ez diren gainerako unitate ziurrak aplikatzeaz gain, unitate gehiago lantzea interesgarria gerta daiteke, baina lanketarako oinarri bezala bi printzipio kontuan hartuta:

- lehenengoa, datu-basean landutakoen kopurua gehiegi ez handitzea. Orokorrean testu laburrak tratatzen direnez, hitz anitzeko unitate desberdin asko egoteko probabilitate txikia dago eta tamaina neurri gabe handitzen bada, erabiltzen diren datuak kargatzeko testu osoa tratatzeko baino denbora gehiago beharko delako. Datu-basea aberasterakoan bereziki maiztasun txikiko unitatean baztertu beharko lirateke.
- bigarrena, berriz, praktikotasunari begira, datu-basean landu beharreko unitateak maiztasun handikoak izatea. Horretarako, Xtract edo LocalMax moduko aplikazioren bat erabil daiteke, datu-basea modu egokian aberasteko.

Azkenik, HAUL ez-ziurrak eta ez-jarraiak integratzeko modu egokia aztertu behar da. Izan ere, osagaiak jarraian agertzen ez direnean, desanbiguazioari begira trabak sortzen dira, osagai bat HAUL dela erabakitzen denean, aldi berean gainerako osagaiak modu berean desanbiguatu behar dira. Hori egitea ez da berehalakoa murriztapen gramatikaren bitartez, are gutxiago metodo estokastikoen bitartez. Horregatik geratu da ikerkuntza-lan honetatik kanpo.

VI Desanbiguazio morfosintaktikoa

Aurreko kapituluak hitz bakunen tratamendu morfosintaktikoari eta hitz anitzeko unitateen tratamenduari eskaini zaizkio. Prozesu horien emaitza testuko unitate bakoitzaren interpretazio morfosintaktiko posible guztiak dira, nahiz eta gainsorkuntzaren ondorioak saihesteko zenbait teknika aplikatu diren, ahalik eta zehaztasun handienaz eta anbiguotasun txikienaz emanez testuinguruari erreparatu gabe.

Tratamendu morfosintaktikoa hainbat tresnaren oinarri izanik, aplikazioetatik independente diseinatu da. Tresna horietako bat, kapitulu honetan proposatzen den desanbiguatzailea da, testuinguruaren arabera zuzenak ez diren interpretazioak baztertu eta, ahal dela, bakarra uzten duena. Izan ere, hitza isolaturik anbiguoa izan ohi da, baina hitzak testuinguru bakoitzean interpretazio bakarra du eta, beraz, gainerakoak baztertu behar dira.

Baina prozesamendu morfologikoak oso emaitza aberatsa ematen duenez, zaila gertatzen da informazio hori guztia aldi berean erabili eta testua erabat desanbiguatzea. Beraz, aurreko prozesuetatik jasotako informazioa nolabait kodetu egingo da, interesatzen zaigun informazioa mantenduz. Horretarako, II. kapituluan deskribatutako etiketa-sistema definitu da, aplikazioaren arabera komeni dena aukeratuz desanbiguatuko baita. Argi gera bedi etiketa-sistema orokorra eta independentea diseinatu dela, eta ez dela desanbiguazioaren emaitzak hobetzearren egokitu, edota ebazten zailak diren anbiguotasunak desagerrarazi.

Desanbiguazioa egiteko moduari dagokionean, teknika desberdinak erabili izan ohi dira. Orokorrean bi ildo jarraitzen dira gaiari buruzko literaturan ikus daitekeenez (van Halteren (*ed*.) 1999); batetik, etiketatzaile linguistikoak daude, eta bestetik, datuetatik erauzitako etiketatzaileak[1] (*data-driven taggers*). Lehenengo taldekoak eskuz garatzen diren bitartean, bigarren taldekoak, corpusetako informazioa automatikoki ustiatzen dute. Azken hauen artean

[1] Corpusetatik automatikoki erauzitako hizkuntz ereduetan oinarritutako etiketatzaileak.

gainbegiratuak eta ez-gainbegiratua bereizten dira, eskuz desanbiguatutako corpus baten beharraren arabera.

Beste sailkapen batzuk ere aurki daitezke bibliografian. Hasierako sailkapenetan erregeletan oinarritutakoak eta datu estatistikoetan oinarritakoak bereizten ziren. Bereizketa honen arrazoi nagusia, erregelak eskuz idatzi ohi zirela zen, baina erregelak automatikoki erauzten zituzten sistemak agertzearekin batera, sailkapena aldatu da. Horrela, erregelak eskuz garatzen dituzten sistemak (etiketatzaile linguistikoak) gainerakoetatik (corpusetatik automatikoki erauzitako hizkuntza ereduak erabiltzen dituzten etiketatzaileak) bereizten hasi dira. Gainera, ikasketa automatikoan (*machine learning*) oinarritutako etiketatzaileak corpusetatik datuak automatikoki erauzten dituzten gainerako etiketatzaileetatik banatzen dituztenak ere badira, etiketatzaile estatistikoak (n-grametan oinarritutakoak eta Markov-en eredu ezkutuetan oinarritutakoak) beste multzo batean bereizita, azken hauek hizkuntz eredu sinpleagoak erabili ohi dituztelako (Màrquez 1999). Dena dela, guk lehenengo sailkapenari jarraituko diogu (van Halteren (*ed.*) 1999) gaiari buruzko azterketa bibliografikoa egiterakoan.

Ildo biei, linguistikoa zein datuetatik erauzitakoa, jarraitzen dieten sistema ugari garatu da, eta bata zein besteaz emaitza onargarriak lortu arren, literaturan kontrajarri izan dira luzaroan. Baina azken aldian teknika desberdinak konbinatzea onuragarria izan daitekeela ikusi da. Bai multzo bereko bai multzo desberdineko etiketatzaileen arteko konbinaketak eginez emaitzak hobetzen dira, aurrerago ikusiko den bezala.

Guk ere hasieratik sistema desberdinen konbinaketa helburu genuela, bi desanbiguazio-sistema garatu ditugu modu independentean, ondoren sistema konbinatua osatu eta kapitulu honetan aurkezten diren emaitzak lortuz.

Ondorengo atala teknika desberdinen azterketari eskaintzen zaio, bakoitzaren ezaugarriak aurkezten dira eta, gure helburuak kontuan izanik, ezaugarri horien arabera zein bidetik jo den azaltzen da. Bigarren atalean, berriz, euskararen desanbiguazioa deskribatzen da, metodo bakarra erabilita teknika bakoitzarekin egindako saiakuntzak eta emaitzak aurkeztuz. Eta, azkenik, zenbait ondorio aipatzen dira.

VI.1 Desanbiguaziorako teknikak

Arestian aipatu den bezala, orokorrean bi ildo bereizi ohi dira desanbiguazio morfosintaktikoari ekiteko tekniketan. Batetik, etiketatzailea linguistikoak daude, eta, bestetik, corpusetatik automatikoki erauzitako hizkuntza ereduak erabiltzen dituztenak edo *data-driven taggers*.

Màrquez-ek (1999) egungo etiketatzaileen inguruko azterketa bibliografiko sakona egin du, lan garrantzitsuenak aztertu eta sailkatu dituelarik. Gainera, etiketatzaileen inguruko informazioa lortzeko helbide garrantzitsuenak ematen ditu. Hori dela eta, teknika desberdinei buruzko informazio sakonagoa lortzeko (Màrquez 1999) tesi-lanera jotzea gomendatzen da (*http://www.lsi.upc.es/~lluism*). Beste informazio iturri garrantzitsua (van Halteren (*ed.*) 1999) lana da. Bertan, besteak beste, etiketatzaile-mota desberdinei lau kapitulu eskaintzen zaizkie, egungo etiketatzaileen egilerik aipagarrienen eskutik. Ondorioz, tesi-lan honetarako erabili diren tekniketan sakonduko da, gainerako tekniketan, berriz, modu laburrean aipatu eta erreferentzia nagusienak emango dira soilik.

VI.1.1 Etiketatzaile linguistikoak

Esan bezala, etiketatzaile linguistikoen[2] ezaugarri nagusiena eskuz idatzitako erregelak erabiltzen dituztela da. Orokorrean, erregela hauek testuinguruaren arabera hautagaiak aukeratu edota baztertu egiten dituzte, baina ez dute gramatika oso bat deskribatzea helburu.

Gramatika idazterakoan, hizkuntzalaria hizkuntzari buruzko orokortasunak ustiatzen saiatzen da eta horretarako oinarri anitz erabiltzen ditu, hala nola, hizkuntzari buruz duen ezagumendua, gramatika deskribatzaileetako informazioa, hiztegiak eta bertako erabilera adibideak, edota testu-lagin handien azterketatik ondorioztatutako ezaugarriak.

Beraz, gramatika idazten duenaren eskuetan egongo da produktuaren kalitatea, hizkuntzalariak duen orokortzeko gaitasunak definituko baitu etiketatzailearen doitasuna eta sendotasuna.

Etiketatzaile hauen ezaugarri nabarmenenak honakoak dira: batetik, pertsona batek bere ezagumenduaz idatziak direnez, deskribapena adierazgarriagoa eta ulerterrazagoa da, eta, bestetik, zehatzak eta doitasun handikoak izan ohi dira.

Gaur egunean, gainera, algoritmo oso eraginkorren bidez erabiltzen dira gramatikak, aspaldian etiketatzaile hauek jasotzen zuten kritikarik gogorrenari, hots, motelak izateari, aurre eginez. Aipatzekoa da, besteak beste, arlo honetan egoera finituko transduktoreen bitartez Roche eta Schabes-ek (1995) egindako lana.

Hala ere, eskuz idatzitako gramatikak orokorrean garestiak dira, batez ere lortu nahi den doitasuna handia bada. Anbiguotasunaren parte esanguratsua erregela gutxi batzuk idatzita

[2] (van Halteren (*ed.*) 1999) liburuko 14. atalean (217-246) mota honetako etiketatzaileei buruzko azterketa aurkezten du Atro Voutilainen-ek.

ken daiteke[3], baina gainerako anbiguotasuna kentzea lana latzagoa izan ohi da, bereziki doitasuna eta zehaztasuna mantendu nahi izanez gero. Beraz, doitasun handiko eta anbiguotasun gutxiko irteera lortu nahi denean, erregela-multzoa birfindu eta handitu behar izango da, lan honek urteak ere eraman ditzakeelarik.

Mota honetako etiketatzaileen aurrekaria 70. hamarkadan *Brown Corpus*a (Kucera eta Francis 1967) etiketatzeko garatutako TAGGIT etiketatzailea da (Greene eta Rubin 1971). TAGGITek 86 etiketa erabiltzen ditu, geroagoko zenbait lanetako etiketa-sistemen oinarri izan dena.

Eskuz landutako 3.300 testuinguru-erregela aplikatzen dira, testuinguru gisa gehienez 5 hitzeko leihoa erabiliz. Programa honen doitasuna txikia da, %77 ingurukoa, eta etiketatzaile automatiko baten helburua testua osorik eta zuzenki etiketatzea denez, sistema honek oraindik urrun du helburua. Hala ere, garatutako lehenetariko sistema da hauxe, eta gainontzeko lanen erreferentzia garrantzitsua izan da beti. Izan ere, estatistikan oinarritutako sistema askotan oinarri gisa erabili da bai *Brown Corpus*a baita bere etiketa-sistema ere.

Sistema honen oinordeko hedatuena Karlsson-en *Constraint Grammar* edo murriztapen-gramatika da (Karlsson *et al.* (*eds.*) 1995). Murriztapen-gramatika formalismoak ataza desberdinak biltzen ditu, nagusienak desanbiguazio morfologikoa, mapaketa morfosintaktikoa eta funtzio sintaktikoen desanbiguazioa izanik; baina desanbiguatzeko heuristikoak ere defini daitezke. Gramatika honen bidez, hitzaren ahalik eta interpretazio gehien baztertu nahi dira. Horretarako, testuinguru-erregelak edo murriztapenak erabiltzen dira analisiak aukeratzeko edota baztertzeko. Testuinguru zabala erabiltzen du, leihoa esaldiaren mugen artekoa da eta, TAGGITen bost hitzetako muga gaindituz.

Hizkuntza anitzetarako idatzi dira murriztapen-gramatikak[4], guztiak publikatu ez badira ere. Horien artean ingeleserako EngCG (Karlsson 1990; Voutilainen *et al.* 1992; Karlsson *et al.* (*eds.*) 1995; Tapanainen 1996), turkierarako (Oflazer eta Kuruöz 1994; Oflazer eta Tür 1996), frantserako (Chanod eta Tapanainen 1995-b), suomierarako, suedierarako, swahilirako (Hurskainen 1996), danierarako, alemanerako, portugeserako, gaztelerarako (Sánchez 1997) eta euskararako (Aduriz 2000) murriztapen-gramatikak aipa daitezke.

Turkierarako desanbiguatzaileak corpusetik erauzitako erregelak eta eskuz idatzitako erregelak erabiltzen ditu (Oflazer eta Kuruöz 1994; Oflazer eta Tür 1996) eta EngCGk duen

---

3 (van Halteren (*ed.*) 1999) lanean Atro Voutilainen-en *NorFa'95 CG 'competition'*i buruzko atalean (240) ikus daiteke 14 ordutan egindako ingeleserako CG gramatika batek anbiguotasunaren 2/3 kentzen zuela %0,58ko errorea eginez.

4 Erreferentzia guztietan ez da *Constraint Grammar* formalismoa erabiltzen baina murriztapen-erregelak aplikatzen dituzten etiketatzaileak dira.

arazoetako bat ebazteko bidea irekitzen du, alegia, erregelen aplikazio-ordenarena. Oflazer eta Tür-ek (1996) corpusetik lortutako erregelei pisuak esleitzen dizkiete eta ordenaren independenteki aplikazio-sekuentziarik onena aukeratzen dute. Turkieraren kasuan, hasierako anbiguotasuna mahiko altua da, tokeneko 1,8-1,9 interpretazio dituela. Hala ere, emaitza onak lortzen dituzte, %95-97 arteko zuzentasuna eta %94-96ko zehaztasuna (1,01 analisi eskas hitzeko).

Ingeleserako ere aplikatu dute, informazioa erauzteko *Brown Corpus*eko 100.000 hitz inguru erabilita, eta eskuz idatzitako erregelak erabili gabe. Emaitza gisa, %96 inguruko zuzentasuna lortzen dute, eta erregela horiek erabilita (800 inguru) batez beste %97,5eko zuzentasuna. Hitzeko 1,05-1,09 interpretazio utziz gero, zuzentasuna %98,48-98,78koa da.

EngCGk 1.100-1.300 erregela ditu. Sarreran batez beste 1,8-2 interpretazio ditu hitzak eta irteeran hitzen %99,7-99,8k interpretazio zuzena mantentzen dute, interpretazio kopurua 1,05-1,09koa izanik. EngCG-2, etiketatzailearen bigarren bertsioak[5], 3.744 desanbiguazio-erregela ditu eta aurrekoak baino errore gutxiago eginez anbiguotasunaren erdia uzten du.

Geratzen den anbiguotasuna ebazteko bestelako metodoetara jo behar izaten da, teknika estatistikoak edota beste nolabaiteko informazioa erabilita. VI.1.3 atalean aipatuko dira emaitzak hobetzeko saiatu diren metodoen konbinaketa batzuk.

VI.1.2 Automatikoki erauzitako datuetan oinarritutako etiketatzaileak

Corpusetatik automatikoki erauzitako hizkuntz ereduetan oinarritutako etiketatzaileek (*data-driven taggers*) testu-lagin handien azterketa estatistikoa automatikoki burutuz hizkuntz ereduak lortzen dituzte. Ereduaren sofistikazio-mailaren arabera, esan bezala, bi talde bereizten dira: batetik, hizkuntz eredu sinpleak dituzten etiketatzaile estatistikoak, eta, bestetik, eredu sinbolikoen paradigma klasikoak erabiltzen dituztenak (Màrquez 1999) edo ikasketa automatikoaren bidezko etiketatzaileak.

Hizkuntz eredu sinpleen multzoan, batzuk etiketatzaile estatistikoak soilik eta beste batzuk Markov-en eredu ezkutuetan oinarritutakoak ere biltzen dituzte. Gainerakoan, informazio-mota desberdina erauzten dituzten etiketatzaileak izango dira, besteak beste erregelak, erabaki-zuhaitzak edota sare neuronalei buruzko informazioa lortzen dutelarik.

[5] Etiketatzailearen dokumentazioa eta demostrazio elkarreragilea URL honetan topa daiteke: http://www.ling.helsinki.fi/~avoutila/engcg-2.html

Ondoren, etiketatzaile-mota nagusienak aurkezten dira eta, historikoki etiketatzean nolabaiteko iraultza edo aurrerapauso nabarmena suposatu duten hiru mota zabalduko dira: lehenengoa, etiketatzaile estatistikoak (VI.1.2.1), bide berria ireki eta ordura arteko emaitzarik onenak lortu zituztenak; ondoren, Markov-en eredu ezkutuetan oinarritutako etiketatzaileak (VI.1.2.2), ahotsaren prozesamenduan arrakasta handia lortu ondoren, testu-tratamenduan aplikatu eta etiketatzaileen garapenean jauzi kualitatibo handia lortu zutenak; azkenik, corpusetatik desanbiguazio-erregelak erauzi eta erroreetan oinarriturik erregelak egokitzen dituen Brill-en etiketatzailea (VI.1.2.3), erregelen bidezko etiketatzaileak modu automatikoan sortzeko bidea ireki zuena.

Gainerako etiketatzaile-motak hauen emaitzen gainetik ekarpen berriak lortu dituzte eta gaur egunean ikasketa automatikoa da, argi eta garbi, gehien jorratzen eta erabiltzen den hurbilpena. Hala ere, gehiegi ez luzatzearren, oinarrizko aurkezpena besterik ez da egingo, baina gaian sakontzeko erreferentzia interesgarriak ematen dira kasu bakoitzean.

Datuetatik erauzitako etiketatzaileen hurbilpenak, beraz, hauexek dira:

- Etiketatzaile estatistikoak edo n-grametan oinarritutakoak: (Garside *et al.* 1987), (Church 1988) eta (deRose 1988) dira atal honetako erreferentzia nagusienak. Erregeletan oinarritzen ez ziren lehenengo etiketatzaileak izanik, ondoren atal bat eskaini zaie.
- Markov-en eredu ezkutuak (HMM)[6]: ahotsaren tratamenduan arrakasta handia izan zuen eredua, idatzizko hizkuntza tratatzeko egokia izan zitekeela ikusi eta ingelesezko etiketatzailea garatu zen (Cutting *et al.* 1992). Geroztik, lortutako emaitzei esker, hizkuntza gehiagotarako ere erabili dira. Batzuk aipatzekotan txinatarra (Chang eta Chen 1993), suediera (Cutting 1994), gaztelera (Sánchez eta Nieto 1995), frantsesa (Chanod eta Tapanainen 1995-b), alemanera (Feldweg 1995) eta euskara (Ezeiza 1997), besteak beste.
- *Transformation-Based Error-Driven rules*[7]: etiketatzaile honek (Brill 1992, 1995) desanbiguazio-erregelak corpusetatik automatikoki lortzen ditu. Aurreko eta atzeko 3 hitzak erabiltzen ditu testuinguru gisa, hitz bati dagokion etiketa aukeratzerakoan. Hasierako bertsioan eskuz markatutako corpusa erabiltzen zuen soilik eta kritikak ere jaso zituen, emaitza txukunak lortzeko eskuz desanbiguatutako corpus handia behar zelako. Brill-ek (van Halteren (*ed.*) 1999) corpus gordinetik erregelak erauzteko algoritmoaren egokitzapena egin du baina emaitza ez dira corpus desanbiguatuarekin bezain onak. Hala ere, corpus gordin handitik erregelak erauzi eta eskuz desanbiguatutako corpus txiki batekin erregelak doituz gero, hasierako etiketatzailearenak baino emaitza hobeak lortzen

6 (van Halteren (*ed.*) 1999) liburuko 16. atalean (263-284) mota honetako etiketatzaileei buruzko azterketa aurkezten dute Marc El-Beze-k eta Bernard Merialdo-k.

7 (van Halteren (*ed.*) 1999) liburuko 15. atalean (247-262) etiketatzaile honi buruzko azterketa aurkezten du Eric Brill-ek.

direla dio. Oso etiketatzaile eraginkorra da, batetik, erregela kopurua nahiko txikia delako, eta, bestetik, erregelak algoritmo azkar batek aplikatzen dituelako. Etiketatzaile honi ere atal bat eskaintzen zaio ondoren.

- *Case-Based / Memory-Based*[8]: adibideetan oinarritutako etiketatzaile honek corpus batean ikasitako kasuak memorian dituela, sarrerako testuan horien antzeko agerpenak etiketatzen ditu (Daelemans *et al.* 1996). Kasuei buruzko informazio anitz erauzten du, ezaugarri bakoitzari pisu desberdina emanez. Pisu horiek, sarrerako testuan topatutako agerpenekin antzekotasunik duen erabakitzeko erabiliko dira, beharrezkoa denean adibideetatik orokortzea eginez. Zuhaitz-adierazpidea erabiltzen du informazio hori modu eraginkorrean kodetzeko eta bilaketak burutzeko.
- Erabaki-zuhaitzak (*Decision Trees*): zuhaitz hauetan erpinak ezaugarriak dira eta arkuak ezaugarri horien balio posibleak, hostoak kategoriak izanik. Kasu honetan, orokortze edo abstrakzioa ikasketa-fasean burutzen da, eta ikasketaren emaitza zuhaitz-multzo bat izango da. Anbiguotasun-klase bakoitzeko zuhaitz bat eratzen da ikasketa-prozesuan, klase horretako hitzen informazioa eta inguruko hitzei buruzko hainbat ezaugarri erabilita. Datu horietan oinarrituz, etiketatzea zuhaitz horien errotik hasi eta sarrerako ezaugarrien balioen arabera hostoetara iritsi eta etiketa esleitzean datza. Lehenengo lana IBMn burutu zen (Black *et al*. 1992), gerora Schmid-ek (1994-b) TREETAGGER eta Magerman-ek (1995) SPATTER aurkeztu zituzten. Azken aldian etiketatzaile asko aurkeztu dira eredu honi jarraituz, hainbat hobekuntza burutuz, besteak beste aipatutako Màrquez-en (1999) lana.
- Sare Neuronalak (*Neural Networks*): hainbat esparrutan erabilitako eredu hau etiketatzeari ere aplikatu zaio. Ingeleserako hurbilpenetako bat Schmid-ena (1994-a) da. Etiketatzaile honek trigrametan oinarritutako etiketatzaile baten antzera funtzionatzen du, baina parametro gutxiagoren estimazioa burutu behar du. Printzipioz, n-grama luzeagoetara ere heda daiteke. Eredu hau portugeserako ere aplikatu da (Marques eta Lopes 1996) lanean azaltzen den bezala. Thai lengoaiarako hurbilpenean (Ma eta Isahara 1998), luzera desberdinetako testuinguruak erabiltzen dira luzera bakoitzeko sare desberdinak erabiliaz (*Multi-Neuro tagger*). Hitz anbiguoak (%34 inguru) soilik kontuan hartuz %94,3ko zuzentasuna lortzen dute, trigrametako etiketatzailearekin lortutako %89,3ren aurrean. Beste hurbilpen bat ere proposatzen dute, sarearen tamaina mugatzeko (Ma *et al*. 1999-a), *Elastic-Neuro tagger* deitzen dutena, aurrekoaren emaitzak pixka bat hobetzen dutena, %94,4. Azkenik, (Pérez-Ortiz eta Forcada 2001) lanean, ikasketarako corpus txikia erabilita (46.000 hitz inguru) ikasketa ez-gainbegiratua burutu da bi eredu erabilita:

[8] (van Halteren (ed) 1999) liburuko 17. atalean (285-304) *case-based learning*, erabaki-zuhaitzak eta sare neuronalak erabiltzen dituzten etiketatzaileei buruzko azterketa aurkezten du Walter Daelemans-ek.

batetik, HMMak eta, bestetik, sare neuronal errekurrenteak erabiltzen dituzte. Ereduen aplikazioaren emaitzak baliokideak dira, %92ko zuzentasuna lortuz. Aurreko hurbilpenekin konparatuz, eskuz desanbiguatutako corpusik erabiltzen ez duen sare neuronal bidezko bakarra dela esan behar da.

- Entropia maximoa (*Maximum Entropy*): etiketatzaile honek (Ratnaparkhi 1996) aurreko eta atzeko 2 hitzak, aurreko 2 hitzen etiketa eta desanbiguatu beharreko hitzaren zenbait ezaugarri ere erabil ditzake etiketa aukeratzeko. Ezaugarriek balio bitarra izango dute, hau da, prozesatzen ari den hitzak ezaugarri hori betetzen du ala ez. Balio horiek ereduaren parametroei aplikatzen zaizkie, hitzaren etiketak testuinguru zehatz horretan duen probabilitatea kalkulatzeko. Parametro horiek lortzeko ikasketa-corpusaren gainean —kasu honetan *Wall Street Journal*eko milioi bat hitz inguru— *Generalized Iterative Scaling* algoritmoa erabiltzen du (Darroch eta Ratcliff 1972). Etiketatze-prozesuari dagokionean, oinarrian *beam search* algoritmo bat da, une oro esaldian orain arte tratatutako hitzen N etiketa-sekuentzia onenak gordetzen dituena. Hurrengo hitza tratatzean, konbinaketa posible guztiak kalkulatu eta berriz ere N onenekin geratzen da, esaldiko hitzen etiketa guztien arteko konbinazio guztiak ez aztertzearren. Sekuentziarik onenak aukeratzeko sekuentziaren entropia maximizatzen dituztenak hartzen dira. Azkenik, esaldi osoa tratatu ondoren, sekuentziarik onena aukeratzen da. Emaitzek %96,63ko zuzentasuna dute eta hitz ezezagunetan %85,56koa, gainerako hurbilpen gehienek baino zertxobait hobea.

Ondoren, etiketatzaile estatistikoak aurkezten dira, metodo estatistikoak erabiltzen dituzten lehen etiketatzaileak izanik, erreferentzia ezinbestekoa direlako. Jarraian, euskararen desanbiguazioan erabilitako Markov-en eredu ezkutuei (HMM) atal bat eskaintzen zaie. Azkenik, Brill-en etiketatzailea ere aurkezten da, erregelak erabiltzen dituzten etiketatzaileen aurkako iritzi guztiak zapuztu zituelako, erregela horiek modu automatikoan erauz daitezkeela frogatuz.

VI.1.2.1 Etiketatzaile estatistikoak

Hizkuntz eredurik sinpleenak n-gramen probabilitateak edo agerkidetza-maiztasunak dira, eta orokorrean agerkidetza-matrizeen bidez adieraziko dira. Aldez aurretik desanbiguatutako corpusetatik lortzen dira maiztasun-neurriak eta emaitzak esanguratsuak izateko corpusa handia izatea komeni da, ehunka mila hitzetatik milioietarako corpusak erabili izan direlarik.

Teknika hau erabiltzen duen sistemetako bat CLAWS1 da (Garside *et al.* 1987). Sistema hau *Lancaster-Oslo/Bergen Corpus of English* (*LOB Corpus*) etiketatzeko garatu zen eta erabilitako etiketa-sistema *Brown Corpus*ean erabilitakoaren antzekoa da. Hala ere, etiketa berri batzuk erantsi dira —guztira 130 etiketa—. Sistema honetan bigramak erabiltzen dira eta aipatutako TAGGIT etiketatzailean oinarriturik dago —TAGGITen bertsio estatistikoa

kontsidera daiteke—. Hain eredu sinplea erabilita oso emaitzak lortu ditu (%96ko zuzentasuna).

Antzeko sistemak proposatu dituzte Church-ek (1988) eta deRose-k (1988). Azken honek, etiketa sekuentzia hoberena lortzeko bide-sare osoa erabili ordez, zatika tratatzen du. Honela, programazio dinamikoaren ezaugarriez baliatzen da abiadura hobetzeko *VOLSUNGA* algoritmoan. Sistema hauek ere emaitza paretsuak lortzen dituzte.

VI.1.2.2 Markov-en eredu ezkutuak (HMM)

Duela urte gutxira arte ahotsaren tratamenduan erabiliak ziren Markoven eredu ezkutuak (HMM). HMMek ez dute aldez aurretik desanbiguatutako corpusen beharrik, parametroen hasierako estimazioa *Baum-Welch* edo *Forward-Backward* (Baum 1972) algoritmoaren bidez egin daitekeelako[9]. Dena dela, askotan komeni da desanbiguatutako corpus zatiren bat erabiltzea, modu horretan lortutako parametroen balioekin emaitzak hobe daitezkeelako[10].

Probabilitateak ikasi direnean, hau da, hizkuntza eredua sortu denean, etiketatze-prozesuan (*tagging*) erabiltzeko prest dago. Prozesu honek esaldiaren esparruan dauden etiketenen sekuentziak jasoko ditu. Desanbiguazio-prozesuak *Viterbi*ren algoritmoaren bidez probabilitate handieneko etiketa-sekuentziak aukeratuko ditu ikasi duen HMMa erabiliaz. Beste aukera bat da *Maximum Likelihood* algoritmoa erabiltzea da. Modu honetan, hitzari dagokion etiketarik egokiena aukeratzen saiatzen da eta ez etiketa-sekuentziarik egokiena. Hala ere, zailago da inplementatzen, gainera sekuentzia inkoherenteak eman ditzake eta emaitza oso antzekoak lortzen dira metodo honen bitartez (Merialdo 1994). Hori dela eta, gehienetan *Viterbi* erabili ohi da.

Hizkuntza askotarako erabiliak izan dira HMMak, emaitza onak lortuz. Ingeleserako, esate baterako, %95-%97ko doitasuna lortu da hizkuntz eredu hauen bitartez, ikasketa gainbegiratua erabilita emaitza hobeak lortu direlarik. Hala ere, eredu hauekin %97ko muga hori gainditzea zaila gertatzen da, etiketa-sistema edota hasierako etiketen esleipena aldatu gabe, behinik behin.

HMMen bidezko hainbat inplementazio definitu dira, arestian aipatu direnak, kasu. Ez dira sistema guztiak hemen aipatuko, luze joko bailuke, baina lan garrantzitsuenen erreferentziak bibliografian aurki daitezke. Markov-en eredu ezkutuei buruzko azalpen orokorra, berriz, (Rabiner eta Huang 1986) lanean aurki daiteke. Bi sistema azterten dira ondoren, *Xerox*

9 Ikasketa-modu honi ez-gainbegiratua (*unsupervised*) esaten zaio.

10 Ikasketa-modu honi gainbegiratua (*supervised*) esaten zaio.

Tagger (XT) etiketatzailea (Cutting *et al*. 1992) eta MULTEXT (Amstrong *et al*. 1995) proiektuko etiketatzailea.

VI.1.2.2.1 XT Xerox etiketatzailea

XT (Cutting *et al*. 1992) lehen ordenako eredu batean oinarritutako etiketatzaile estokastikoa da, hau da, bigramen neurrietan oinarritutakoa. Sistema honek duen berrikuntzarik garrantzitsuena anbiguotasun-klaseen[11] erabilera da. Hau da, ereduak sortzen dituen sinboloak hitzak izan beharrean, anbiguotasun-klasea osatzen duen etiketa-multzoak izango dira. Honela, eredua lexikoarekiko independentea izango da. Ezaugarri hau garrantzitsua da hitz-formen zerrenda bat lortzerik ez dagoenean.

Berrikuntza honi esker, kalkulatu beharreko probabilitateak askoz gutxiago dira, anbiguotasun-klase batean dagoen etiketa kopurua etiketa bera duten hitzen kopurua baino askoz txikiagoa delako. Kupiec-ek erabili zuen lehenengoz teknika hau (Kupiec 1989, 1992) *Brown Corpus*ari aplikatuta. Corpus honek 50.000 hitzetako lexikoa darabil eta metodo honen bitartez 400 anbiguotasun-klase ingurutan bil daitezke. Honek ematen digu hobekuntza honen neurria.

XT etiketatzailearen arkitektura erabat modularra da, modulu bakoitzak funtzio zehatza izanik eta protokolo estandarrei jarraiki. Honela, moduluren baten inplementazioa ordezkatu nahi izanez gero ez da gainerako guztia ukitu behar.

Sarrerako modulua tokenizatzailea da. Modulu honek sarrerako testua unitateetan banatuko du, baina hitzak eta esaldia mugatzeko markak besterik ez dira bereiztuko. Tokenizazio sinplea egiten da baina token-mota gehiago bereiztu nahi bada, beti dago modulua ordezkatzeko aukera.

Aurreko pausoan bereiztutako tokenak lexikoian bilatzen dira dagozkien etiketak eskuratu ahal izateko. Honela lortuko da hitzen anbiguotasun-klasea. Modulu hau analizatzaile morfologiko batekin ordezka daiteke, hiztegi-bilaketa hutsa nahikoa ez duten hizkuntzetan erabili ahal izateko. (Schiller 1996) lanean, gaztelera, italiera, portugesa, alemanera eta frantsesa tratatzeko erabili da aukera hau. Horretarako, transduktore lexikoak erabili dute.

Behin token bakoitzak dagokion anbiguotasun-klasea izanik, ikasketa-prozesuari ekingo zaio —aldez aurretik ereduak ikasi ez badu behintzat—. Ikasketa-moduluak anbiguotasun-klaseen sekuentzia luzeak hartuko ditu ikasketarako. Esaldiaren mugen artean

[11] Hitz baten anbiguotasun-klasea hitz horrek izan ditzakeen etiketa posible guztiek osatutako multzoa da.

ikasten dela esan ohi da, baina orokorrean bi token ez-anbiguoen arteko anbiguotasun-klaseen sekuentziak hartzen dira ikasketarako.

XT edonork eskura dezake eta bere erabilera oso zabaldua da, batez ere oso inplementazio eraginkorra lortu delako. Dena dela, aurreprozesua oso sinplea du, hizkuntzekiko independentea dela diote, baina hizkuntza bakoitzaren berezitasunak modu egokian tratatu ahal izatea komeni da.

VI.1.2.2.2 MULTEXT proiektuko tresnak

MULTEXT proiektuan hizkuntza desberdinetarako baliabideak lortu eta hauek tratatzeko tresnak diseinatu eta inplementatzen dira, besteak beste, analizatzaile morfologikoa eta etiketatze-prozesurako tresnak. Lan honetan azken hauei buruzko ezaugarri nagusiak aztertuko ditugu.

Eredu markoviarrak sistema askotan erabili izan dira, arestian esan bezala, doitasun-maila onargarria lortzen saiatu direlarik —gutxienez %95—. Baina sistema gehienak analizatzaile morfologikoaren beharrik ez dute aurreikusten, hiztegi-bilaketa hutsean oinarrituz, eta, gainera, etiketa-sistema finko bati lotuak diseinatu dira.

MULTEXT proiektuan, aldiz, etiketa-sistema desberdinak erabiltzen direla ikusirik, sarrerako testuaren formatuak etiketa-sistemarekiko independentzia izan dezan, testua ikasketarako zein etiketatzerako etiketa-multzoaren arabera automatikoki prestatzen duen tresna eskaintzen da. Bestetik, sistema publiko askotan hizkuntzarekin lotutako baldintzak ezartzen dira, adibidez, tokenizazio eta segmentazioari dagozkien baldintzak. Beste zenbait sistematan tokenizazioa etiketatzailean integraturik dagoenez, ez dago beste moduluekin konbinatzerik eta aldaketak egin nahi badira, zailtasun handiak sortzen dira. Hau ez da MULTEXTeko inplementazioan gertatzen.

Bestalde, sistema askotan etiketa posibleak egokitzeko zerrenda bat —edo hiztegia— erabiltzen da. MULTEXTeko tresnak, berriz, analizatzaile morfologikoaren moduluaren ondoren aplikatzeko diseinatu dira. Analisi morfologikoa eskatzen duten hizkuntza edota aplikazioetarako oso egokia da ezaugarri hau, eta behar ez dutenei ez die trabarik egiten, hiztegi-bilaketaren bidez aukerak esleitzen ahal direlako.

Laburbilduz, MULTEXTeko tresnen ezaugarri nagusiak honakoak dira:

- **Hizkuntzarekiko independentzia**: etiketatze eleanitza egin nahi denean, hizkuntzaren baliabideak —lexikoiak eta etiketen definizioak— eta hauek erabiliko dituzten programak garbi bereizten dira.

- **Modulartasuna**: ataza guztiak ondo bereizten dira, horretarako programa bakoitzak ataza bakarra betetzen du, eta programen arteko komunikazioa modu zehatzean dago definituta. Ataza-banaketa honek erabiltzaileari moduluak bakarka zein taldeka aztertzeko aukera ematen dio. Honela, nahi izanez gero, zenbait ataza ordezkatu edota kentzeko arazorik ez du.
- **Malgutasuna**: batetik, hizkuntza bakoitzak izan ditzakeen berezitasunak alde batera utzi eta tratamendu orokorra bilatu da, eta, bestetik, etiketa-sistema desberdinekin saiatzeko aukera ematen du, behin-betiko sistema definitu bitartean probak egiteko aldaketa handiak egin beharra izan gabe sarrerako datuetan. Etiketa-multzo anitz erabili nahi izanez gero ere oso interesgarria da ezaugarri hau.

Eskaintzen diren tresnen artean, sarrerako testua ikasketarako zein etiketatzerako prestatzeko programa dago. Honela, sarrerako testuan analizatzaile morfologikoak ematen digun informazio guztia mantenduko dugu eta programak analisietatik etiketetarako bihurketa automatikoa egingo du.

Bestalde, ikasketa-prozesuan lortzen diren matrizeak aldatu nahi izanez gero, matrizeen datuak testu-fitxategian ematen ditu, eta aldaketak egin ondoren matrizeak konpilatzeko programa bat ere eskaintzen da. Matrizeak zuzenean aldatu nahi ez badira, baina datuak zenbait testuingurutan ukitu nahi badira, erregela (*bias*) batzuen bitartez trantsizioen pisuak alda daitezke. Ideia hau XT etiketatzailean ere agertzen da.

Behin ikasketa amaitzen denean, eredu hori erabilita testuak etiketatuko dira. Jatorrizko testua eskuz desanbiguatuta badago, emaitzak ebaluatzeko, testuaren etiketatzean eginiko erroreen eta haien testuinguruen estatistikak ematen dituen programa ere eskaintzen da. Honela testu osoa begiratu beharrik ez da izango, erroreak sortarazten dituzten testuinguruak aztertu besterik ez baita egin beharko.

VI.1.2.3 Brill-en etiketatzailea

Brill-ek (1992, 1995) erregeletan oinarritutako etiketatzailea garatu du. Etiketatzaile honek estokastikoen antzeko emaitzak ematen ditu. Erregeletan oinarritutako etiketatzaile gehienak ez bezala, Brill-ek garatutakoa sendoa da eta erregelak automatikoki erauzten ditu. Hurbilpen honen ezaugarririk garrantzitsuena erabilitako memoria da, eredu estokastikoek erabiltzen dutena baino gutxiago behar baitu. Eredu estokastikoetan datu-taula itzelak erabiltzen diren bitartean, etiketatzaile honek erregela kopurua oso txikia erabiltzen du, aurreprozesu luzea duen arren. Saiakuntzetarako *Brown Corpus*a (Kucera eta Francis 1967) erabili da.

Etiketatzaile honek okerrak diren etiketak ezabatzen joan ordez, edo egokia dena aukeratu eta bere horretan utzi ordez, beti ere etiketa bakarra aukeratzen duten erregelak erabiltzen

dira. Ondoren, etiketa hori zuzena ez denean, beste batekin ordezkatzen da. Hortik datorkio metodoari izena (*transformation-based error-driven*). Etiketatzeko hiru modulu erabiltzen ditu:

- Etiketatze lexikala. Modulu honek hitz bakoitzari testuingurua kontuan hartu gabeko etiketa probableena egokitzen dio. Datu horiek corpusetik lortutako lexiko batean gordeta daude.
- Hitz ezezagunen etiketatzea. Aurreko pausoan etiketatu ez diren hitzei etiketa egokitzen die. Honela, maiuskulaz idatzitakoak izen bereziak kontsideratzen dira, eta gainerako hitzak bukaerako hiru letren arabera corpuseko etiketarik probableenarekin etiketatzen ditu.
- Testuinguru-erregelak. Erregela hauek aurreko moduluetan lortutako etiketa-sekuentziei aplikatzen zaizkie, testuinguruaren arabera aurreko pausoetan egokitutako etiketa aldatzeko.

Etiketatze lexikala egiteko ikasketarako corpus bat erabiltzen da, bertatik hitz bakoitzaren etiketarik probableena lortuz eta hiztegi bat osatuz. Pauso honetan ez da testuingurua kontuan hartuko. Ikasketarako *Brown Corpus*aren %90a erabili da, beste %5 erregelak erauzteko eta gainerako %5a probetarako.

Erregelak automatikoki lortzeko ondoko prozesua jarraitzen da:

- Erregeletarako corpus-zatia etiketatzen da. Horretarako ikasketarako corpusetik hitzei dagokien maiztasun handieneko etiketa lortzen da, testuinguruari buruzko informazioa erabili gabe.
- Errore bakoitzari dagozkion datuak lortzen dira. Aurreko pausoan lortutako etiketak zuzenekin konparatuz egingo da. Datu horiek hirukote baten bidez adierazten dira; hirukoteak <*etk1,etk2,zenbat*> modukoak dira, non lehen etiketa etiketatzaileak egokitutakoa izango den, bigarrena, berriz, etiketa zuzena, eta azkena errorea zenbatetan gertatu den.
- Errore bakoitzeko erregela lortu. Erregela hauek lortzeko, errorea gertatu den adibide guztiak aztertu eta testuinguruari begira errorea zuzentzeko erregela idatziko da. Erregela hauek *etk1 etk2* etiketen arteko bihurketa zehazten dute.
- Lortutako erregelak corpusari aplikatzen zaizkio, errore-tasa berriz neurtuz. Errore gehien zuzentzen duena testuinguru-erregelen multzoan sartzen da. Gehienetan, ondo zeuden zenbait etiketa aldatzen dira, baina erregela-multzoan sartzeko errore-tasaren jaitsiera aurreko etiketatzean lortutakoarekin alderatzen da. Honela, zuzendutakoak okertutakoak baino gutxiago direnean, erregela baztertu egingo litzateke.

Prozesu honetan jarraituko da errore-tasa gutxi daitekeen bitartean. Erregela bat lortzen denean corpusari aplikatzen zaio erregela gehiago bilatzen jarraitu aurretik.

Behin erregela-multzoa erauzia, edozein testu etiketatzeko prest dago sistema. Horretarako, etiketatzaile lexikaletik pasako da eta, ondoren, testuinguru-erregelak aplikatzen dira.

Etiketatzaile honek doitasun ona du, %95 ingurukoa, edozein etiketatzaile estokastikok baino datu gutxiago erabilita ere. Gainera, erregelak automatikoki lortu arren, haien irakurketa linguistikoa ulergarria izan daiteke, batzuetan esanguratsua ez izan arren. Hiru saiakuntzen ondorioz 300 erregela inguru lortu ziren. Bestalde, erabat garraiagarria da, beste edozein eremutan aplikaturik ere sendoa izango da, ikasketa eta erregeletarako corpusetatik erauzten baitira datuak.

Hala ere, sistema motela zen, erregelak aplikatzeko algoritmoa ez zelako batere eraginkorra. Hau konpontzeko egoera finituko transduktoreen bidezko soluzioa proposatu zuten Roche eta Schabes-ek (1995), erregela-multzoa transduktore determinista bihurtzean, trantsizio bakoitzeko hitz bat tratatzen delarik. Eredu honen bidez etiketatzaile estokastikoak baino azkarragoa bihurtzen da, literaturan erregeletan oinarritutako hasierako sistemei egotzitako oztopo guztiak gaindituz.

Hala ere, Brill-ek sistema hobetzeari ekin zion. Erroreak aztertuz erregela lexikalak gehitzea proposatzen du, hau da, etiketak aldatzeko erregelen osagaiak inguruko hitzen etiketak ez ezik, inguruko hitzak edota hitza bera ere kontuan hartzea (Brill 1995). Gainera, hitz ezezagunen tratamendua hobetu du. Erregela berri hauek aplikatuz eta hitz ezezagunen etiketak zorrotzago egokituz, %97,5 inguruko doitasuna lortu du.

Laburbilduz, Brill-en etiketatzaileak erregelen bidezko sistemei egotzitako traba gehienak gainditu ditu, nahiz eta emaitzak hobetzeko ikasketarako corpus handia behar duen. Hala ere, corpusaren zati txikia eskuz desanbiguaturik egotea nahikoa da. Desanbiguatu gabe corpus handia erabil daiteke erregela-multzoa eraikitzeko eta, ondoren, corpus gainbegiratu txikia erregelak doitzeko erabili. Brill-ek dioenez, modu horretan emaitza onak lor daitezke eskuz lan txikia eginez (van Halteren (*ed.*) 1999).

VI.1.3 Metodoen konbinaketa bidezko etiketatzaileak

Orain arte aztertu diren sistemetan ezagumendu mota bakarra erabiltzen zen. Datu estatistikoak erabiltzen dutenek doitasun ona lortzen badute ere, ezagumendu linguistikoan oinarritutako azken sistemek balio horiek hobetzen dituzte. Badirudi datu estatistikoetan oinarriturik %96ko doitasunaren muga hori gaindi ezina dela.

Gauzak horrela, sistema estokastiko hutsak garatu beharrean, datu estatistikoei oinarrizko ezagumendu linguistikoa gehituz emaitzak hobetzen saiatu dira. Doitasun-maila hori gainditu nahian sistema konbinatuak sortu dira. Konbinaketa honen helburua anbiguotasuna gutxitu eta, honela, errore-tasa murriztea da. Horretarako metodo desberdinak erabili dira.

Metodoetako bat arestian aipatutako CLAWS1 sistemaren CLAWS4 bertsioan (Leech *et al*. 1994) erabilitakoa da. Sistema honen azken bertsioan anbiguotasun-tasa gutxitzeko zenbait hitz anitzeko unitate tratatzen dira. Honela, etiketatze estatistikoak sortutako akats batzuk zuzendu eta kasu askotan anbiguotasun-tasa jaistea ere lortzen da.

Beste metodo bat etiketatzaile estokastikoa eta linguistikoa konbinatzea da. Aukera hau logikoena dirudi baina ez da berehalakoa. Sistema bakoitzak etiketa-sistema desberdina erabil dezake, eta erabilitako informazioa desberdina izan ohi da. Hala ere, bide honetatik saiakuntzak egin dira.

Horietako bat Tapanainen eta Voutilainen-ek (1994) aurkeztu dute. Lan honetan arestian aipatutako EngCG eta XT etiketatzaile estokastikoaren emaitzak elkarturik desanbiguazioaren emaitzak hobetzen dira.

Izaera desberdineko informazioa ere konbina daiteke. (Màrquez eta Padró 1997) lanean, etiketatzaileak unigramak, trigramak, automatikoki erauzitako murriztapenak eta eskuz idatzitako murriztapenak konbinatzen dituzte etiketatze-prozesu bakarrean, emaitzen zuzentasuna hobetzearren. Beste hurbilpena batean (Màrquez *et al*. 1998) etiketatzaileen konbinaketa erabiltzen dute datu gutxi izanik etiketatzaile baten ikasketa-prozesua burutzeko (*boostrapping* teknika erabilita).

Baina aztertutako gainerako sistema hibridoetan, Tapanainen eta Voutilainen-en proposamenari jarraituz, hainbat etiketatzaileen emaitzak aztertzen dituzte desanbiguazioaren emaitza emateko. Horrela, paraleloan egindako bi lan badira, (van Halteren *et al*. 1998) eta (Brill eta Wu 1998), etiketatzaile desberdinak maila bateraino behintzat osagarriak direla ikusirik, emaitzak hobetzen saiatu direnak. Lehenengoak, Markov-en eredu ezkutuetan oinarritutako etiketatzailea, MBT (*Memory Based Tagger*), Brill-en etiketatzailea eta entropia maximoan oinarritutakoa erabiltzen ditu. Bigarrenak, aldiz, unigramen etiketatzailea, trigramena, Brill-ena eta entropia maximoan oinarritutakoa.

Guztien emaitzak konbinatzeko, oinarrizko bozketa (gehienek hautatu duten etiketa eman) edo neurri sofistikatuagoak erabiltzen dituzte. Azken emaitzetan, erreferentzia gisa entropia maximoan oinarritutako etiketatzailearen emaitzak hartu dira bi lanetan, etiketatzaileetan errore gutxien egiten duena delako. van Halterenen taldeak (%97,92ko zuzentasuna) erroreen %19 eta Brill-ek eta Wu-k (%97,2ko zuzentasuna) erroreen %10 ekiditea lortzen dute.

Ondorengo lan batean (Màrquez *et al.* 1999) iturburu desberdineko ereduak erabilita bi etiketatzaile proposatzen dituzte eta Brill eta Wu-ren (1998) etiketatzailearen emaitza berak lortzen dituzte informazio estatistiko hutsa erabiliz. Ondorioz, etorkizunari begira emaitza horiek beste etiketatzaile batzuenarekin konbinatuz egungo etiketatzaile hoberenen neurriak berdindu edo gainditzea posible dela diote.

Azkenik, (Ma *et al*. 1999-b) lanean, aipatutako sare neuronal elastikoen bidezko desanbiguatzailea (Ma *et al*. 1999-a) eta Brill-en etiketatzailea konbinatzen dira. Azken hurbilpen honek, (Tapanainen eta Voutilainen 1994), (van Halteren *et al*. 1998) eta (Brill eta Wu 1998) lanetan ez bezala, etiketatzaileak sekuentzialki aplikatzen ditu, lehenengoz sare neuronalak eta ondoren desanbiguazio-erregelak. Hitz anbiguoen desanbiguazioa %1,1 hobea dela diote (%95,5).

Guk planteatzen ditugun hobekuntzekin erlazio handiagoa duten bi konbinaketa zabaltzen dira ondoren, batetik, etiketatzearen akats batzuk zuzentzeko hitz anitzeko unitateen prozesamendua planteatzen duen CLAWS4 sistema eta, bestetik, murriztapen-gramatika eta Markov-en eredu ezkutuen arteko konbinaketa.

VI.1.3.1 CLAWS4 sistema

Esan bezala, CLAWS4 sisteman (Leech *et al.* 1994) hitz anitzeko unitateak markatzen dira anbiguotasuna eta errore-tasa murrizteko. Lan honetan hitz anitzeko unitateen multzoa definitzeko corpusetako maiztasun handieneko agerkidetzak neurtu dira. Azterketa horren ondorioz, HAUL ez diren zenbait patroi ere lortu dira. Horregatik, CLAWS4 sistemaren aurkezpenean egileek onartzen dute hasiera bateko IDIOMTAG izena ez dela oso egokia. Azken hurbilpenean erregelen bidezko etiketen egokitzapena dela esaten da ("*rule-driven contextual part-of-speech assignment*").

Zenbait erregelaren bitartez, hitz anitzeko unitatearen osagaiak elkarrekin etiketatzen dira, eta osagai bakoitzari sekuentzia-zenbaki bat egokitzen zaio. Beste zenbaitetan, berriz, osagai bakoitzari interpretazioa aldatzen zaio. Horrela, testuinguru zehatzetan interpretazio arraroak (*rare*) dituzten hitzei, hasieratik aukeran ematen ez zaizkion etiketa bereziak erregelen bidez egokituko dira, hitzaren hasierako anbiguotasuna kasu gutxi batzuengatik ez handitzeko.

Bestalde, testuinguru batzuetan interpretazioak ezabatzen dira. Inguruko etiketak ezagutu aurretik ezin dira baztertu, baina testuingurua ikusirik jakin daiteke zein interpretazio baztertu. Zenbait kasutan irakurketa bakarra geratuko da, baina ez beti.

Sistema honek etiketatzea lau pausotan egiten du. Hasiera batean lexikoko etiketak egokitzen zaizkie hitzei. Ondoren desanbiguazio estokastikoa egiten da, baina etiketa posibleen artean erabaki argi bat ezin denean hartu, ez da desanbiguatuko. Hurrengo pausoa

erregelak aplikatzea izango da. Esan bezala, prozesu honek zenbait kasutan ez du testua erabat desanbiguatuko. Horregatik, laugarren pausoan desanbiguazio estokastikoa aplikatuko da bigarrenez.

Lau pausoak aplikatu ondoren testuan anbiguotasunik gera daiteke, baina azkeneko bi pausoak emango ez balira, anbiguotasun-tasa askoz ere handiagoa izango litzateke.

Gure kasuan, hitz anitzeko unitate seguruak morfologiarekin batera tratatzea erabaki dugu, desanbiguazioan egin daitezkeen akats batzuk ekiditeko eta anbiguotasuna jaistearren. Hala ere, hitz anitzeko unitate ez seguruak modu honetan tratatzea irtenbide ona izan daitekeela uste dugu.

VI.1.3.2 EngCG eta XTren arteko konbinaketa

Hurbilpen honetan beste lanetan bakarka erabilitako bi metodo konbinatuz desanbiguazio-prozesuaren doitasuna hobetzen saiatu dira (Tapanainen eta Voutilainen 1994). Horretarako, arestian azaldutako EngCG eta XT konbinatu dituzte.

Bakoitza bere aldetik diseinatu denez, bi etiketatzaileak konbinatzerakoan zenbait arazo sortzen dira. Batetik, token-ezagutzaile desberdinak erabiltzen dituzte, baina orokorrean bi etiketatzaileen irteerak erraz lerroka daitezke. Arazoak zenbait tokenetan gertatzen dira. Adibidez, XTk *aren't* token bakarra kontsideratzen du eta EngCGk '*are not*' tokenetan banatzen du. Beste adibide bat hitz anitzeko adierazpenen etiketatzea da, EngCGk zenbait espresio token bakartzat jotzen dituen bitartean, XTk tokenetan banatzen ditu beti; '*in spite of*' espresioa, esate baterako, EngCGk osorik etiketatzen du eta XTk hiru token bereizten ditu. Hala ere, token-ezagutzaileen diferentziak ezagutzen direlarik, egokiena aukeratu eta gainerako hitzei ekitea da irtenbide zuzenena.

Bestetik, etiketa-sistema desberdinak erabiltzen dituzte, EngCGen erabilitako deskribapena askoz aberatsagoa da. XT etiketatzaileak *Brown Corpus*eko etiketa-sistema erabiltzen du. EngCGk, aldiz, EngTwol analizatzaile morfologikoak ematen duen informazioa erabiltzen du. Bi etiketa-multzoen arteko mapaketa egiteko, XTk EngCGk baino etiketa gutxiago duenez, modu errazena EngCG sistemaren etiketak *Brown Corpus*enetara bihurtzea da. Mapaketan gertatzen diren arazoak bi motakoak izan daitezke:

- Bereizketa-diferentziak. EngCGk interpretazio xeheagoa izan arren, *Brown Corpus*ean dauden bereizketa batzuk ez dira agertzen, edota alderantziz.
- Etiketa bera modu desberdinean erabiltzen da batzuetan.

Bi etiketatzaileak konbinatzeko prozedura honakoa da: testuari XT eta EngCG etiketatzaileak aplikatzen zaizkio independenteki; EngCGk interpretazio bat baino gehiago uzten duenean XTk proposatzen duen etiketa kontsultatzen da, eta XTk aukeratutako

interpretaziotik gertuen dagoen EngCGren interpretazioa hobesten da. Honela testu osoa desanbiguatzen da, hitz bakoitzeko interpretazio bakarra utzirik. Sistema honen emaitzak oso onak dira. Oraintxe esan dugun prozeduraz, aurretik ikusi gabeko 27.000 hitzeko corpus zatiari aplikatuta %98,5eko doitasuna lortu da.

Gure kasuan, murriztapen-gramatikak uzten duen anbiguotasuna oso handia da eta etiketatzaile estokastikoaren emaitza oso kaxkarra da ingeleserako emaitzekin alderatuta, VI.2 atalean ikusiko den bezala, beraz, metodoak paraleloan erabili ordez, bata bestearen atzetik aplikatzea proposatzen dugu, VI.3 atalean azaltzen den moduan.

VI.2 Euskararen desanbiguazioa metodo bakarrarekin

Testuak desanbiguatzeko metodoak azterturik, gure beharretara egokitu eta erabil ditzakegunen artean hautatu behar izan dugu. Hasteko, testuetan dagoen anbiguotasuna kontuan hartu behar da, interpretazio morfologiko osoa erabiltzea interesgarria den arren, horrek konplexutasun handia suposatzen du. Horregatik, aplikazio desberdinetarako egokiak izan daitezkeen etiketa-multzoak diseinatu dira, ondorioz, anbiguotasun-tasa, aukeratzen den etiketa-multzoaren araberakoa da eta, zehazkiago esateko, etiketa-multzoan erabiltzen den deskribapen linguistikoaren araberakoa.

Desanbiguazio-metodoa erabakitzeko, batetik, metodo estokastikoa erabiliz gero, programazio eta eskuzko lan ugari aurrez daitekeela ematen du, baina, bestetik, MORFEUSek ematen duen informazio guztia kudeatu nahi badugu, metodo linguistikoek zuzenago dirudite. Arlo honetan egindako azken lanen joerarekin bat etorriz, bi metodoak jorratu eta konbinatzeko aukera izan dugu.

Ezagumendu linguistikoan oinarritutako desanbiguaziorako Karlssonen *Constraint Grammar* formalismoa aukeratu da. Eta desanbiguazio estokastikorako, berriz, MULTEXT proiektuko tresnak erabiliko ditugu, ondoren zehaztuko diren arrazoiengatik. Ondorengo ataletan formalismo bakoitza modu independentean aplikatuta egindako saiakuntzak eta lortutako emaitzak aurkezten dira. Emaitza horiek, VI.3 atalean aurkezten den metodo-konbinaketa egiteko modua ulertzeko balio dute.

VI.2.1 Desanbiguazio linguistikoa: murriztapen-gramatika

Murriztapen-gramatikaren (MG) azken helburua analisi sintaktikoa egitea den arren, lehen urratsa desanbiguazio morfosintaktikoa da. Horretarako erregela-multzo bat definitzen da, fenomeno batzuk aztertzearen ondorioz zenbait interpretazio baztertzea helburu izanik.

Arestian EngCGren aurkezpenean azaldu denez, metodo honen bidez ingeleserako oso emaitza onak lortu dira. Euskararen kasuan, hasierako anbiguotasuna handiagoa denez, testuak erabat desanbiguatu nahi izanez gero, ingeleserako diseinatu direnak baino erregela gehiago beharko liratekeela ikusi da. Lan horretan diharduen hizkuntzalari-taldeak corpus-azterketan oinarrituta 1.000 erregela inguru garatu ditu. Horietatik 500 inguru erabili dira kapitulu honetan aurkezten diren emaitzak lortzeko, eta emaitzak onak dira, ondoren aurkezten den legez, baina irteerako anbiguotasun-tasa ingelesaren hasierakoaren parekoa da. MGk 4. etiketa-multzoa erabiltzen du, hau da, deskribapen morfosintaktiko osoan oinarritzen da desanbiguatzeko.

Desanbiguatzaile honen deskribapen sakona Aduriz-en tesian (2000) egiten da eta bertan, fenomeno desberdinen azterketa egin ondoren definitutako erregelak aurkezten dira. Metodologia eta ebaluazioaren inguruan informazio gehiago lortzeko honako argitalpen hauek ere kontsulta daitezke: (Aduriz *et al*. 1996-d) eta (Aduriz *et al*. 1997).

Erregelak diseinatzeko erreferentzia-corpuseko EEBSko 27.000 tokenetatik erdia bakarrik erabili da, beste erdia egiaztapenerako erabili izan dutelarik. Erregelen azken doiketa egiteko, berriz, erreferentzia-corpuseko *Euskaldunon Egunkaria*ko 9.000 inguru tokenak erabili dituzte.

Aipatzekoa da, bestalde, gramatika garatzeko tesi-lan honetan planteatzen diren hobekuntzetatik kontuan hartu diren bakarrak IV. kapituluko hitz ez-estandarren tratamenduari dagozkionak izan direla. Beraz, ez dira erregela konkretuak idatzi III. kapituluan aurkeztu den analizatzaile hedatuak gehitutako anbiguotasuna ebazteko.

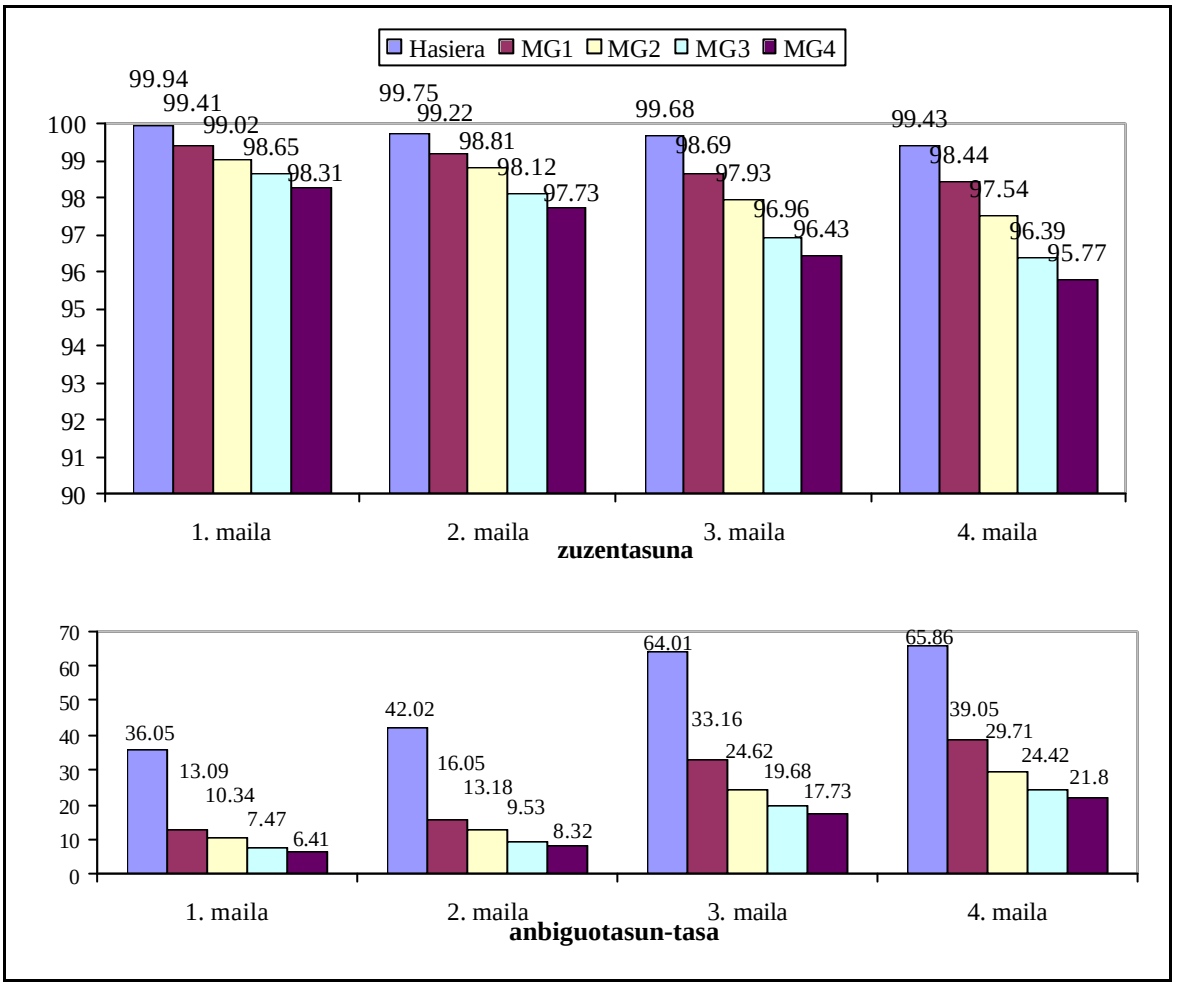
Hala ere, gramatika diseinatzerakoan erabilitako metodologia nahiko orokorra izan da, aplikazio zehatzei lotu gabekoa, eta, horregatik, III., IV. eta V. kapituluetako hobekuntzen gainean aplikatuta ere, emaitza egokiak lortzen dira. Hori dela eta, ondorengo atalean aurkezten diren emaitzak, aipatutako hobekuntza guztien gainean lortutakoak dira.

VI.2.1.1 Ebaluazioa

Esan bezala, gramatika lau azpimultzotan banaturik dago, ziurtasun-mailaren arabera antolatuta. VI.1 irudiko grafikoetan, egiaztapen-corpusaren gaineko desanbiguazioaren emaitzak agertzen dira, zuzentasunaren eta anbiguotasun-tasaren bilakaera hain zuzen ere[12]. Desanbiguazio-tasari dagokionean, 4. mailan %57-58, %71-72, %78-79 eta %81-82 ingurukoa da MG1, MG2, MG3 eta MG4 erabilita, hurrenez hurren. Prozesuaren errore-tasa nahiko txikia da, %0,5-0,7 artekoa MG1ena, %1,25-1,4 MG2rena, %1,75-2 MG3rena eta %2-2,4koa MG4ren kasuan.

Etiketa-multzoaren araberako emaitzak aztertuz gero, argi ikusten da erregela gehienak kategoria eta azpikategoriaren desanbiguazioari begira diseinaturik daudela. VI.1 irudiko grafikoetan 1. eta 2. mailako desanbiguazioa askoz ere nabarmenagoa dela ikus daiteke, eta 3. mailako anbiguotasuna 2. mailakoaren bikoitza da gutxi gorabehera.

[12] Erreferentzia- eta egiaztapen-corpusetan lortutako emaitza osoak C eranskinean ematen dira, C.4 atalean, zuzentasuna eta anbiguotasun-tasaz gain, batezbesteko analisi kopurua emanik.



VI.1 irudia.- MG aplikatu aurretik eta ondorengo emaitzak etiketatze-mailaren arabera.

VI.2.2 Desanbiguazio estokastikoa: *MULTEXT*en aplikazioa

Metodo estatistikoei buruz aritu garenean hauetan oinarritzen diren zenbait sistema aipatu ditugu eta, esan bezala, sistema asko publikoak dira. Horrek sistema guztien artean gure beharretara gehien hurbiltzen zena lortzeko aukera eman zigun. Sistema batzuk baztertu behar izan genituen lexikoaren dependentzia zuzena izanik, hitz-formen maiztasunak behar izateagatik. Dependentzia hori saihesteko etiketen maiztasunak erabiltzen zituzten sistemen artean aukeratu behar izan genuen. Gainera erabilera publikoa izatea ere ezinbestekotzat jo genuen. Sistema horien guztien artean arestian azaldu dugun *MULTEXT* sistema hautatu genuen.

Tresna-multzo hau aukeratzean ondokoak hartu genituen kontuan:

- Analisi morfologikorako modulu baten ondoren aplikatzeko pentsatuta dago.
- XT etiketatzaileak bezala, anbiguotasun-klaseak erabiltzen ditu, lexikoarekiko independentzia lortzearren.

- Etiketa-sistema lan handirik gabe alda daiteke, sarrerako testuan aldaketarik egin gabe.
- Testuak ikasketa zein etiketatzerako prestatzeko tresna lagungarriak eskaintzen ditu.
- Emaitzen azterketa estatistikoa egiteko tresnak eskaintzen ditu, erroreak gertatu diren testuinguruan kokatuz.
- Matrizeetako emaitzak ukitzeko aukera ematen du, bai matrizeen gainean zuzenean aldatuz, bai erregela lokalen bidez pisuak aldatuz.

Testua prestatzeko tresnak anbiguotasun-klaseen multzoa automatikoki erauzten du, eta ikasketarako corpusean ez dauden anbiguotasun-klaseak agertzen badira, automatikoki gehitu dakizkioke multzoari.

Etiketa berriak gehitzea ere zilegi da, automatikoki detektatu eta gehitzen direlako hala nahi izanez gero. Ezaugarri hau interesgarria da bereziki hirugarren mailan etiketatu nahi denean. Izan ere, ikasketarako corpusaren tamaina dela eta, hirugarren mailako etiketa posible guztiak ez dira agertzen, ondorioz, testu berriak tratatzen direnean askotan etiketa berririk agertuko da. Hori dela eta, etiketa berri horien agerpen gutxi batzuen gainean ikasi behar izaten du sistemak, kasu askotan agerpen bakarraren gainean, eta, beraz, testuinguru posible guztien berri izan ezik, probabilitateen pisuak ez dira esanguratsuak izango kasu horietan.

Dena dela, arazo hau orokorra da etiketatzaile estatistikoetan. Matrizeak eraikitzean, etiketa-bikote posible guztien probabilitateak kalkulatzen dira, baina testuinguru horietako askorentzat probabilitate minimoa esleitzen da. Etiketa kopurua handitzen den heinean, probabilitate minimo horrekin betetako posizio gero eta gehiago agertzen dira matrizean, benetako testuinguruen artean banatzen den probabilitatea geroa eta txikiagoa izanik. Arazo honi ingelesez *data sparseness* deitzen zaio, hau da, datu-sakabanaketa.

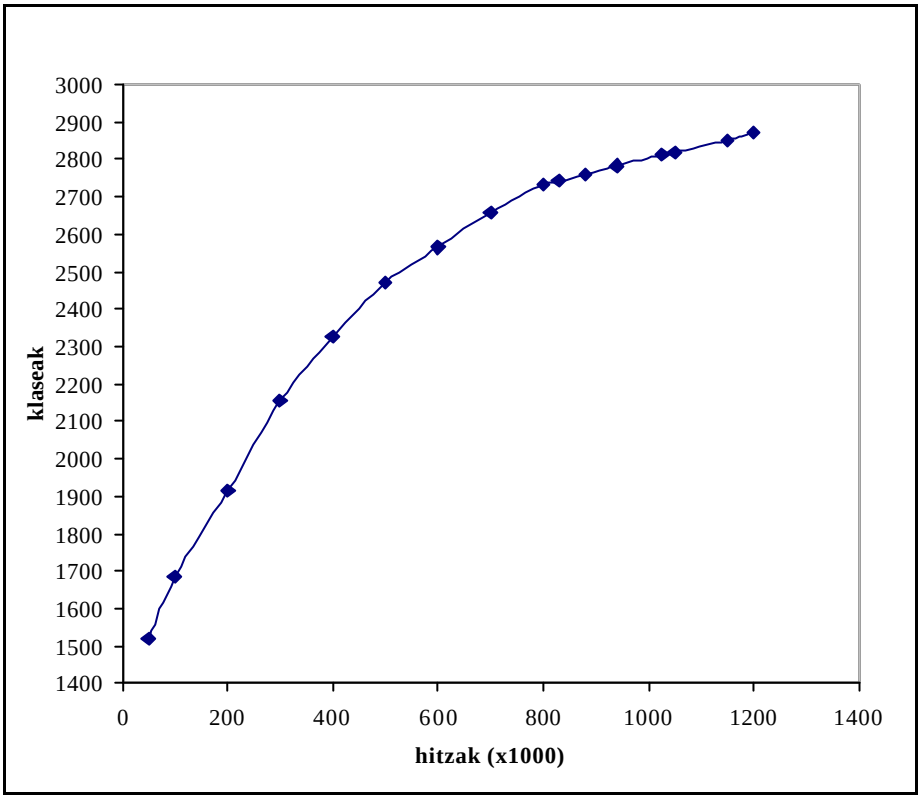
Horrelako arazoak zein beste batzuk konpontzeko erregela lokalak (*bias*) idatz daitezke. Erregela hauen bidez testuinguru jakin batzuetan ikasitako probabilitateak alda daitezke, ikasketa-testuan gutxitan agertu diren etiketen pisuak indartzeko edota indar handia dutenenak ahultzeko. Baina hau ez da oso irtenbide txukuna etiketa kopurua oso handia denean.

Azkenik, erabilerak erakutsi digu MULTEXTen tresnek hainbat alde txar ere badituztela, besteak beste, honakoak aipatu nahi genituzke:

- Inplementazioa ez da datuetatik independentea, etiketa-kopuru maximoa edo bestelako konstante bat aldatuz gero, programa birkonpilatzea ez da nahikoa, matrizeak berriz ere sortu behar dira.
- Ikasketa-corpusak etiketa zilegi guztien eta anbiguotasun-klase guztien agerpenak —gutxienez bat— izan behar ditu.

Bigarren hau da, izatez, arazorik larriena. Izan ere, klase edo etiketa berriak agertzean, matrizeen pisuak birkalkulatu behar dira, ikasketa ez-gainbegiratuaren bitartez, eta horrek desanbiguazioaren kalitatea kaskar dezake.

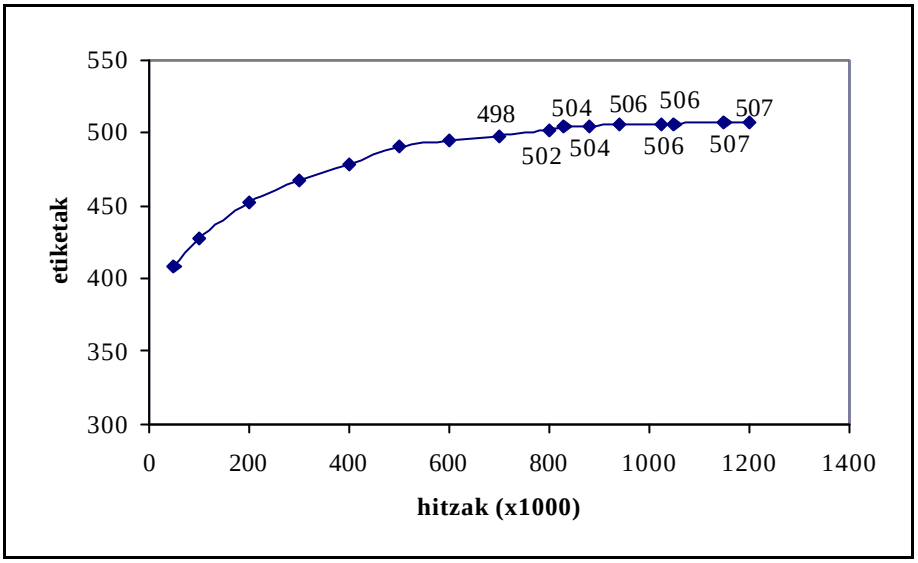
Beste sistema askotan biguntze-teknikak (*smoothing*) erabiltzen dira aldez aurretik agertu ez diren klase eta etiketa berriak agertzen direnerako probabilitatearen proportzio txiki bat gordeaz. Baina MULTEXTeko tresnetan ez da horrelakorik inplementatu, eta gure kasuan, hori desabantaila nabarmena da, gehienbat 3. mailako desanbiguazioa burutu nahi denean.



VI.2 irudia.- 3. mailako etiketa-multzoa erabilita agertzen diren anbiguotasun-klaseak.

VI.2 irudian EEBS (800.000) eta *Euskaldunon Egunkaria*ko (400.000) 1.200.000 inguru hitzeko corpusean agertutako anbiguotasun-klaseen irudia aurkezten da. Ikasketarako corpusean 1000 anbiguotasun-klase inguru agertzen diren bitartean, corpus mardulagoa tratatzean ikusten da kopuru hori ia hirukoiztu dela eta hazkundea pixkanaka jaisten doan arren, ez da erabat egonkortzen.

Etiketa kopuruari dagokionean, VI.3 irudian ikus daitekeenez, hazkundea ia egonkortu dela dirudi, baina etiketen zerrenda aztertzean ikusi da badirela oraindik ere testuetan ager daitezkeen etiketa zilegiak, zenbait kasu-flexio ez baitira oso maiz agertzen, baina edozein kasutan ez dira konbinaketa askoz gehiago agertuko. Hortaz, biguntze-teknikaren bat erabiliz gero, ager daitezkeen etiketa berri horien probabilitateak ikasketa berririk egin gabe kudea daitezke. Dena dela, ikasketa-corpusean etiketa horietatik 400 eskas agertzen dira soilik.

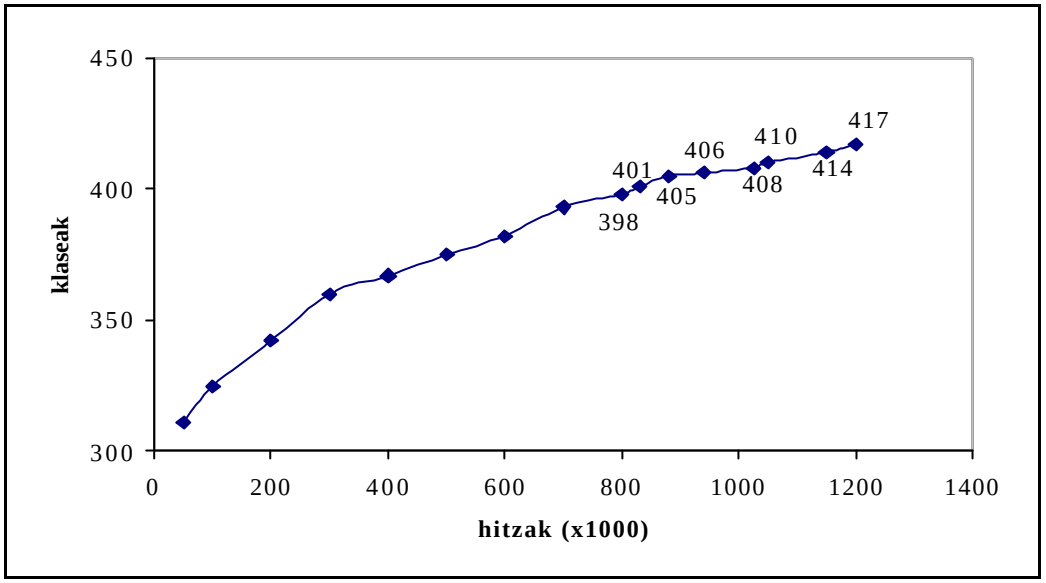


VI.3 irudia.- 3. mailako etiketa-multzoa erabilita agertzen diren etiketak.

Bigarren mailako etiketatzeak ez du horrenbesteko arazorik ematen. Etiketei dagokienean, zilegi da corpus ez oso handian etiketa guztiak agertzea, ikasketarako corpusa kasu, baina anbiguotasun-klaseei buruz ezin da horrelakorik esan. VI.4 irudian ikusten denez, corpus handituz batera, klase kopurua ere handitzen doa, baina ez 3. mailakoan bezain azkar.

Hala ere, ikasketarako corpusean 300 klase inguru besterik ez dira agertzen, eta, irudian ikus daitekeenez, corpus handiagoa hartuz gero, 400etik gora aurki daitezke. Hortaz, biguntze-teknikaren bat aplikatzea komeniko litzateke, baina kasu honetan ez dago hainbesterainoko diferentziarik, eta, ikasketa ez-gainbegiratua aplikatuz gero, ia klase guztien agerpenak lor daitezke, biguntze-teknikarik eza arazo txikiagoa bihurtuz 2. mailako desanbiguazioan.

Arazo hauen guztien ebazpena etorkizunerako lehentasunezko lan gisa planteatu da. Izan ere, metodoen konbinaketaren ondorioz arazoa murrizten den arren, oraindik ere klase eta etiketa berrien agerpenak hizkuntz ereduaren birkalkulatzea ekartzen baitu.



VI.4 irudia.- 2. mailako etiketa-multzoa erabilita agertzen diren anbiguotasun-klaseak.

VI.2.2.1 Ikasketarako corpusaren tamaina aukeratzen

Ikasketa-prozesuan erabili beharreko corpusa aukeratzea eta komeni den tamaina erabakitzea ez da gauza erraza. Beste hizkuntzekin egindako saiakuntzak aztertzean denetatik aurki daiteke. Kasu batzuetan eskuz desanbiguatutako corpus itzela erabiltzen dute (milioi bat inguru hitzekoa) eta beste kasu batzuetan mila edo bi mila hitz eskuz desanbiguatuta nahikotzat jotzen da.

Gure ustez, ahalik eta corpus desanbiguatu handiena erabiltzea interesgarriena litzateke — Merialdoren ildo beretik (1994)–, horrelako testuetan esaldi-mota guztiak eta anbiguotasun-klase guztien agerpen nahikoa aurki daitezen. Baina guk ezin dugu milioi bat hitzeko corpus bat eskuz markatu, lan neketsu eta garestia delako erdi-automatikoki etiketa daitekeen arren. Beste hizkuntza batzuetan, bereziki ingelesez, markatutako corpus publikoak direla medio, ez dute horrelako arazorik.

Beste zenbait kasutan, eskuz desanbiguatutako zati txikitxo bat lortu eta hasierako matrizeen pisuak kalkulatzeko erabiltzen dituzte. Ondoren corpus anbiguo handi bat erabiltzen dute ikasketarako metodo ez-gainbegiratuen bitartez.

Honekin zera adierazi nahi dugu, ez dagoela corpusa aukeratzeko nolabaiteko neurri estandarrik, bakoitzak eskutan duenari ahalik eta etekin handiena ateratzen saiatzen da.

Gogora dezagun ikasketarako erabili den erreferentzia-corpusak EEBSko 27.000 inguru tokenek eta *Euskaldunon Egunkaria*ko 9.000 inguru tokenek osatzen dutela, eta ebaluaziorako edo egiaztapenerako, berriz, EEBSko 1.300 token inguru eta *Euskaldunon*

*Egunkaria*ko 5.800 token inguruko corpusa erabiliko dela. Ikasketa ez-gainbegiratuaren saiakuntzak burutzeko, arestian aipatutako 1.200.000 hitzetako corpusa erabili da.

Erreferentzia- eta egiaztapen-corpusetako banaketak ez dira proportzionalak. Izan ere, hasiera batean genuen testu-masa EEBStik hartutakoa zen eta testu biak corpus orekatu horretakoak ziren. Dena dela, eskuz desanbiguatutako corpusaren tamaina oso txikia zen eta zenbait proiektu burutzeko interesgarri jo genuen *Euskaldunon Egunkaria*tik lortutako testuekin mamitzea. Baina egiaztapenerako testua 1.300 tokenez osaturik zegoenez, handiagoa egitea ere komeni zen eta, II. kapituluan aipatu den bezala, izen berezien tratamendua landu eta modu egokian ebaluatu ahal izateko egunkarietako testuak soilik gehitzea erabaki genuen. Horrek emaitzetan eragin negatiboa izan dezakeen arren, aplikazioei begira ebaluazio sendoagoa egiten lagunduko du.

VI.2.2.2 Saiakuntzak

Hasierako saiakuntzak (Ezeiza 1997) lanean aurki daitezke. Bertan, 1. eta 2. mailako etiketa-multzoekin egindako saiakuntzak deskribatzen dira, ikasketa mota biak erabilita eta emaitzak hobetzeko erregelak edo *bias*-ak erabili zirelarik. Ordutik hona, deskribapen morfologikoa eguneratu da, eta, horrekin batera, anbiguotasuna handitu ere. Dena dela, metodo estatistiko hutsak erabiltzean emaitza onak lortzeko esperantza handirik ez genuen hasieratik, eta saiakuntza hauek hobetu beharreko maila minimoa zein den neurtzeko balio izan dute.

Ondoren aurkezten diren emaitzak lortzeko hiru saiakuntza-multzo burutu dira. Lehenengoan, ikasketarako corpusaren EEBSko atala soilik erabili da, egiaztapenerako corpus ere EEBSko zatira mugatuz. Bigarren multzoan, berriz, corpus osoak erabili dira, bai ikasketarako baita egiaztapenerako ere. Azkenik, bai EEBS (800.000) bai *Euskaldunon Egunkaria*ko (400.000) testuak gehitu dira ikasketa-corpusean, anbiguotasun-klase eta etiketa gehiagoren isla izatearren. Lehenengo bi multzoetan ikasketa gainbegiratua egin da eta hirugarrenean, erreferentzia-corpusari dagokion atalarekin ikasketa gainbegiratua eta gainerakoarekin, interpretazio zuzena markaturik ez duenez, ez-gainbegiratua.

Aipatu, lan honetan ez ditugula *bias*-ak aplikatu, erregelak idazteko murriztapen-gramatikaren bidezkoa modu egokiagoa iruditzen zaigulako eta, gainera, helburua metodo biak konbinatzea delako.

VI.2.2.3 Emaitzak

Lehenengo eta bigarren saiakuntza-multzoen emaitzak VI.5 irudian azaltzen dira. Maila bakoitzeko bi emaitza ematen dira, testua erabat desanbiguatuta lortutakoa (*recall* =

precision) eta bi etiketen probabilitateak oso hurbil daudenean, hitza anbiguo utzita lortutakoa (*recall* > *precision*). Horrela, 1. mailako emaitzetan batezbesteko analisi kopurua 1,1ekoa da, 2. mailakoetan 1,09koa eta 3. mailakoetan, berriz, 1,13koa (1,08 bigarren multzoan).

Bigarren saiakuntza-multzoan, esan bezala, ikasketarako corpus osoa erabili da eta ebaluazioa, aurrekoarekin konparatu ahal izateko, bi zatitan ematen da: lehenengoan, egiaztapen-corpuseko EEBS zatiaren gaineko emaitzak, eta, bigarrenean, *Euskaldunon Egunkaria*ko zatiaren gainekoa. Azkenik, egiaztapen-corpus osoaren emaitzak ere ematen dira. Kasu hauetan, anbiguotasuna uztean, 3. mailako emaitzetan 1,08 analisi uzten dira, gainerakoak, aurretik aipatutako berberak izanik.

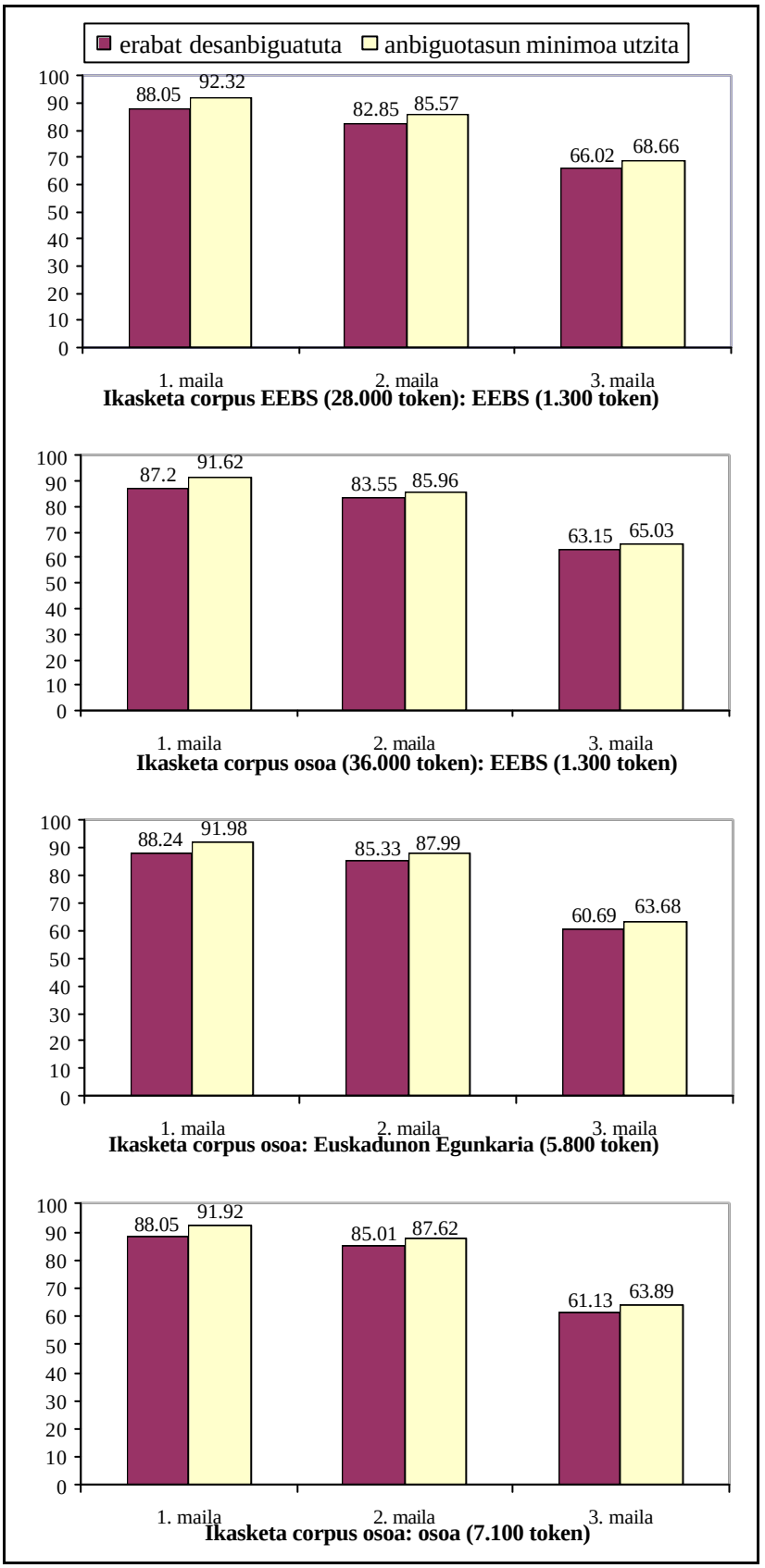
Ikus daitekeenez, lehenengo saiakuntzen emaitzak EEBS zatian hobeak dira bigarren multzokoak baino, 2. mailakoak salbu. Aldiz, bigarren multzoko saiakuntzetan, *Euskaldunon Egunkaria*ko zatiak EEBSkoak baino emaitza hobeak ditu, 3. mailan ezik.

Batezbesteko emaitzak aztertuz gero, lehenengo mailako desanbiguazioan lortutako emaitzak EEBS ikasketa-corpusa soilik erabilitakoen parekoak direla ikusten da. Bigarren mailari dagokionean, 2 puntuko hobekuntza lortzen da, beraz, corpusaren tamaina handitzeak laguntzen du. Azkenik, 3. mailakoetan EEBSko emaitzetatik atzera pausoa ematen da, gehienbat izen berezietan sortzen diren anbiguotasunak direla medio. Esate baterako, pertsona-izena genitiboan/leku-izena inesiboan bikotea askotan gertatzen da (*Kurt-en, Baudot-en, Altdorfer-en*...), baina EEBSko zatian gehienetan pertsona-izenak dira eta hobeto etiketatzen ditu. *Euskaldunon Egunkaria*ko zatian, aldiz, leku-izen eta pertsona-izenen arteko oreka handiago izanik, etiketen probabilitateak ere askoz hurbilago daude, eta, ondorioz, nahasketa gehiago emaitzan.

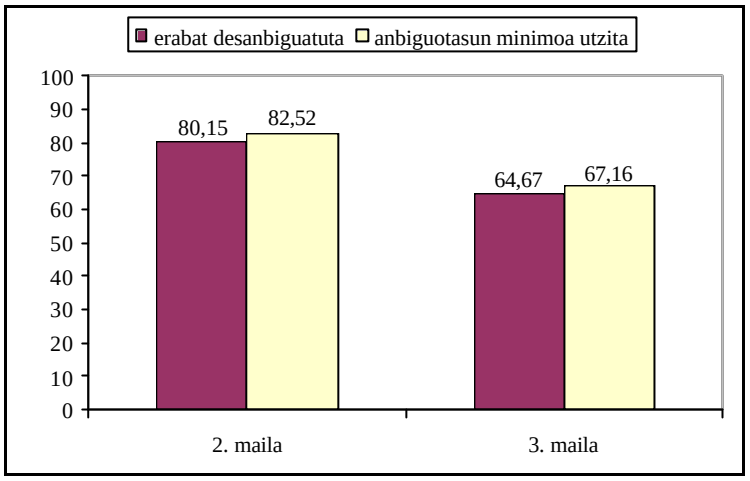
Ikus daitekeenez, kasu batean zein bestean, emaitzak nahiko eskasak dira, ia %12, %15 eta %40ko errorea egiten delako. Dena dela, emaitza hauek ulertzeko bi faktore hartu behar dira kontuan:

- hasierako anbiguotasuna oso altua dela, tokeneko 1,49 analisi 1. mailan, 1,56 2.ean eta 2,28 3.ean baitaude.
- ikasketarako corpusaren tamaina nahiko txikia dela, gehienbat 3. mailako etiketatzeari dagokionean. Gainera, arestian azaldu bezala, corpusean ez dira anbiguotasun-klase guztiak azaltzen, ezta egiaztapenerako corpuseko guztiak ere.

Beraz, emaitzak hobetzeko bi bide jarrai daitezke: batetik, hasierako anbiguotasun hori jaistea, ondoren aurkezten den metodoen konbinaketaren bitartez, esate baterako, eta, bestetik, ikasketarako corpus handiagoa erabiliz.



VI.5 irudia.- Ikasketa gainbegiratuaren ebaluazioa.



VI.6 irudia.- Ikasketa ez-gainbegiratuaren ebaluazioa.

Ikasketa ez-gainbegiraturi dagokionean, aurretik aipatutako 1.200.000 inguru hitzeko corpusarekin 2. eta 3. mailako desanbiguaziorako ikasketa ez-gainbegiratu egin da, eta erreferentzia-corpusarekin gainbegiratua burutu da eta, emaitzak —VI.6 irudia— aztertzean honako ondorioak atera daitezke:

- 2. mailako emaitzetan zuzentasunak okerrera egiten du, %80 inguruko zuzentasunera degradatuz. Anbiguotasun minimoa utzita, berriz, %82,5 ingurura iristen da.
- 3. mailako emaitzetan hobekuntza gertatzen da, logikoa denez, erreferentzia-corpusean oso anbiguotasun-klase gutxi agertzen zirelako, baita klaseen agerpen gutxi ere. Zuzentasuna %64,7ra igotzen da, anbiguotasuna utzita %67,3 ingurukoa lortuz.
- Corpus tamaina handitzen den heinean hobekuntza gertatzen da, 900.000 eta milioi bat hitz ingurura bitartean, baina hortik aurrera oszilazio txikiak gertatzen badira ere, emaitzak egonkortzen dira.

Pentsa daiteke ikasketa gainbegiratuaren proportzioa handituz gero, ikasketa gainbegiratu eta ez-gainbegiratu honen emaitzak hobe daitezkeela neurri batean, baina eskuz desanbiguatutako corpus handia erabiltzen ez bada, ez dirudi hobekuntza handiegia lor daitekeenik:

> *"A large untagged text and small tagged text can be combined by using the unsupervised algorithm to adjust the parameters in order to increase the probability the model assigns to the untagged corpus. In practice, this method of combining the two information sources has not proven to be very effective. Merialdo (1994) shows that in such a situation simply training on the small tagged corpus*[13] *typically gives better performance than can be achieved by combining this with unsupervised learning." (van Halteren (eds.) 1999:261-262).*

[13] Merialdok erabilitako ikasketa-corpusa txikia dela badiote ere, 450.000 hitzetakoa da. Ingeleserako, agian, txikia kontsidera dezakete, baina gure baliabideetatik oso urrun dago.

VI.3 Euskararen desanbiguazioa metodoen konbinaketarekin

Garatutako bi metodoen emaitzak aztertuta biak konbinatzeko aukera bakarra ikusten da, hau da, lehenengo murriztapen-gramatika aplikatzea, honela, errore gutxi eginda anbiguotasuna dezente mugatzen delako; eta, ondoren, Markov-en eredu ezkutuak aplikatzea geratzen den anbiguotasuna ebazteko.

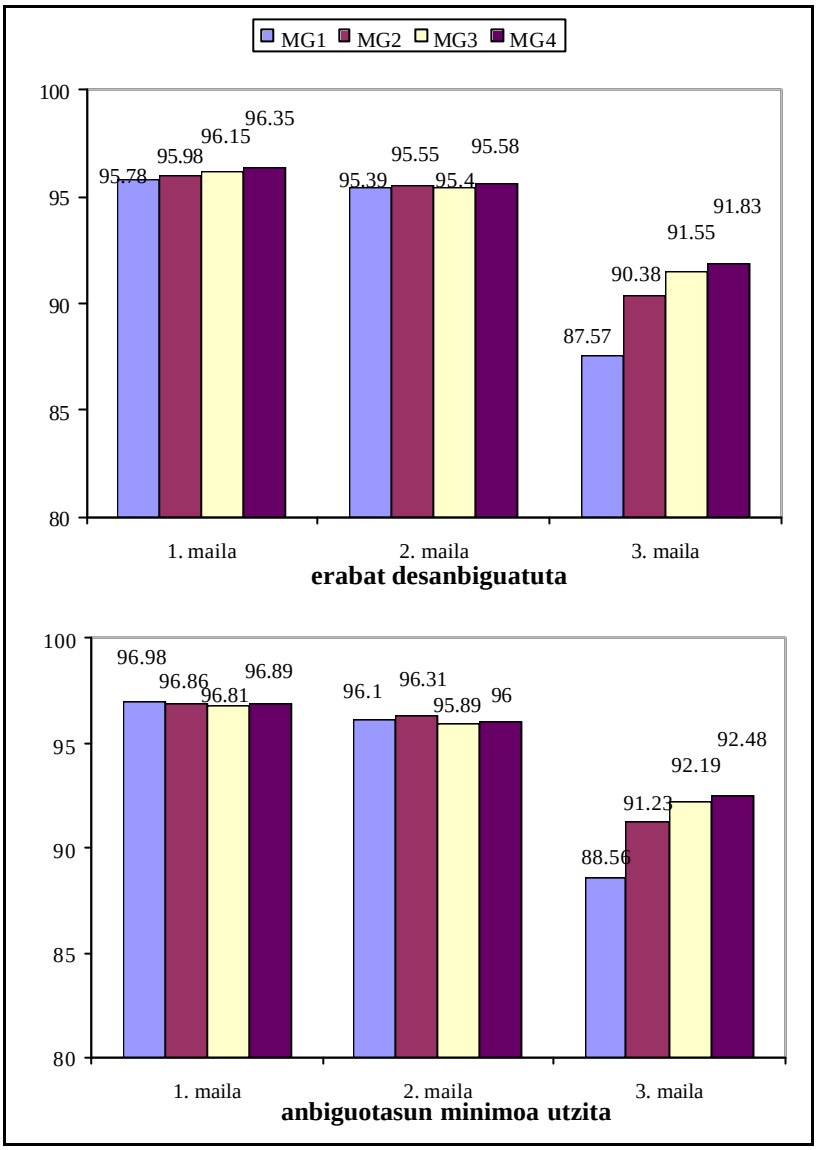
Murriztapen-gramatikari dagokion atalean aipatu da gramatikak lau azpigramatika dituela, lehenengo biak kategoria eta azpikategoria desanbiguatzeko balio dute eta beste bietan kasua, perpaus-mota, mugatasuna eta bestelako informazioa desanbiguatzeko balio dute. Azpimultzoak gehitzen diren heinean errore ere handitzen doa, baina etiketatze-mailaren arabera egokiena dena aukeratu behar da. Horretarako, saiakuntzak burutu dira eta kasu bakoitzean EUSLEMen erabili beharreko azpigramatika-multzo egokiena zein den azaltzen da ondoren.

VI.3.1 Saiakuntzak

EUSLEMen ebaluazioari begira 2. mailako desanbiguazioa interesgarriena den arren, maila guztietako desanbiguazioaren ebaluazioa ere burutu da, eta atal honetan jarraian aztertzen dira desanbiguazioa burutzeko aukera desberdinak.

VI.7 irudian MG1, MG2, MG3 eta MG4 aplikatu ondoren, erreferentzia-corpusaren gainean ikasketa gainbegiratua burutu eta corpus bera etiketatuz lortutako emaitzak aurkezten dira, lortutako emaitza guztiz desanbiguatutakoa da lehen grafikoan, eta tokeneko 1,02 analisi utzita bigarrenean. Desanbiguazio estokastiko hutsa erabilita baino askoz emaitza hobeak lortzen direla argi ikusten da.

Aurreko atalean azaldu bezala, MULTEXTek 3. mailan zuzenean aplikatuta %34ko errorea egiten du. Aldiz, MG1 aplikatu ondoren %11,5ean, MG2 ondoren %8ean, MG3 ondoren %6an eta MG4 ondoren %5,4an soilik huts egiten du. Errore honi aurreko prozesu guztiek egindakoa gehituta ere, desanbiguazio estokastiko hutsak baino kalitate handiagoa duen desanbiguazioa lortzen da.



VI.7 irudia.- Ebaluazioa ikasketa-corpusean.

Azken finean, 3. mailako anbiguotasuna handiegia da ikasketarako hain corpus txikia erabilita emaitza onargarriak lortzeko. Horregatik, nahiz eta murriztapen-gramatikak gero eta errore gehiago egin, desanbiguazioaren emaitzak hobetzen jarraitzen dute. Dena dela, gure ustez ikasketarako corpusa askoz ere handiagoa beharko luke anbiguotasun-klase eta etiketa guztiak —edo gehien-gehienak behintzat— ager daitezen. Gainera, gramatikaren 3. eta 4. azpimultzoetan definitutako erregelek analisi asko baztertu arren, errore gehiegi egiten dituzte eta ebaluazio sakonagoa eginez etorkizunean hobetzea espero dugu.

Bigarren mailako saiakuntzei dagokienean, murriztapen-gramatikaren emaitzei erreparatuta[14], ikus daiteke MG1 eta MG2 artean anbiguotasuna %3 jaisteko %0,5eko errorea

[14] Murriztapen-gramatikari buruzko atalean agertzen ez diren arren, erreferentzia- eta ikasketa-corpusari dagozkion emaitzak C eranskinean aurki daitezke.

gehitzen da, eta MG2 eta MG3 artean beste %3 jaisteko %0,7 gehitzen da. Errorea proportzionalki igotzen bada ere, kontuan hartu behar da beste hizkuntzetako emaitzak aztertuta, desanbiguazio estokastikoak %2,5-3,5ko errorea gehitu dezakeela. Beraz, MG2 edo MG3 aplikatu ondoren espero daitezkeen emaitzak %94,5-96,5 ingurukoak izango dira gehienez ere.

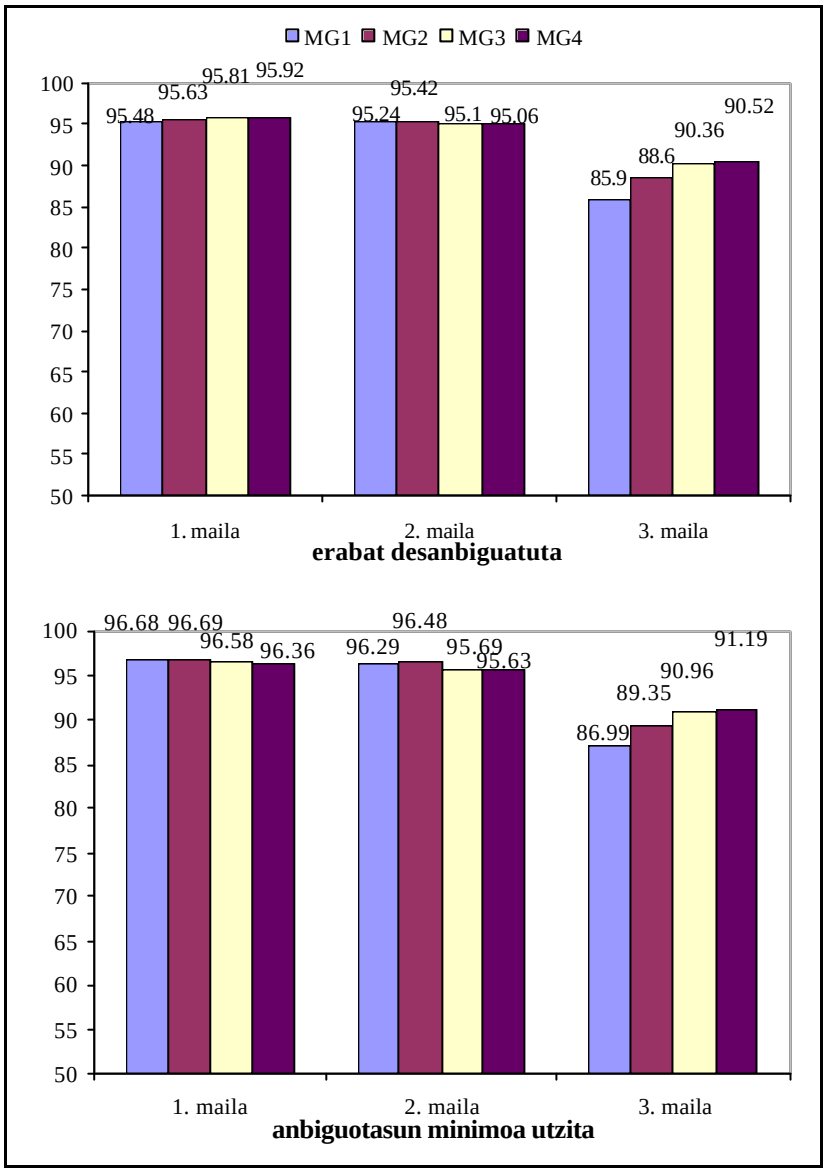
VI.7 irudian ematen diren zuzentasun-neurriak tarte horretan daude. Guztiz desanbiguatutako emaitzetan hoberena MG4ren ondorengoa da, MULTEXTeko tresnek %2,2ko errorea besterik egiten ez dutelarik. 3. mailan gertatu bezala, zenbat eta anbiguotasun txikiagoa, orduan eta hobeto desanbiguatzen da Markov-en eredu ezkutuen bitartez. Hala ere, hobekuntzarako aukerarik ia ez da geratzen MG4ren kasuan. Gainera, bi etiketen probabilitateak oso hurbil daudenean aukera guztiak ematen direnean, emaitzarik onena MG2ren ondoren aplikatutakoak ematen du —ikus VI.7ko bigarren grafikoa—; bestalde, MG3ren kasuan MG2rekiko gehitutako erroreak (%0,7) ez du konpentsatzen estokastikoaren hobekuntza (%0,15 erabat desanbiguatuz eta %0,42 bi aukera utzita). Ondorioz, 3. azpigramatikako erregelak doitzen diren bitartean, 2. mailako desanbiguaziorako egokiena MG2 dela dirudi.

Azkenik, 1. mailako desanbiguazioari dagokionean, ondorioak ateratzea ez da hain erraza. Emaitzak aztertuz gero 2. mailako errore-tasaren parekoa du eta, printzipioz, hobeto egin beharko luke. Izan ere, hasierako anbiguotasuna %1-2 txikiagoa du, etiketa kopurua erdia da, anbiguotasun-klaseen kopurua ere askoz txikiagoa da, baina oso antzeko emaitzak lortzen ditu. Emaitzei buruzko atalean kategoria-mailako desanbiguazioaren inguruko hausnarketa egiten da, baina mugan gaudela adieraz lezake honek.

VI.3.2 Emaitzak

Aipatutako saiakuntzen emaitzen ebaluazioa egiaztapenerako corpusaren gainean burutu da. VI.8 irudian lehenengo grafikoak erabateko desanbiguazioa egitearen emaitzak agertzen dira eta bigarrenean, 1,01-1,02 analisi utzita lortzen direnak. Ikasketarako corpusaren emaitzetan azaldutako gauza bera gertatzen da, baina zuzentasun-maila zertxobait okerragoa da.

Konbinaketaren ondorioz, desanbiguazio estokastiko hutsa aplikatuta egindako erroreen %62-66 ekiditen dira lehenengo mailako etiketatzean, MULTEXTen bidez egindako errorea ia %12tik %2,4-%4ra jaisten delarik. Bigarren mailari dagokionean, erroreen %68-70 inguru ekiditen dira, desanbiguazio estokastikoak egindako errorea ia %15etik %2,7-3,4ra jaitsita. Azkenik, hirugarren mailan erroreen %64-76 ekiditen dira, errorea ia %40tik %5,3-12,5era jaitsita.

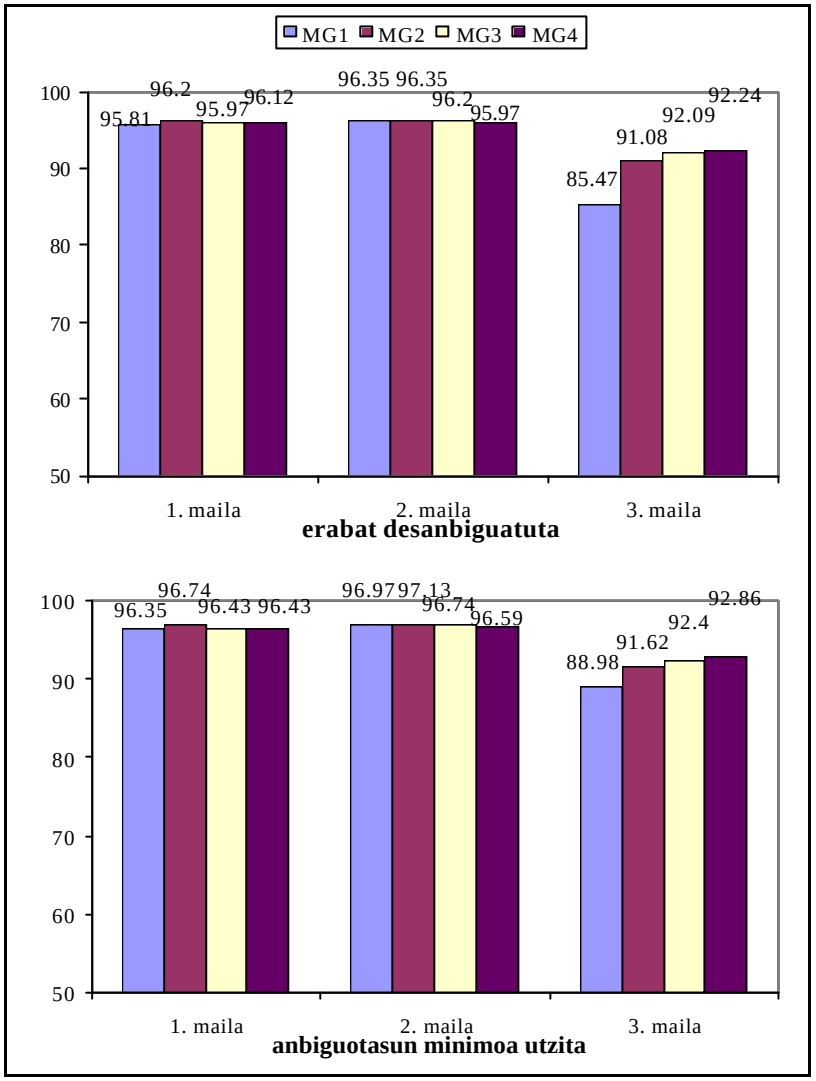


VI.8 irudia.- Ebaluazioa egiaztapen-corpusean.

Murriztapen-gramatikaren aplikazioaren emaitzak aztertzeko, 2. mailari dagokionean, corpus honetan hobeto islatzen da saiakuntzen aurkezpenean MG3 eta MG4ri buruz esandakoa, hots, gehitutako erroreak hobekuntzarako aukerak murrizten dituztela. Kasu honetan, MG3 eta MG4 aplikatzean, gehiago desanbiguatu arren, MG1ekin baino emaitza okerragoak lortzen dira. Probabilitate antzekoak dituzten aukera guztiak mantenduz gero, VI.8 irudiko bigarren grafikoan ikus daitekeenez, MG1 eta MG2ren ondoren desanbiguatutakoan %1eko hobekuntza lortzen da, %96,5 inguruko zuzentasuna lortuz.

Hala ere, emaitzak hobeak espero genituen eta, horregatik, zergatiak aztertu ahal izateko egiaztapenerako corpusa bi zatitan banatu dugu: lehenengoan EEBSri dagokion 1.300 token inguruko zatia, eta, bigarrena, *Euskaldunon Egunkaria*ko 5.800 tokenak. Izan ere, murriztapen-gramatikaren diseinu ia osoa EEBStik hartutako corpusean oinarrituta burutu da,

eta, printzipioz, bertako fenomenoei erantzuteko egina denez, emaitza hobeak eman beharko lituzke corpusaren zati horretan.



VI.9 irudia.- Ebaluazioa egiaztapen-corpuseko EEBSko zatian.

VI.9 irudian EEBSko zatiari dagozkion emaitzak aurkezten dira. Ikus daitekeenez, kasurik gehienetan emaitzak egokiagoak dira, guztiz desanbiguatutakoetan, gainera, 2. mailako emaitzak 1. mailakoak baino hobeak dira eta 3. mailakoetan %2ko aldea dago testu osoarekin konparatuz gero.

Guztiz desanbiguatu gabeko emaitzetan, 2. mailan %97ra iristen da, %1,65eko errorea gehituz soilik. Egiten diren akats asko eta geratzen diren anbiguotasun asko *izen/adjektibo*, *izen berezi/izen arrunt* eta, neurri txikiagoan, *leku-izen/pertsona-izen* motakoak dira eta hauek anbiguotasun gogorrak (*hard ambiguities*) dira, ebazten zailak direnak. Izen berezi konposatuen identifikazio eta sailkapena aplikatuz gero, hauetako batzuk ebatz daitezke, baita egindako errore batzuk zuzendu ere.

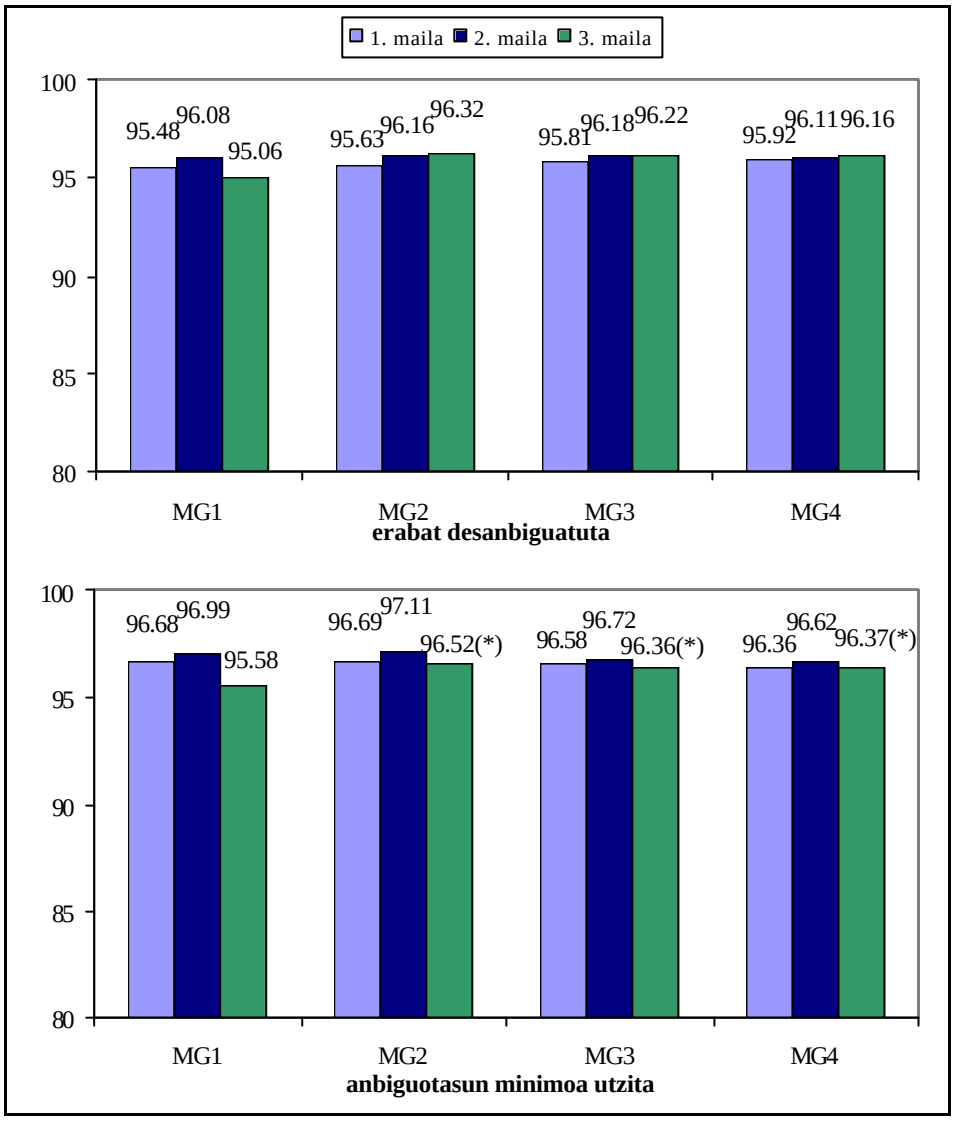
*Euskaldunon Egunkaria*ko zatiko akatsei dagokienean, izenei buruz aipatutakoez gain, loturazko osagaietan eta adberbioetan akats nabarmenak burutzen dira. Hauek gramatikan erregela egokiak gehituta ekidin daitezke, neurri handi batean behintzat. Aditzetan egindako akats batzuk ere hitz anitzeko unitateen tratamenduan erabili ez diren aditz konposatuak landuta ekidingo lirateke. Horrekin guztiarekin %97-97,5 ingurura iristea espero dugu, hori izango bailitzateke hasierako anbiguotasuna kontuan izanda lor daitekeen neurririk egokiena.

1. mailako desanbiguazio osoa burutzean, ikasketa-corpusaren emaitzen tendentzia bera ikusten da, hau da, geroz eta anbiguotasun txikiagoa, orduan eta errore gutxiago. Hala ere, ez da %96ko zuzentasun-mailara iristen. Beharbada, kategoria ez da nahikoa eredu estokastikoak testuinguru desberdinak bereizteko, izatez oso desberdinak diren anbiguotasun-klaseen sekuentziak berdintzat jo eta aukera desberdinen arteko probabilitateen banaketa oso uniformea izango da. Bestela, gaitza gertatzen zaigu 2. mailako emaitzekin alderatuta hain diferentzia txikia gertatzea, problema errazagoaren aurrean egonda ere.

(Ezeiza *et al.* 1998) lanean aipatzen zen maila bateko emaitzak hobe zitezkeela hurrengo mailan desanbiguatu eta aukeratutako etiketa aurreko mailara bihurtuta. Zehazki, lan horretan 3. mailako desanbiguazioa burutu eta emaitzak 2.era itzulita 2. mailako emaitzak hobe zitezkeela aipatzen zen. Nahiz eta hemen aurkeztutako analizatzailearen emaitzak lan horretan erabilitakoak baino anbiguotasun handiago izan, ikasketarako corpusa handiagoa denez, teknika hori aplikatu dugu, gehien bat 1. mailako emaitzak hobetzeko.

VI.10 irudian ikus daitekeenez, 2. eta 3. mailan desanbiguatu eta aukeratutako etiketaren kategoria hartuta, 1. mailan desanbiguatuta baino emaitza hobeak lortzen dira. Gainera, 1,01-1,02 analisi utzita, joera mantentzen da 2. mailan desanbiguatuta. Bestalde, 3. mailan 1,02 analisi utzita, lehenengo mailan batez beste analisi bakarra geratzen da, anbiguotasuna %0,5era iristen ez delako, eta beraz, MG2 aplikatu eta 3. mailan desanbiguatuz gero, %96,5era iristen da kategoria mailan —1. mailan desanbiguatuz gero, emaitzarik onena ez da %96ra iristen—. Anbiguotasun txikia utziz gero, emaitzarik onena MG2 aplikatu ondoren 2. mailan desanbiguatzean lortzen da, %97tik gorako zuzentasuna lortuta.

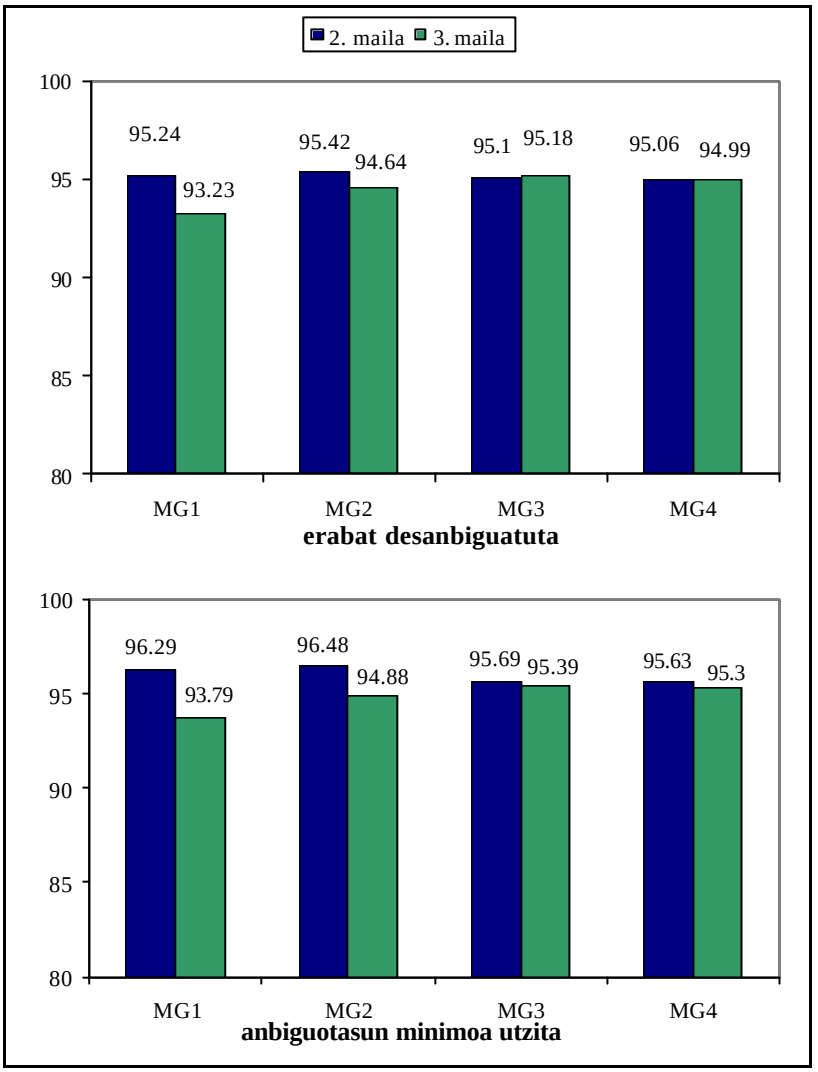
2. mailari dagokionean, saiakuntzen emaitzak VI.11 irudian aurkezten dira. Ikus daitekeenez, analisi bakarra utzita MG3 eta MG4rako emaitzak oso parekoak dira, nahiz eta MG2rekin 2. mailan lortutako emaitzak baino kaskarragoak izan. Hala ere, 3. mailako emaitzak aztertuta, MG2 eta MG3ren artean %1eko hobekuntza lortzen da eta, hori, 2. mailan %0,5eko igoeran gauzatzen da. Beraz, gure ustez, 3. mailako emaitzak hobetuz gero, bai 1. bai 2. mailako emaitzak hobetzea lor daiteke.



VI.10 irudia.- 1. mailako ebaluazioa egiaztapen-corpusean.

Horretarako, bi alderdi landu beharko dira sakonean:

- Murriztapen-gramatikaren 3. eta 4. azpimultzoetan egindako erroreak ahal den neurrian ekidin eta erregela egokiak landu.
- Ikasketarako corpus handiagoa erabili, ahal den neurrian eskuz desanbiguatuta, bereziki 3. mailako desanbiguazio estokastikoan ahalik eta anbiguotasun-klase eta etiketa gehien edukitzeko.



VI.11 irudia.- 2. mailako ebaluazioa egiaztapen-corpusean.

Laburbilduz, maila sakonean desanbiguatu eta maila sinpleagoetara bihurtzea emaitzak hobetzeko aukera interesgarria izan daiteke, sakoneko mailaren emaitzak onak (%95) diren neurrian, behinik behin.

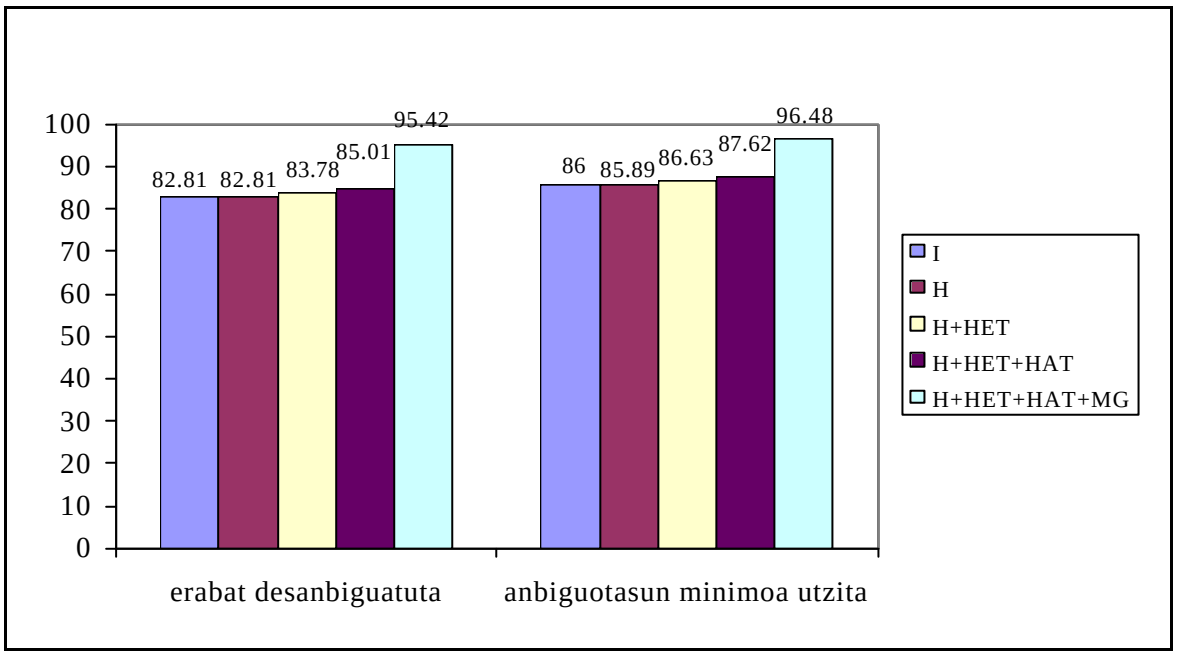
VI.5 Ebaluazioa orokorra

Tesi-lanean zehar aurkeztutako hobekuntza guztiak desanbiguazioan laguntzeko helburuarekin gehitu dira eta helburu hori zein neurritan betetzen duten ebaluatuko da jarraian. Izan ere, kapituluan zehar egindako ebaluazioa hobekuntza guztiak aplikatuz egin da. Lanaren kokapenean, hasierako analizatzaile morfologiko inkrementaletik (I) abiatuta proposatzen ziren hobekuntzak honakoak ziren (Alegria *et al*. 1996):

- Analizatzaile morfologikoaren hobekuntza —III. kapituluan—, besteak beste, eta desanbiguazioari dagokionean, zuzentasun-neurriaren hobekuntza (H).

- Hitz ez-estandarren tratamendua —IV. kapituluan— zehaztasun-neurria handitzeko (HET).
- Hitz anitzeko unitateen tratamendua —V. kapituluan—, analisiaren emaitza egokiagoa izateaz gain, unitate ziurren tratamenduaren bitartez zehaztasunean hobetzeko (HAT).
- Desanbiguazio linguistikoa (MG) eta estokastikoaren konbinaketa (HMM), zuzentasunean hobetzeko —VI. kapitulu honetan—.

Banan-banan aztertu aurretik, hasierako egoeratik abiatuta kapitulu bakoitzean proposatutako tresna guztiak bateratuta, desanbiguazioaren azken emaitzetan duten eragina aurkezten da VI.12 irudian. Kasu guztietan, 2. mailako desanbiguazio estokastiko gainbegiratua aplikatzen da. Abiapuntu gisa, analizatzaile morfologiko inkrementalaren emaitzetan oinarritutakoak aurkezten dira; ondoren, analisi hedatua erabilita lortutakoak; jarraian, analizatzaile hedatuari hitz ez-estandarren tratamendua aplikatuta lortutakoak; gero, hitz anitzeko unitate ziurren tratamendua gehituta lortutakoak, eta, azken emaitza gisa, tresna guztiei desanbiguazio-metodoen konbinaketa (MG2) aplikatzearen emaitzak ematen dira.



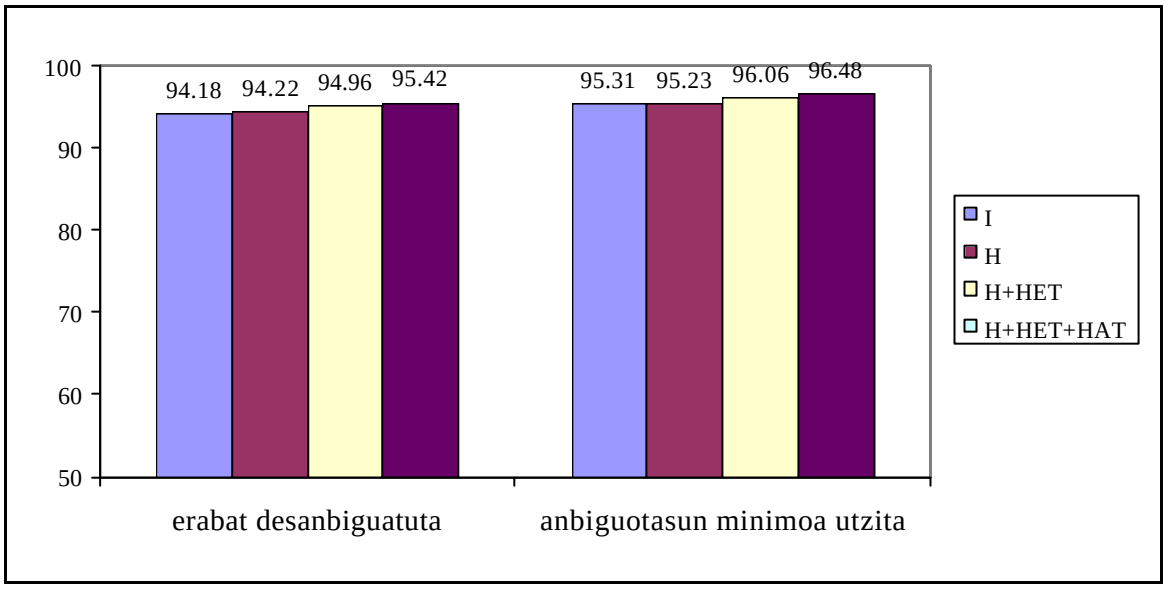
VI.12 irudia.- 2. mailako emaitzak (HMM) proposatutako hobekuntzak aplikatuta.

Irudian ikus daiteke III. kapituluan proposatutako analizatzaile hedatuak desanbiguazioan ezer gutxi laguntzen duela. Dena dela, zuzentasunean hobetzean anbiguotasuna areagotu da, eta faktore hori desanbiguazio estokastikoaren emaitzetan erabakiorra izan daitekeela ikusi da. Aurrerago gai honen inguruko hausnarketa egiten da.

Gainerako tresnek, desanbiguazio estokastikoaren sarreraren anbiguotasuna are gehiago jaistearekin batera, azken emaitzen zuzentasunean igotzea dakarte. Zehaztasunean gehien irabazten duen prozesua desanbiguazio linguistikoa denez, irudian jauzi kuantitatibo handiago

ikusten da VI. kapitulu honi dagozkion emaitzetan. Argi dago, beraz, kapitulu honetan proposatutako desanbiguazio-metodoen konbinaketa, esku artean ditugun baliabideei etekin handiena ateratzeko modurik onena dela. Hala ere, etorkizunerako beste teknika batzuekin saiakuntzak burutu beharko ditugu, gehienbat 3. mailako desanbiguazioa ahalik eta gehien doitzeko.

Dena dela, emaitzak modu egokiagoan ebaluatzeko, desanbiguazio estokastiko hutsa baino, metodoen konbinaketa burutzea egokiagoa da, hasierako analizatzaile morfologikoari proposatutako tresnak gehitzen lortzen diren hobekuntzak aztertzeko. Horretarako 2. mailako emaitzetan duten eragina aztertuko dugu jarraian, beti ere konbinaketarik egokiena erreferentzia gisa hartuta, hots, murriztapen-gramatikaren lehenengo bi azpimultzoak erabilita (MG2).

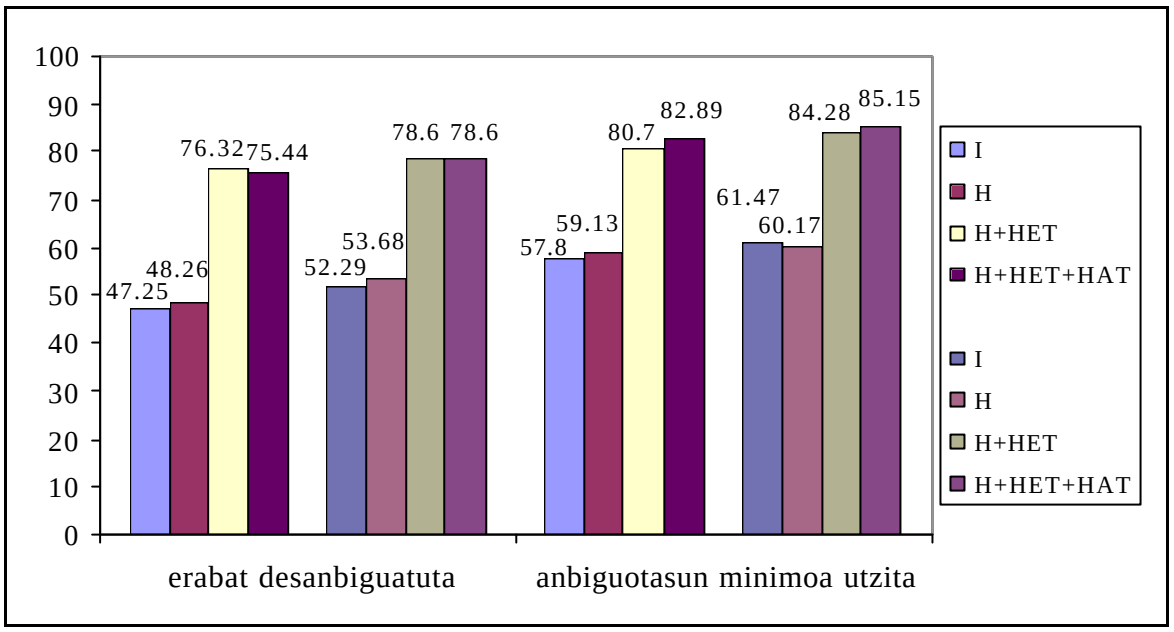


VI.13 irudia.- 2. mailako emaitzak (MG2+HMM) proposatutako hobekuntzak aplikatuta.

VI.13 irudian emaitza horiek guztiak aurkezten dira. Ikus daitekeenez, proposatutako tresna guztiek desanbiguazioren azken emaitza hobetzen laguntzen dute. Diferentzia handiena hitz ez-estandarren tratamenduak egiten du. Analizatzaile inkrementalean soilik oinarrituta egindako erroreen %13,6 inguru ekiditen dira, nahiz eta tratamenduak erroreak gehitu.

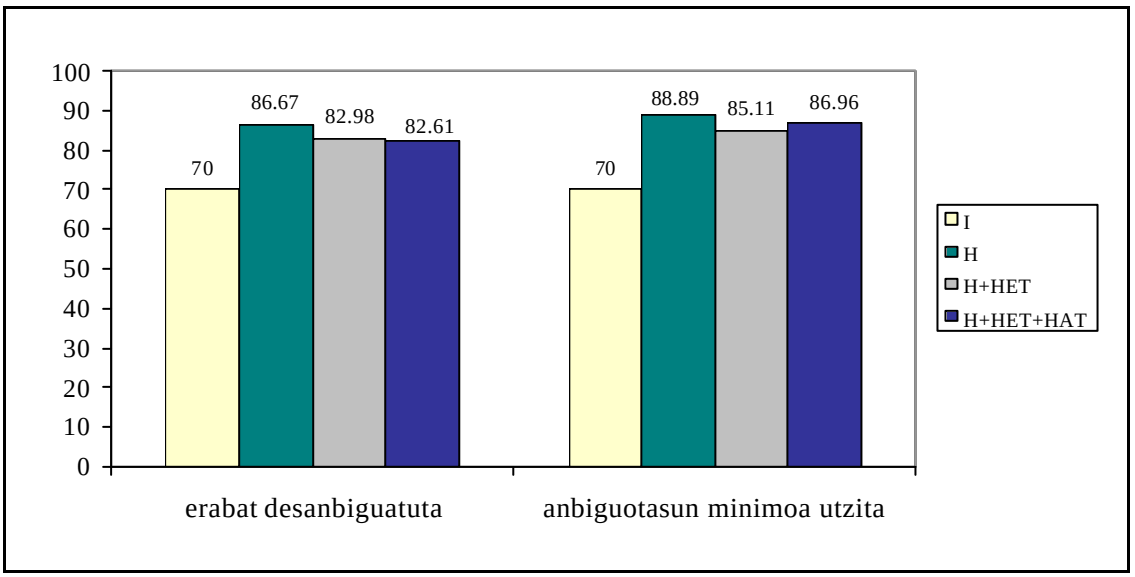
Analizatzaile hedatuari dagokionean, ez da hobekuntza handirik lortzen eta 3. mailako emaitzetan, okerrera egiten duela ere ikusi da. Horren arrazoi nagusia hauxe da, gehitu diren anbiguotasun-mota berriak gramatikaren erregelak diseinatzerakoan ez direla kontuan hartu. Hala ere, gure ustez, analizatzaile hedatuaren beharra argia da, bereziki izen bereziak modu egokian analizatu eta tratatu ahal izateko. Beraz, bere erabilera masiboa egin aurretik, gramatika egokitu beharko da.

Azkenik, hitz anitzeko unitateen tratamendua gehitzeak ere hobekuntza handia eragiten du, ia %0,5ekoa. Orotara, aurreko kapitulu guztietako hobekuntzak kontuan izanik, bakarrik analizatzaile inkrementalarekin egindako erroreen %22,8 ekiditen dira.



VI.14 irudia.- Hitz ezezagunen 2. mailako ebaluazioa.

Hitz ezezagunen emaitzei dagokienean, hasierako zuzentasuna erabatekoa dela kontuan izanik, hitz ez-estandarren tratamenduaren ondorioz erroreen %50 inguru ekiditen dira, proposatutako prozesu guztien artean erabat desanbiguatuta %62era eta anbiguotasun minimoa utzita ia %70era iristen delarik hobekuntza hori. VI.14 irudian desanbiguazio estokastiko hutsa (zutabe koloredunak) eta desanbiguazio konbinatuaren emaitzak (zutabe marradunak) aurkezten dira.



VI.15 irudia.- Aldaeren 2. mailako ebaluazioa.

Aldaeren kasuan, desanbiguazioan ez da errorerik egiten analizatzaile inkrementala erabilita, hau da, hasierako %70eko zuzentasun-maila mantentzen dela. VI.15 irudian ikusten den bezala, analizatzaile hedatua erabiltzean, ordea, aldaera gisa tratatzen ziren hitz batzuk hitz ezezagun gisa prozesatzen dira, hitzen portzentaia jaitsi eta zuzentasun-maila handitzea lortzen delarik. Hitz ez-estandarren tratamendua egiterakoan, berriz, aldaera eta hitz ezezagunen analisiak dituzten zenbait hitzek aldaeraren aukera soilik mantentzen dutenez, berriz ere portzentaia aldatzen da, eta, errorerik egiten ez den arren, zuzentasun-maila jaisten da. Azkenik, hitz anitzeko unitateen tratamendua egitean ere, aldaera batzuk zenbakien adierazpideetan agertzen dira eta osagai bakarra aldaera izanik, hitz estandar gisa kontatu dira. Izatez, errore bakarra gehitzen da azken zutabean agertzen diren emaitzetan, baina hitz kopurua oso txikia denez, erroreak zuzentasunean eragina du.

VI.5.1 Analizatzaile hedatuari buruzko hausnarketa

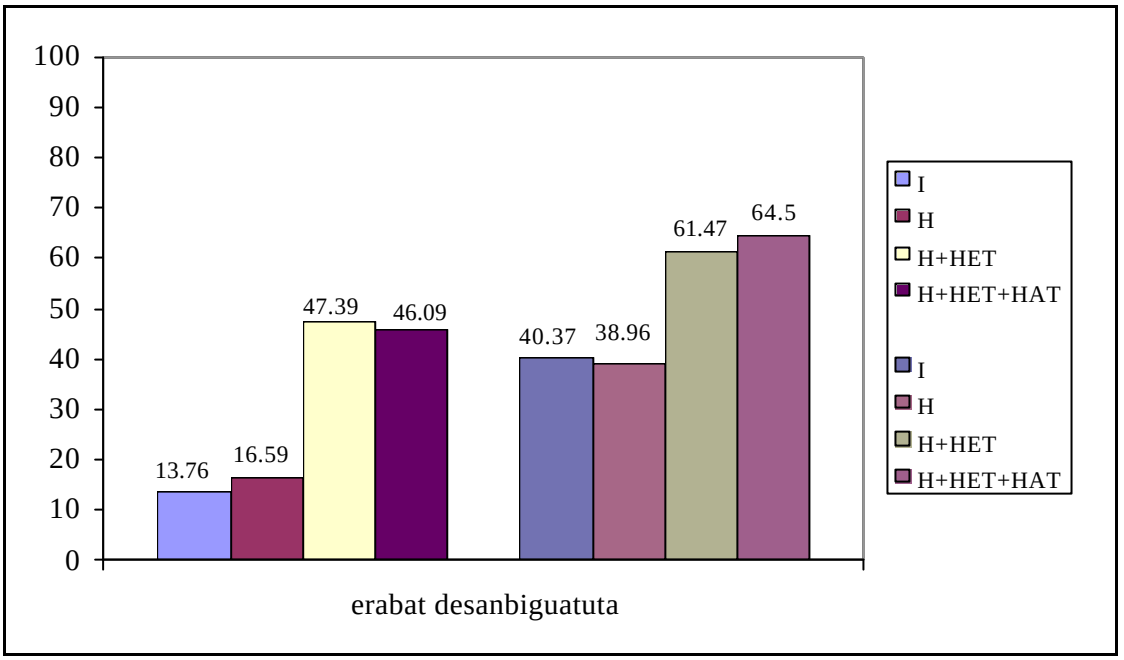
Emaitzen azterketa sakona egin ondoren hainbat ondorio atera daitezke III. kapituluan proposatutako analizatzailearen inguruan:

- Emaitzak izatez okerragoak ez diren arren, desanbiguazioa burutzerakoan errore handiagoa egiten da, hasierako zuzentasuna analizatzaile inkrementalarena baino handiagoa delako, bai hitz estandarrena baita ez-estandarrena ere. Lehenengoei dagokienean, murriztapen-gramatika analizatzaile inkrementalerako diseinatuta dago eta ez dira erregela berriak gehitu kasuistika berria kontuan hartzeko. Ondorioz, hobekuntzarako aukera ematen du.
- Hala ere, VI.12 eta VI.13 irudietan ikusi denez, analizatzaile inkrementalaren emaitzen antzekoak ematen ditu, nahiz eta hitz ezezagunen proportzioa handiagoa izan. Horiek, aukera gehien dituzten hitzak izanik, desanbiguatzen zailenak dira. Baina analizatzaile bat zein beste erabilita lortzen diren emaitzak konparagarriak dira.
- Hitz ez-estandarrei dagokienean ere, aipatu bezala, heuristikoak analizatzaile inkrementalaren irteerarako diseinatu dira, eta, IV. kapituluaren amaieran ikusi den bezala, errore gehiago egiten dute analizatzaile hedatuaren emaitzen gainean. Erreferentzia-corpusean egindako erroreak berraztertu beharko lirateke, izen berezietan gertatzen diren akats berriak nola ekidin erabaki eta prozedurak doitzeko.

VI.5.2 Hitz ez-estandarren tratamenduaren ekarpena

Tratamenduari esker, 2. mailan hitz ez-estandarren desanbiguazioan egiten ziren erroreetatik erdia ekiditen direla ikusi da. Dena dela, 3. mailako emaitzek ere azterketa merezi dutela uste

dugu. Izan ere, maila horretako desanbiguazioaren emaitzetan ekiditen diren erroreen proportzioa oso antzekoa den arren (%40 inguru), zuzentasunean ematen den jauzia askoz nabarmenagoa da, batetik, anbiguotasuna proportzionalki askoz altuagoa delako orokorrean, eta, bestetik, aipatu den bezala, maila honetan desanbiguazio estokastikoak emaitza onargarriak emateko erabiltzen den corpusa txikiegia delako.



VI.16 irudia.- 3. mailako emaitzak hitz ezezagunetan[15].

Ikus daitekeenez, hasierako emaitzak oso kaskarrak ziren eta hitz ez-estandarren tratamenduaren ondorioz, emaitzen zuzentasuna %30ean hobetzen da. Gainera, tresna guztien aplikazioa kontuan hartzen badugu, zuzentasunean lortzen den jauzia %50ekoa da orotara.

Bestalde, VI.16 irudian ikus daitekeenez, tratamenduarekin lortutako emaitzak (3. zutabe koloreduna) murriztapen-gramatikaren erregela guztiak aplikatuta (1. zutabe marraduna) baino hobeak dira[16], askoz ere metodo sinpleagoak erabilita (%47ko zuzentasuna %40ren aurrean). Horrek hitz ezezagunen desanbiguazioaren zailtasuna berresten du, IV. kapituluan esan den bezala, analizatzaileak ematen dituen interpretazio-multzoak ez direlako hitz estandarretan aurkitzen direnak, eta, beraz, horientzako erregelak diseinatzea ez delako ez intuitiboa ezta berehalakoa ere.

15 Koloredun zutabeetan desanbiguazio estokastiko hutsaren emaitzak azaltzen dira eta marradunetan desanbiguazio konbinatuarena.

16 2. mailako emaitzetan nabariagoa da diferentzia, %76koa (3. zutabe koloreduna) %52ren (1. zutabe marraduna) aurrean, VI.14 taulan, erabat desanbiguatutakoaren emaitzetan, ikus daitekeenez.

Hala ere, hitz estandarren emaitzetatik oraindik ere nahiko urrun geratzen dira emaitza hauek, beraz, etorkizunean hauen desanbiguazioa bideratzeko modu ezberdinak aurkitzea komeni da.

VI.6 Ondorioak

Kapitulu honetan aurkeztutako desanbiguazio-metodoen konbinaketak oso emaitza onak ematen ditu. Izan ere, metodo estatistiko hutsen bidez desanbiguatu nahi izanez gero, askoz ere baliabide gehiago beharko genituzke. Orokorrean, emaitzarik onenak ingeleserako lortutakoak direla esan daiteke, %97,5-%98,5 inguruko zuzentasun-mailara iritsi direlarik, hainbat etiketatzaile edota hizkuntz eredu konbinatuta.

Dena dela, ezin dugu gure etiketatzailea ingelesekoekin konparatu, arrazoi anitz direla medio. Hasteko, flexio-morfologia oso sinplea du eta esaldiaren osagaien ordena nahiko finkoa da. Gainera baliabide asko garatu dira ingeleserako, bai corpus aldetik baita tresnen aldetik ere, horretarako mundu osoan zehar hainbat talderen esfortzua horretan inbertitu delarik.

Konparazioa bestelako ezaugarriak dituzten hizkuntzekin burutzea egokiago iruditzen zaigu eta, jarraian, morfologia konplexuagoa eta, esaten dutenaren arabera, oso baliabide gutxi dituzten hiru hizkuntzetarako burututako lana aurkezten da jarraian, gure lanarekin konparatzeko asmotan.

VI.6.1 Beste hizkuntzetarako emaitzak

Ingelesa ez bezalako hizkuntzetan, eranskariak izanagatik edota eratorpen eta flexio-sistema emankorra izanagatik morfologia aberatsa duten hizkuntzetan edota osagaien ordena (ia) librea duten hizkuntzetan, alegia, desanbiguazio-emaitza onargarriak lortzeko sortzen diren arazoak anitzak dira. Ondoren, hurbilpena batzuk aurkezten dira eta emaitzen inguruko hausnarketa egiten da.

VI.6.1.1 Turkiera

Turkierak morfologia aberatsa du, hizkuntza eranskaria da eta eratorpen eta flexio oso emankorra du, beraz, ezin da hitz-formen araberako desanbiguazioa planteatu, euskaraz gertatzen den bezala, hiztegiaren tamaina izugarri handitzen delako. Hala ere, desanbiguazio estatistikoa aplikatzea interesgarritzat jo dute. Aurretik erregeletan oinarritutako etiketatzaile

on bat badago turkierarako (Oflazer eta Kuruöz 1994; Oflazer eta Tür 1996; Oflazer eta Tür 1997), baina erregela-sistema mantentzea eta hobetzea garesti gertatu ohi da —nahiz eta guztiak eskuz diseinatu behar ez diren—. Hori dela eta, teknika estatistikoak aztertzea erabaki zuten.

(Hakkani-Tür *et al*. 2000) lanean eredu klasikoak erabiltzeko arazoak planteatzen dira, hiztegi-tamaina handiegia, desanbiguaziorako interesgarria den informazio guztia etiketa bakarrean islatzeko arazoak, eta, orokorrean, euskararen kasuan agertzen den datu-sakabanaketaren arazoa (*data sparseness*). Sarrerako batezbesteko analisi kopurua 1,53-1,55 ingurukoa da, zenbait kolokazio identifikatu eta argi eta garbi ezinezkoak diren analisi batzuk baztertu ondoren.

Informazio morfologikoa etiketan islatu ordez, hitzaren segmentazio morfologikoan oinarrituz, erro eta *inflectional group* (IG) edo flexio-multzoen sekuentziak desanbiguatzen dituzte. Ikasketarako corpuseko sekuentzia ez-anbiguoak (1 milioi hitzeko corpusetik 650.000 hitz) baliatzen dituzte eta, eskuz desanbiguatutako 12.000 eta 20.000 hitzetako beste bi corpus gaineratzen dituzte. Horrela, hitz baten erroa, bere flexio-multzoekiko eta aurreko hitzaren erroarekiko dependentzia du soilik. Erro eta IG ereduen bitartez, %93,95eko zuzentasuna lortzen da.

VI.6.1.2 Errumaniera

Errumanierarako Tufis eta Mason-ek (1998) etiketatze mailakatua (*tiered tagging*) proposatzen dute. Kasu honetan, euskararen antzera, deskribapen morfologiko osoa erabiltzeak etiketatzearen konplexutasuna areagotzen du eta ikasketa-corpusaren tamaina handia eskatzen du horrek. Horregatik, interpretazio morfologikoetan oinarritutako etiketa-multzo laburtua erabiltzen dute desanbiguaziorako, guztira 89 etiketa.

Multzo murriztuaren bidezko desanbiguazioaren emaitza, jatorrizko interpretazio morfologiko oso bakarrera %90 inguruan zuzenean mapa daiteke. Gainerako %10etan interpretazio bat baino gehiago ematen dira. Bigarren fase batean, testuinguru-erregela sinpleak erabilita hitz horietatik %98 inguru ondo desanbiguatzen dira.

Ikasketa-corpusaren tamaina 225.000 hitzekoa da, eta egiaztapenerakoak 25.000 ditu. Ikasketa-corpusean deskribapen morfosintaktiko guztien (611) %72 inguru eta deskribapen osoen gaineko anbiguotasun-klase guztien (869) %71 inguru agertzen dira. Beraz, lortzen den hizkuntz eredua nahiko esanguratsua da.

Desanbiguatzailearen sarrerako anbiguotasunari dagokionean, 1,55-1,63 interpretazio direla esaten du, baina ez du zehazten deskribapen morfologiko osoaz ala etiketa-multzo

murriztuan oinarrituriko neurriak diren. Dena dela, bigarrenean oinarrituta dagoela suposatzen da.

Hasierako emaitzetan %95,63ko zuzentasuna lortzen dute. Etiketa batzuen artean bereiztea (desanbiguatzea) zaila izanik errore asko gehitzen zirenez, etiketa bakarrean biltzea erabaki zuten. Etiketa-biltzearen adibide gisa konjuntzio eta adberbioaren artekoa ematen du. Beste interpretazio morfosintaktiko batzuk lexikotik ere kendu zituzten. Horren ondoren, %96,22ko zuzentasuna lortzen dute.

Dena dela, mapaketa zuzena lortzen duten hitzei dagokien neurria da hau. Gainerako hitzen erregelen bidezko desanbiguazioan (%10 gehienez) %98ko zuzentasuna lortzen dela diote. Ondorioz, batezbesteko zuzentasuna %95,87 eta %96,4 izango da gehienez ere.

VI.6.1.3 Txekiera

Txekierarako Hajic eta Hladká-k (1997) ere deskribapen linguistikoen kopuru altua arazo nagusitzat jotzen dute. Bere kasuan, 600.000 hitzeko corpusean 1.100 interpretazio desberdin daudela diote, batezbesteko analisi kopurua 3,65ekoa izanik. Dena dela, deskribapen hori ez zetorren bat euren kategoria-sistemarekin eta 100.000 hitzeko corpus egokitua 882 interpretazio desberdin aurkitu zituzten, batez beste 2,36 etiketa izanik.

Hitz-formetan oinarritutako lehenengo eta bigarren ordenako Markov-en eredu ezkutuak aplikatuta, hau da, bigramak eta trigramak erabilita, %81,53 eta %81,14ko zuzentasuna lortu zuten, hurrenez hurren.

Emaitza hauek hobetzearren, anbiguotasun-klaseetan oinarritutako eredu markoviarrak erabili zituzten, zehatzago esanda XT etiketatzailea. Gainera, saiakuntzak etiketa-multzo desberdinekin burutu zituzten. Horrela, VI.1 taulako emaitzak lortu zituzten.

T	AR	R
47	%39	%91,7
43	%36	%93
34	%14	%96,2

VI.1 taula.- Desanbiguazio-emaitzak[17].

Dena dela, etiketa kopurua murriztean deskribapenaren aberastasuna galtzen denez, interpretazio morfologikoetan agertzen diren ezaugarri desberdinak modu independentean

[17] T = etiketa kopurua (*tags*);
AR = anbiguotasun-tasa (*ambiguity rate*);
R = zuzentasuna (*recall*)

desanbiguatu eta ezinezkoak diren ezaugarri konbinaketak ekidinez, hasierako %81 inguruko emaitzak hobetzeari ekin eta (Hajic eta Hladká 1998) lanean aurkeztu dute. Horretarako, Markov-en ereduak alde batera utzi eta entropia maximoaren bidezko etiketatzailea eratu dute, eta zuzentasuna %93,8raino igotzea lortu dute.

VI.6.2 Konparazioa eta ondorioak

Orokorrean, ingelesa baino morfologia aberatsagoa duten hizkuntzak tratatzen dituzten taldeek bi arazo nagusi plazaratzen dituzte: batetik, baliabide falta, gehienbat corpus etiketatuen falta, eta, bestetik, interpretazio morfologikoen kopuru handia.

Lehenengoari buruz, aurkeztutako adibideetan gure kasuan baino corpus handiagoak dituztela argi geratu da. Eta, bigarrenari dagokionean, 4. mailako etiketa kopuruarekin konparatuz gero, euskararen parekoa izan daitekeen bakarra turkiera da —aurkeztutako lanerako corpusean 10.000 desberdin daudela diote egileek—, hau ere eranskaria baita. Gainerakoetan aipatzen diren kopuruak 1.000 ingurukoak dira eta, arestian ikusi den bezala, 3. mailako etiketa-multzoan 500 etiketa inguru eta 4. mailan erabili nahi izanez gero, 30.000tik gora ere agertzen dira.

Errumanieraren emaitzak gure 2. mailako emaitzekin konparagarriak dira. Etiketa kopurua handiagoa eta anbiguotasun-klase gehiago badira ere, askoz corpus txikiagoan guztien agerpenak biltzen dituzte. Euskararen kasuan, berriz, milioi bat hitz ingurura bitartean 500 klase inguru agertzen dira, eta askoren ebidentzia gutxi dago hala ere.

Txekierari dagokionean, XTren bitarteko emaitzak aztertuta, 47 etiketekin lortzen diren emaitzak gure 2. mailakoekin konparatuz gero, gure emaitza hobea da, baina entropia maximoaren bidezkoak ezin dira gurearekin konparatu. Ereduan gehitutako morfemei buruzko ezaugarri desberdinen informazioak asko laguntzen duela argi dago, nahiz eta etiketa kopurua handia izan.

Azkenik, turkierari dagokionean, aipatu desanbiguazio estatistikoaren emaitzak gureak baino okerragoak direla, baina guk 2. mailan baino informazio aberatsagoa erabiltzen dute. Gainera, eurek ez dute erregeletan oinarritutako desanbiguaziorik aurretik aplikatu. Berez, azken etiketatzaile honek oso emaitza onak lortzen dituela ikusi dugu etiketatzaile hauei eskainitako atalean.

Ondorio gisa, baliabide aldetik beste hizkuntza batzuen atzetik egon arren, dauzkagunak ondo baliatuz, beste hizkuntza batzuetan lortzen diren emaitzen parekoak lortzen ditugu. Beraz, baliabideak areagotzen ditugun heinean, hobekuntzarako aukera ere izango dugu, gehienbat 3. mailako emaitzak kontuan hartuta.

VII Lematizazioaren eta etiketatzearen aplikazioak

Tesi-lan honetan azaltzen diren oinarrizko tresnak zenbait aplikaziotan integratu dira, batzuetan ikerkuntza-mailan edo modu esperimentalean, eta, beste batzuetan, modu komertzialean. Murriztapen-gramatika aplikatzen duen programaren eskubideak ikerkuntzarako baino ez dauzkagunez, aplikazio komertzialetan ezin izan da integratu, beraz kasu horietan lematizazioaren emaitza anbiguoak erabili dira. Dena den, hainbat kasutan, informazioaren berreskurapenean adibidez, ez dirudi desanbiguazioa aurrerapen handiegirik dakarrenik, galdera laburretan oso zaila baita osagaiak desanbiguatzea.

Integrazio honekin frogatuta geratzen da diseinu-mailan aipatu diren orokortasuna, modulartasuna, ahalmena eta estaldura; arazorik gabe eta emaitza egokiekin barneratu baitira oinarrizko tresnak aplikazio mota oso desberdinetan.

Horrez gain, hemen azaltzen diren tresnak *Cycit*-ek onartutako ITEM (TIC96-1243-C03) eta *MiCyT*-ek HERMES (TIC2000-0335-C03-03) proiektuetan erabili dira tresna eleanitzak egiteko asmoarekin eta, noski, sintaxia ekiteko garaian aurreprozesu gisa (Aldezabal *et al*. 1999-c). Gainera, SGMLn oinarritutako formatu orokor bat proposatu da (Artola *et al*. 2000, Aldezabal *et al*. 2002), tresna hauek erabilpen zabala eta internazionala izan dezaten.

Ondoren azaltzen dira lematizatzaile/etiketatzailea integratzen duten aplikazio desberdinak, dagozkien arloen arabera sailkatuta: lexikografia, informazioaren berreskurapena eta erauzketa, eta hizketaren sorkuntza.

VII.1 Lexikografia

Hau izan zen lehen aplikazioa, etiketatzaile/lematizatzailearen ideia sortu baitzen, besteak beste, UZEIk zuen behar batetik: corpus orekatu erraldoi (Urkia eta Sagarna, 1991) bat

lematizatu eta etiketatu behar zuten, eta modu semiautomatikoan egiten bazuten ere (oso metodo sinpleak erabiliz) tresna linguistiko automatikoekin egitea proposatu ziguten eraginkortasuna handitzeko asmoz. IXA taldeak garatu dituen Office-ko *plugin*-etan integratzea beste urrats bat izan da.

VII.1.1 EEBS: egungo euskararen bilketa sistematikoa

Egungo Euskararen Bilketa-Lan Sistematikoa (EEBS) Euskaltzaindiaren eskariz UZEIk garatzen duen lantegia da (*http://www.uzei.com*).

VII.1.1.1 Ezaugarriak

Ekimen honen ezaugarri nagusiak ondoko puntuetan labur daitezke:

1. XX. mendeko euskara jasotzen duen corpus estatistikoa, ia 4.237.000 testu-hitzez osatua (1900-1998 urteetakoa). Erabili izan den eta erabiltzen den euskararen lekuko eta erakusgarri izatea du egiteko nagusi eta ia bakarra, eta ez ereduzko hizkuntza proposatzea.
2. XX. mendeko euskal argitalpenen inbentario osoa da abiapuntua. Argitalpenek osatzen duten unibertsotik abiatuta, corpus orekatu bat selekzionatu da, jasotako 6.047 obra-zatik osatzen dutelarik lagina.
3. Corpus irekia da oraingoz, urtero eguneratzen dena, nahiz mendea bukatzean corpus itxi izatera pasako den, mende oso baten erakusgarri. Bestalde, euskara idatzia jaso da hor, ez ahozkoa.
4. Testu-zatiak SGML (*Standard Generalized Mark-up Language*) formatu estandarrean ezarri dira.
5. Corpus lematizatua eta etiketatua da. Lematizazio hau, bestalde, ez da hitz bakunetara mugatzen; hitz soilez gain, hitz elkartuak, eratorriak eta bestelako hitz anitzeko unitate lexikalak ere markatu dira: *etxe* lema soilaren ondoan, *etxe orratz, etxe-abere, etxe-tresna, etxeko, etxeko jaun, etxekoandre, etxepe, etxetxo, etxeño, etxezain* bezalako lemak ere adieraziz. Edo, *hala* soilaz gain, *hala ere, hala eta guztiz ere, hala... nola, hala nola* modukoak ere zehaztuz.
6. Horiek horrela, 98.800 lema desberdin jasotzen ditu corpusak.

VII.1.1.2 Etiketatze-prozesua

Lematizatze/etiketatze lana modu semiautomatikoan burutu da, EUSLEMen oinarrituta. SGMLz markatutako testua EUSLEMek prozesatzen du eta honen irteera datu-base batean jasotzen da.

Txostenean zehar azaldutako EUSLEMi ezaugarri berri bat gehitu zaio: erabiltzailearen lexikoa. Erabiltzailearen lexikoak lexiko orokorra osatzen du baina ez da batera konpilatu behar, horrela denbora aurreztu eta malgutasuna irabazten delarik (ikus III.1.3). Erabiltzailearen lexiko honen bitartez testu ez-estandarretan (euskalkiren batez idatzitakoak edo testu zaharrak) edo berezietan (pertsona-izen zein leku-izen berezi anitz dutenetan) analisi zehatzagoak lortzen dira. Beraz, UZEIn testu berezi horiei tratamendu gehigarria ematen zaie, etiketatu aurretik ez-analizatutako hitzen zerrenda, maiztasunaren arabera sailkatuta, lortzen baita, eta horretan oinarrituta erabiltzailearen lexikoa aberasten da.

VII.1.1.3 Ebaluazioa

UZEIk idatzitako txosten bat (komunikazio pertsonala) ebaluazio kualitatiborako baliagarria dela uste dugu. Hona hemen txosten horren pasarte batzuk:

Dudarik gabe, EEBS corpusaren lematizazioan lagungarri da EUSLEM: oso helduleku ona da bertatik abiatzeko, nahiz eskuzko zuzenketak egiten diren sistematikoki.

- *testu-hitzak multzokatuta iristen zaizkigu (forma deklinatuak eta aditz jokatuak batetik, eta lemen aldaerak bestetik). Hemendik abiatuta, errepasoa azkarra da.*
- *HAULak, kasurik emankorrenetan behintzat, detektatu egiten ditu eta, beraz, lematizatu: testuinguruaren irakurketa murriztu batekin errepasoa azkarra da. Gainera, askotan maiztasun handiko hitzak dira HAULen osagarriak eta forma soilen lematizatze-errepasoa arintzen du (adib. "hala ere" lema datorrenez, "hala" eta "ere" forma soil gutxiago dira eta horiek errazago kontrola daitezke).*
- *Kategoria/azpikategoriaren arabera lematizatuta datozenez, helduleku horrek ere behin-betiko lematizazioa errazten du.*

Beraz, EUSLEMetik datorren guztia berriro begiratzen bada ere, helduleku egokia da guztia azkar errepasatzeko. Noski, euskara batuan dauden formetan gertatzen da hau, askoz zailagoa da euskalki desberdinetan idatzita dauden testuetan, baina hauen tamaina ere ez da handia, eta bereiz prestatutako hiztegitxoekin neurri batean behintzat konpontzen da.

(...)

1991tik hona hasi ginen EUSLEMekin lematizatzen eta testu berriak ziren: normalizatuak. Gainera, bi urte hartzen dituzten multzoak dira batera lematizatzen direnak, ez testu-masa handiegiak.

Hala ere, aurtengo datuetan oinarrituta esan daiteke EUSLEMen lematizatutako testu berriak errepasatzean, egunean pertsona batek 5.000 lema inguru ikusi eta zuzentzen dituela.

VII.1.2 Hiztegien kontsulta

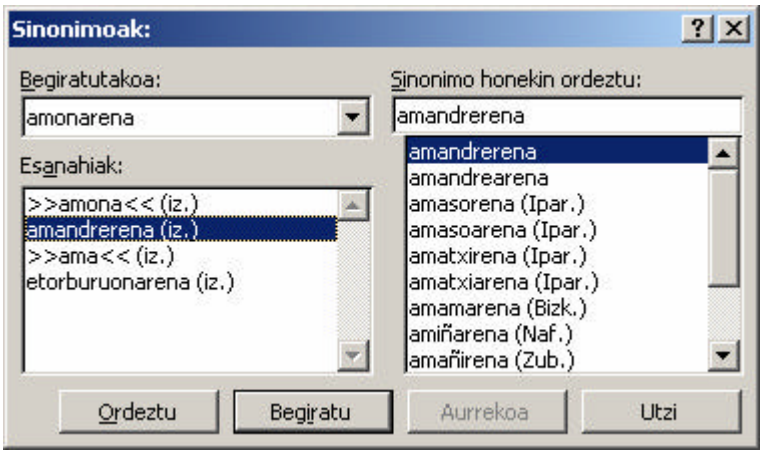
IXA taldeak euskararako aplikazio orokorrak egiten hasi zen 1994an Xuxen zuzentzaile ortografikoa kaleratu zuenean Hizkiarekin lankidetzaz. Azken urteetan lan hori areagotu egin da hiztegien kontsulta integratuz hedapen handieneko bulego-aplikazioetan. Lematizazioa barneratzen duten bi produktu burutzen joan dira azken urteetan: hiztegi elebiduna eta sinonimoen hiztegia.

Bi aplikazioak Unibertsitate-Enpresa proiektu gisa aurkeztu izan dira Eusko Jaurlaritzaren Hezkuntza eta Industria sailetan eta biek finantziazioa lortu dute. Hona hemen proiektuen deskribapena:

- *Elhuyar Hiztegia eta testu-prozesamenduko programen integrazioa. Eusko Jaurlaritza. 1999-2000. Elhuyar enpresarekin batera.*
- *UZEI sinonimoen Hiztegiaren integrazioa Word testu-prozesadorean. Eusko Jaurlaritza. 2001-2002. UZEI institutuarekin batera.*

Hiztegi elebidunean euskarazko zein gaztelaniazko lematizazio sinplea sartu izan den bitartean, sinonimoen hiztegian, berriz, sorkuntza morfologikoa ere barneratu egin da. Beraz, hitz baten sinonimoa bilatzen denean, lehenengoz lematizatzen da eta lema eta kategoria kontuan hartuta hiztegian bilatu ondoren, lema baliokidetik abiatuta eta aurretik erauzitako informazio morfologikoa erabiliz sorkuntza egiten da. Zenbait kasutan, gutxienekoetan, ez da lortzen sorkuntza morfologikorik eta lema hutsa ematen da, lemaren marka gehiturik.

VII.1 irudian sinonimoen hiztegiaren erabileraren adibide bat azter daiteke. "*amonarena*" da kontsultatu nahi den hitza eta lematizazio prozesuaren ondorioz bi izen lortzen dira *amona* eta *ama*. Bigarrena harrigarri samar irudi lezake, baina genitibo plural hurbila eta nominatiboa eranstea ez da hain arraroa. "*amona*" hitzetik abiatzen bada erabiltzailea, sorkuntza gertatuko da eta, genitiboaren analisi bikoitza zenez (mugagabea eta mugatu singularra), sinonimo bakoitzak (*a* organikoa ez dutenean) bi forma sortuko du: *amandre*-tik *amandrerena* eta *amandrearena*, *amatxi*-tik *amatxirena* eta *amatxiarena*, etab.



VII.1 irudia.- Sinonimo hiztegiaren kontsulta baten adibidea

Ez elebidunean ez sinonimoenean, ez da desanbiguazioa barneratu aipatutako lizentzien arazoarengatik, baina modu esperimentalean egina dago.

VII.2 Informazioaren berreskurapena eta erauzketa

Arlo honetan taldea ari da lan anitz garatzen eta horren oinarrian, EUSLEMez gain, txostenean aipatutako entitateen ezagutzaileak (datak, zenbakiak, ...) oso tresna baliagarriak dira eta izango dira etorkizunean. Beste aplikazio batzuen garapena eskuartean ditugun arren, bi dira orain arte garatu direnak: GaIn izeneko Internet/Intranet bilatzailea (Aizpurua *et al*. 2000) eta terminologia erauzketarako aplikazioa (Alegria *et al*. 2002-a).

VII.2.1 GaIn

Euskarazko testuetarako Internet/Intraneteko bilatzaile aurreratua da GaIn (Aizpurua *et al*. 2000). Bere garapenerako Eusko Jaurlaritzaren eta Gipuzkoak Foru Aldundiaren laguntza izan du 2000 eta 2001 urteetan zehar unibertsitate eta enpresaren arteko teknologia-transferentziaren programaren barruan. Plazagune S.L izeneko enpresa izan da aplikazioa merkaturatu duena. Proiektuaren izen ofiziala hauxe da:

- *GaInternet: Interneteko euskarazko testuentzako eduki-arakatzaile adimenduna. Eusko Jaurlaritza, Gipuzkoako Foru Aldundia. 2000-01.*

Tresnaren funtsa *Swish-E* (Swish-E 2000) bilatzaile askearen egokitzapena da, eta hiru osagai nagusi ditu: robota, indexatzailea eta bilatzailea. Banan-banan aztertuko ditugu.

VII.2.1.1 Robota

Sarea (edo fitxategi-sistema) formatu zehatzetako testuen bila (*txt*, *html*, *pdf*, *doc*, *rtf*, ...) usnatzen du eta interesgarriak hautatzen ditu. Egokitzapen batzuk egin dira eta interesgarriena, lan honetatik kanpo geratzen bada ere, izan da euskarazko testuen hizkuntza-ezagutzailearen integrazioa. Izan ere, euskarazko testuak baino ez ditu hautatuko eta indexatzaileari bidaliko.

VII.2.1.2 Indexatzailea

Alderantzizko indizeak sortzeko modulua da. Jatorrizko programan hitzaren arabera egiten bazen ere, egindako aldaketarekin euskarazko lemen arabera indexatzen da. Hauxe da tresnaren gure egokitzapeneko atal nagusia, eta horretan desanbiguaziorik gabeko lematizazioa barneratu da bere hiru urratsetan: estandarra, aldaerak eta lexikorik gabekoa.

Desanbiguazioa ez integratzeko arrazoia bikoitza da kasu honetan: lizentzia-eskubidearekin momentuz dugun arazoaz gain, kontuan hartu behar da galderak laburrak izango direla (bilatzaile moduluan), hitz bakarrak sarri, eta horretan oso zaila izaten dela desanbiguazioa burutzea.

Lematizazioa integratzean modulartasuna oso garrantzitsua zelakoan lematizatzailea zerbitzari moduan inplementatu da, beraz, makina berean zein sareko beste batean egon daiteke eta aplikazio-protokolo baten bitartez lematizazio-lan hori burutuko da aplikazio honetarako edo beste edozertarako. Hau lortzeko, kontuan hartuz hiru urratsez gain zenbait iragazketa eta heuristiko aplikatzen direla, lematizatzailearen berrosatzea burutu behar izan da.

Lematizatzaileak lema eskaintzeaz gain kategoria ere eskaintzen duenez beste hobekuntza bat egin ahal izan da, *stop-list*a erabili beharrean kategoriaren araberako hautapena, edo bazterketa, burutzen baita, hitz/kategoria funtzionalak baztertuz (determinatzaileak, konjuntzioak, aditz laguntzailea, ...).

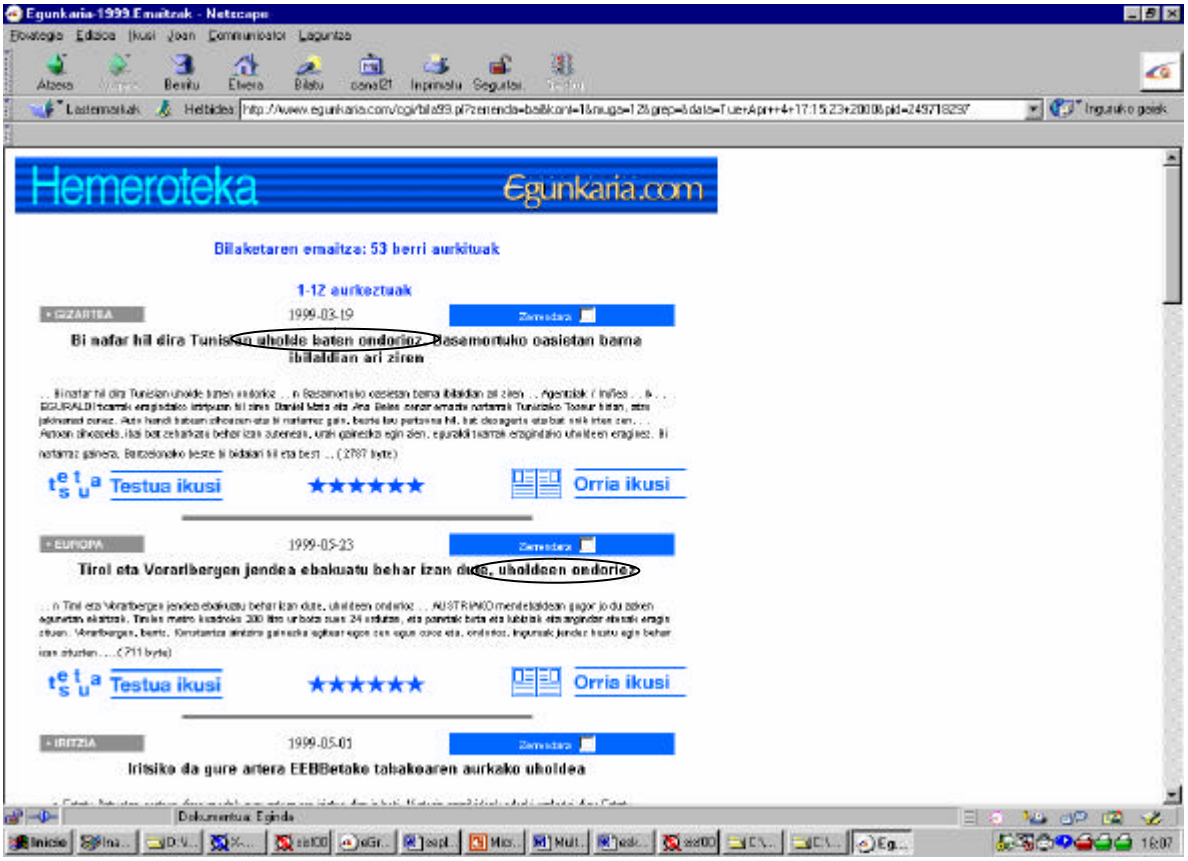
VII.2.1.3 Bilatzailea

Erabiltzailearekiko elkarrizketa eta indizeen gaineko bilaketa da modulu honen helburua. Gure egokitzapenean, interfazearen itzulpena eta egokitzapenaz gain, indexazio ere barneratu da, izan ere barneratzeaz baino lematizazioaren zerbitzariari eskaera eginez burutuko da, zeren indizeak lemak izanik ezin baita espero erabiltzailearen galderetan lemak agertzea. Beraz, galderako hitzak lematizatzen dira eta lema horien agerpenak bilatzen dira testuetan ohiko algoritmoak erabiliz aurkezteko garaian.

Sistema batzuetan, indexatzean lematizazioa egin beharrean, bilaketan sorkuntza morfologikoa burutzen da, baina hori euskararen kasuan ez da oso egokia, eranskaria izatean sorkuntza aberatsegia baita.

VII.2.1.4 Ebaluazioa eta adibideak

Tresnak bi bertsio izan ditu (lehenengoan lematizazio estandarra baino ez zen egiten) eta nahiko arrakastatsua gertatu da. Dagoeneko hiru web-gune garrantzitsutan erabiltzen da: *www.jalgi.com* euskarazko direktorio nagusietako bat, *www.egunkaria.com/hemeroteka* euskarazko egunkari bakarraren hemeroteka eta *www.zientzia.net* irakaskuntzari begirako web-gune handia. Epe laburrean beste batzuetan erabiliko da.



VII.2 irudia.- kontsulta baten adibidea GaIn-en

Zehaztasunaren ebaluazio kuantitaboa egitea zaila da eta horretan ari da IXA taldea, baina, hala ere, euskara bezalako hizkuntza baterako lematizazioa behar-beharrezkoa dela ezin da ukatu. Adibide gisa *egia* hitza erabil daiteke. *Stemming* edota espresio erregularrak erabiliko balira *egi* sasilema lortuko litzateke baina *egi*-tik abiatuta lortzen diren formak (eta benetako lemak) izugarri handia da: *egiazki, egiaztatu, egile, egin, egitasmo, egitamu, egitura, ...*

Funtzionamenduaren adibide gisa VII.2 irudia dugu. "*uholdeen ondorioa*" galderari *Euskaldunon Egunkaria*ko hemerotekan dagozkion lehen dokumentuak agertzen dira bertan,

eta lehen biek izenburua honako zati hauek dituzte: "*uholde baten ondorioz*" eta "*uholdeen ondorioz*". Argi dago lematizazioaren eragin positiboa.

Etorkizunari begira bilaketaren zehaztasuna handitu nahi da, entitate eta terminologiaren araberako indizeak erabiliz.

VII.2.2 Terminologiaren erauzketa

Proiektu honetan lematizatze/etiketatze lana aurreprozesuaren partea da, termino hipotetikoak egitura morfosintaktikoaren arabera aukeratzen dira-eta (Urizar *et al*. 2000). Lematizatze/etiketatze lana analisiaren 3 urrats inkrementalez eta desanbiguazio-prozesu osoaren bitartez burutzen da.

Hiztegi teknikoetan zein corpus berezituetan aztertutakoaren arabera euskarazko hitz anitzeko termino gehienak espresio erregular honen bitartez identifika daitezke[1]:

```
(Nnc|Aprep)+ (N|Apos+) | (ADV | (Nnc* Nabs)) V
```

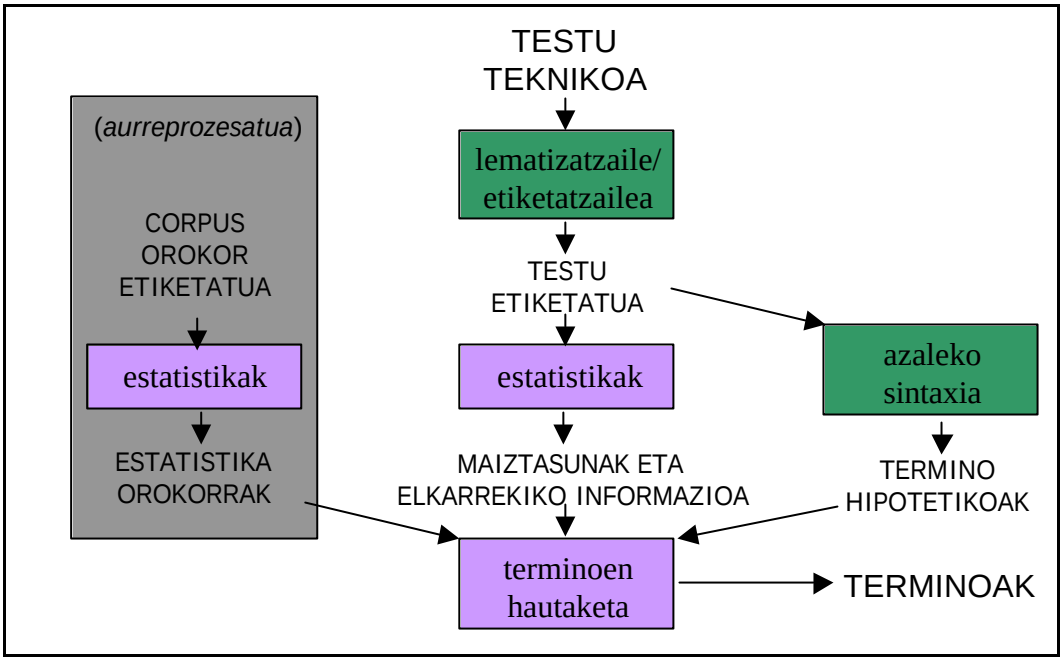
Hain trinko izan gabe honako hauek dira patroi ohikoenak:

```
Nnc N | Aprep Nnc? N | Nnc Nnc? Apos

Nnc? Nabs V | ADV V
```

Hau horrela izanik, patroiak bilatzeko testua EUSLEMetik pasatzen da eta ondoren espresio erregularretako tratamendutik (aipatutako Xeroxeko *xfst* programan oinarrituta). Horrela lortzen dira hautagai diren terminoak. Hautagai hauen artean sailkapen probabilistikoa egin behar da, baina urrats hau garatu gabe dago.

1 N = izena;
Nnc = izena kasu-markarik gabe;
Nabs = izena absolutiboan;
Aprep = adjektibo izenlaguna (edo izenlagunaren funtzioa duen multzoa)
Apos = adjektibo izenondoa;
ADV = adberbioa;
V = aditza.



VII.3 irudia.- Terminoak erauzteko erabilitako arkitektura

Dena den, eginkizun hau planifikatuta dago eta neurri estatistiko sinpleetan oinarrituko da lehen urrats batean: maiztasun absolutua, erlatiboa eta elkarrekiko informazioa konbinatuz lortuko dena. Sistema arkitektura orokorra VII.3 irudian azaltzen da.

VII.3 Hizketaren tratamendua

Hizketaren tratamenduari dagokionean, bi eremu bereizten dira: batetik, hizketaren sorkuntza, testu batetik abiatuta testu hori "irakurtzen" duen tresna garatzean datzana eta, bestetik, hizketaren ezagumendua, alderantzizko prozesua burutzen duena, hau da, pertsona batek esandakoa prozesatu eta testu bihurtzen du.

Hizketaren sorkuntzaren aplikazioak anitzak dira, besteak beste, informazio-gune elkarreragileak, bidaietako txartelen erosketarako sistemak, edota posta elektronikoa irakurtzen duten tresnak. Ezagumenduari dagokionean, guztiok ezagunak ditugun aplikazioetatik hauek aipa daitezke: telefono-sistemak, menuak ahots bidez aukeratzeko eta eskaerak egiteko balio dutenak; telefono eramangarrietako katalogoan dauden zenbakiak ahotsaren bidez aukeratu eta markatzeko balio dutenak; eta, nola ez, testua mikrofono baten aurrean esan eta testu-prozesadore baten dokumentu bihurtzeko tresnak.

VII.3.1 Hizketaren sorkuntza

Testuekin eta hizketarekin lan egiten duten EHUko bi talde, *Ixa* eta *Aholab*, bildu ginen proiektu honen inguruan:

- *Euskarazko testu-ahots bihurketa prozesuan analisi morfosintaktikoa barneratzea. UPV/EHU. 1999-2000.*

Helburua IXA taldeko tresnek, eta bereziki EUSLEMek, eskaintzen duten informazioa erabiltzea testu-ahots euskarazko sistemen ezaugarriak hobetzeko. Esperimentu desberdinak egin dira, eta intonaziorako emaitzek EUSLEM baino haratago doan zerbaiten beharra (azaleko sintaxia, adibidez) adierazten badute ere, isiluneen detekziorako oso emaitza onak lortu dira (Navas *et al*. 2002).

Isiluneen detekziorako sailkapen-zuhaitz bitarrak erabili dira eta ezaugarrien artean hitzen kategoriak hartu dira kontuan. Beraz, ikasketa-corpusa zein ebaluaziokoa EUSLEMetik pasatu dira sailkapen-zuhaitzak erauzi eta aplikatu ahal izateko. Dena den, etiketen egokitzapen bat egin da, gehienbat 1. mailakoak erabiliz, baina loturazkoetan 2. maila ere kontuan hartuz. Erabilitako kategorien zerrenda azaltzen da VII.1 taulan.

Etiketa	**Deskribapena**
ADB	Adberbioa
ADI	Aditz nagusia
ADJ	Adjektiboa
ADL	Aditz laguntzailea
ADT	Aditz trinkoa
BEREIZ	Puntuazio-marka berezia
DET	Determinantea
IOR	Izenordaina
ITJ	Interjekzioa
IZE	Izena
LOT_JNT	Juntagailua
LOT_LOK	Lokailua
LOT_MEN	Menpeko lotura
PUNT	Puntuazio-marka

VII.1 taula.- Isiluneen sorkuntzan erabilitako etiketak

Ebaluazioak frogatzen du zehaztasuna %92 baino handiagoa dela. Dena den, erroreek sortzen duten ondorioen arabera azterketa sakonagoa eginez bi errore-mota bereiztu dira: saihestekoak eta onargarriak. Lehenak gehiago gertatzen dira isilune desegokiak txertatzen

direnean (%84) eta gutxiago ipini beharreko isiluneak ez direnean jartzen (%15). Hori dela eta, ikasketa-zuhaitz berri bat eraiki da eta emaitzak orokorrean antzekoak izan arren, saihesteko errore gutxiago agertzen da.

Kontuan hartu behar da esperimentu guztiak etiketatze-erroreak zuzendu gabe egin direla.

VII.3.2 Hizketaren ezagumendua

Hizketaren ezagumendurako sistemen garapenerako ezinbesteko pausua unitate lexikalen hautapena da. Unitate lexikal hauek hizkuntz eredua osatzeko ez ezik, eredu akustiko-fonetikoak bil ditzaketen hiztegiak definitzeko ere erabiliko dira.

Unitate lexikal klasikoa hitz-forma da, hitzetan oinarrituta ereduak erabiliko duen hiztegia eratzen duena. Baina esaldi barruan hitzak modu argian banatuta ez dauden hizkuntzetan, japoniera kasu, edota hitz barruan nolabaiteko egitura dutenetan, alemanera, suomiera edo euskara kasu, hitz-forma ez den beste unitateren bat erabili behar da, neurriko hiztegia eratzeko eta ereduak osatzeko egokiagoa den unitatea erabiltzeko.

Proposamenen artean unitate gisa morfemak erabiltzea dago. Euskararen kasuan, unitate egokia izan daitekeelakoan, hainbat saiakuntza egin dira MORFEUSen oinarriturik. Dena dela, segmentazio morfologikoa, hurbilpen gisa egokia dirudien arren, morfemen sekuentzietan hizketan gauzatzen ez diren osagaiak agertzen dira.

Adibidez, *osabei* hitzaren segmentazioa *osaba+ei* izango litzateke, baina tarteko *a* hori ez da ahoskatzen. Bestalde, aldaketa fonologikoak gertatzen dira morfema-mugetan eta, ondorioz, esaten denarekin alderatuta oso desberdina den segmentazioa ere ager daiteke, *honi* hitzean bezala, *hau+i* segmentazio morfologikoa duena. Beraz, segmentazio morfologikoari hainbat transformazio-erregela aplikatzen zaizkio, hizketaren tratamendurako unitate egokiak lortzearren.

VII.2 taulan hitz-formak erabilita eta morfema egokitu hauek erabilita hainbat corpusekin lortzen diren hiztegien tamaina aurkezten dira. Lehenengo corpusa, EEBSko ia 200.000 hitz dira, bigarrena, *Euskaldunon Egunkaria*ko 167.000 hitz eta hirugarrena, berriz, ETBko *Gaur Egun* berrietako 180.000 hitz.

	hitzak	**morfemak**
EEBS	50.121	20.117
Euskaldunon Egunkaria	38.696	15.302
ETBko berriak	41.085	17.983

VII.2 taula.- Hiztegien tamaina unitate lexikal desberdinekin.

Argi ikus daiteke hiztegiaren tamaina %60 inguru txikiagoa dela morfemak erabilita. Dena dela, morfema horietatik %10 lau ikur baino laburragoak dira eta euren artean %40ko maiztasun metatua biltzen dute. Eta, azkenik, oso laburrak diren morfemen artean, asko akustikoki oso antzekoak dira eta, kasu askotan, hitz mugetan agertzen dira (*ak*, *ek*, *ok*, *ko*, *tu*, *du*...). Horien maiztasun metatua %25ekoa da eta oso zaila gertatzen da horiek bereiztea.

Modu honetan, hainbat hobekuntza lortzen dira. Lehenengo, hitzen hiztegia erabiliz gero, hiztegian ez dauden hitzen kopurua (*out of vocabulary*) gehiegizkoa litzateke. Gainera, ikasketa burutzean, bigramen estatistikak kalkulatzeko askoz corpus handiago beharko litzateke, estatistika esanguratsuak lortzeko unitate bakoitzeko agerpen kopuru handiagoa izatearren. Hala ere, corpus handiagoak edukita ere, datu-sakabanaketaren arazoari aurre egin beharko litzaioke (*data sparseness*), desanbiguaziorako hizkuntz ereduetan gertatzen den bezala. Azkenik, hiztegi-tamaina txikiagoa erabilita ikasketa-prozesua asko azkartzen da, unitateen arteko nahasketak ere gutxiago dira, eta, ondorioz, emaitzak hobetzen dira.

Lan honen inguruko informazio zehatzagoa lortzeko (López de Ipiña *et al.* 2000, 2002-ab) argitalpenak kontsulta daitezke.

VIII Ondorioak eta zabaldutako ikerlerroak

VIII.1 Ondorioak

Ikerlan honen emaitza gisa euskararen prozesamendu automatikorako oinarrizkoa den tresna modu arrakastatsuan diseinatu eta garatu da, EUSLEM lematizatzaile/etiketatzailea, euskararen tratamendurako lehena dena.

Tresna hau, IXA taldeak garatutako gainerakoak bezalaxe, euskararen prozesaketa automatikorako egitasmo orokor baten barruan kokatzen da, eta bide horretan aurrera egiteko oinarrizko tresnatzat jo daiteke, analisi sintaktikoa, terminologiaren erauzketa, informazioaren erauzketa eta berreskurapenaren abiapuntua delarik.

Hori dela eta, erabilera orokorreko tresna garatu da, bai diseinuan bai garapenean ere euskara estandarra prozesatzeko helburuari jarraituz. Bere ezaugarri nagusiak hiru dira: sendotasuna, estaldura eta egokitzeko gaitasuna.

Etiketatzaileari sendotasuna emateko, diseinu eta garapenerako baliabide neutroak erabili dira, aplikazioetatik independenteak. Erabilitako baliabide lexikoak Euskararen Datu-Base Lexikalean (EDBL) landutakoa dira. Helburu orokorrekoak datu-basea izanik, euskararen tratamenduaren urrats ezberdinetan aplikatzeko diseinatu eta aberastutako baliabidea da.

Etiketa-sistema diseinatzerakoan ere, EDBLn oinarritutako analisi morfologikoaren informazioa modu inkrementalean gehituz eratutako lau maila definitu dira. Hau egiteko, aplikazioetan beharrezko edo interesgarri gerta zitekeen informazioa aztertu da, desanbiguazio-atazari lagunduko ote dion begiratu gabe, orokortasuna bermatu asmoz.

Bestalde, tresnaren garapenerako erreferentzia izan den corpusa aukeratzean ere, espezializatu gabe nahiko orokorra izan zedin EEBS corpus orekatuaren atal bat erabili da.

Horretan oinarriturik, euskararen murriztapen-gramatika garatu da, eta EUSLEMen ebaluazioan ikus ahal izan den bezala, EEBSkoak ez diren testuak ere oso modu egokian desanbiguatzeko gai den tresna garrantzitsua.

Hala ere, testu espezializatuen prozesamendua burutu nahi denean, orokortasun hori dela medio, lexiko aldetik hutsune nabarmena dago. Hori dela eta, erabiltzaileen lexikoak gehitzea ezinbestekotzat jo dugu. Lexikoa eskuz edota modu erdiautomatikoan egin daiteke, eta mantentzen erraza da, beste hizkuntza askotan gertatzen ez den bezala, hitz-forma guztiak sartzearen beharrik ez dagoelako. Modu horretan, tresna hizkuntz espezializatuetara modu nahiko errazean egokitzeko aukera ematen du.

Horri guztiari esker, erabateko estaldura lortu da, tokenizazio-akats batzuk ekarritako prozesaketa okerra alde batera utzita. Hori lortzeko garrantzitsua izan da aldaera eta desbideratzeen tratamendua burutzea eta, lexikoan agertzen ez diren hitzak modu egokian tratatzea.

Horretarako, ebaluazioan argi ikusi den bezala, aurrerapauso nabarmena eman da ikerkuntza-lan honen ekarpena den hitz ez-estandarren tratamenduari esker. Hitz ezezagunen emaitzen hobekuntza bereziki interesgarria da. Batetik, desanbiguazioaren emaitza orokorraren kalitatea hobetzen laguntzen du, eta bestetik, lematizazio eta etiketatzean oinarritutako aplikazioen kalitatea ere areagotzeko balio izango du.

Desanbiguazioaren emaitzak hobetzearren ere, analizatzaile morfologikoaren emaitzak nola hobetu aztertu da. Baina, zuzentasun aldetik nahiko maila altua izanik, zaila gertatzen zen anbiguotasuna gehiegi areagotu gabe, erroreak ekiditea. Dena dela, anbiguotasuna modu orekatuan gehituta, hasierako analizatzailearen erroreen erdia inguru ekiditea lortu da. Analizatzaile morfologiko hedatua ere tesi-lan honen ekarpena da.

Aipatzekoa da hitz anitzeko unitateen tratamenduaren garrantzia, emaitzak areagotzen laguntzeaz gain, aplikazioei begira oso interesgarria delako. Aldez aurretik garatuta zegoen hitz anitzeko unitate lexikalen tratamendua, HABIL, EUSLEMen integratu ez ezik, bere estaldura areagotu ere egin da, data eta zenbakien tratamendua gehituz.

Desanbiguazio-atazari dagokionean, hasierako hipotesia egiaztatu da, hau da, metodo linguistiko eta estatistikoen bidezko konbinaketa emaitzetan oso onuragarria dela. Kontuan hartu behar da desanbiguazioaren arrakasta hasierako anbiguotasunak baldintzatzen duela. Eragozpen horri metodo estokastikoen bitartez aurre egiteko ikasketa-corpusaren tamaina txikia erabakiorra da, baina murriztapen-gramatikari esker anbiguotasuna modu esanguratsuan jaisten denez, emaitza onak lortu dira. Hala ere, 3. mailako emaitza egokiak lortzeko bi arlo lantzea eskatzen du, batetik, corpusaren tamaina handitzea, eta bestetik testua are gehiago desanbiguatzearren erregelak fintzea.

Horrekin guztiarekin batera, IXA taldean tresnen sarrera/irteerak SGML/XMLz etiketatzeko proposamena diseinatu eta gauzatu da, tartean, MORFEUS eta EUSLEMen interfazeak inplementatu direlarik.

Horrez gain, tresna guztien emaitzen ebaluazio orokor sakona burutu da, aurreikusitako arazo guztien aurrean planteatutako soluzioak egokiak diren neurtzeko. Ebaluazioaren ondorioz, emaitzak beste hizkuntzetakoekin konparagarriak direla esan daiteke. Gainera, ebaluazioa borobiltzeko aplikazio errealetan probatua izan da, egokitasun hori berretsi delarik.

Azkenik, etiketatze-prozesua prestatzeko eman beharreko urratsei buruzko bibliografia-bilketa interesgarria delakoan gaude.

VIII.2 Zabaldutako ikerlerroak eta etorkizuneko lanak

Lan honen ondorioz zabaldutako ikerlerroak anitz dira. Alde batetik lanaren alde ahulenak hobetzeko bideak eta lanean zehar hausnartutako etorkizuneko hobekuntza posibleak daude. Beste aldetik, proiektu zabalago batean integratuta egotetik datozen ikerlerroak daude, egindako tresnak beste urratsetan oinarri gisa erabiliko baitira. Aldez aurretik esan behar da ez zaizkigula denak berdin interesatzen, eta zenbaitetan dagoeneko lanean hasiak garen bitartean, beste batzuk aipatu besterik ez ditugu egingo.

Aurkezteko orduan hiru multzotan banatu ditugu etorkizuneko lan hauek: lehenengoan, ikerlan honen hobekuntzarekin zuzenean lotutako gaiak hartzen dira mintzagai; bigarrenean, lematizatzailea/etiketatzailearen egokitzapena burutzeko eman beharreko pausoak aurkezten dira; eta, azkenik, etorkizunerako lanen artean, desanbiguazio-eredu berriak eta egindako tresnak oinarritzat hartuko dituzten aplikazioak aipatzen dira.

VIII.2.1 Lematizatzaile/etiketatzailearen hobekuntza

Ebaluazioari buruzko atalean aipatu den bezala, lematizatzaile/etiketatzailearen emaitzak onak direla esan daiteke eta, erabiltzen ditugun metodoen bitartez behinik behin, emaitzak hobetzea nahikoa zaila gerta daitekeela uste dugu, batez ere 2. mailakoak. Hala ere, oraindik hobekuntzarako hainbat bide geratzen dira eta horiek jorratzea izango da aurrerantzean helburuetako bat.

Lehenik eta behin, tesi-lanean zehar hainbatetan adierazi den bezala, analizatzaile hedatuak sortutako anbiguotasunak era egokian ebazteko murriztapen-gramatikaren erregeletan ukituak egitea komeni da, bereziki izen bereziei dagokien erregeletan. Hartara, erroreak gutxiarekin batera, besteak beste, izen berezi konposatuen identifikazioa ere erraztuko da.

Bestalde, orokorrean desanbiguazioaren emaitzak areagotzeko erregela berriak diseinatzea komeni da, batez ere kategoria itxietako hitzetarako. Egiaztapen-corpuseko akatsak aztertzean, errore asko loturazko elementu eta adberbioetan gertatzen zirela ikus dugu. Horrelakoak desanbiguatzeko etiketa ez da nahikoa izaten, izan ere, osagai guztiek ez dute portaera bera, eta lemaren informazioa oso esanguratsua izan daiteke etiketa posibleen artean aukeratzeko.

Esan gabe doa, erreferentzia- eta egiaztapen-corpusak handitu egin behar direla, alde batetik, desanbiguazio estokastikorako ikasketa-corpusa esanguratsuagoa izan dadin, hau da, anbiguotasun-klase eta etiketa gehiagoren agerpenak eduki ditzan, eta, beste aldetik, gainerako tratamenduen garapen eta ebaluaziorako testu-masa handiagoa izateko.

Azkenik, hitz anitzeko unitateen tratamenduaren aplikazioak ikerlerro berriak ireki ditu. Batetik, unitate gehiago lantzea interesgarria izan daitekeela uste dugu, gehienbat 3. mailako etiketatzeari begira. Beraz, maiztasun handienekoak identifikatzeko tresnaren baten bitartez, interesgarrienak gerta daitezkeenak soilik lantzea izango litzateke hitz anitzeko unitateen tratamendua osatzeko modurik egokiena. Gainera, egiteke geratu den izendun entitateen tratamenduari dagokionean, garapen fasean egonik, etorkizun hurbilean integratzeko moduan izango da. Bukatzeko, HAUL anbiguoen integrazioa burutzea geratzen da.

VIII.2.2 Lematizatzaile/etiketatzailearen egokitzapena

Esan den bezala, EUSLEMen ezaugarrietako bat egokitzeko gaitasuna da. Egokitzapen hori bi ikuspegi desberdinetatik buru daiteke, euskalkien ikuspegitik, eta hizkuntza teknikoaren ikuspegitik.

Egia da analizatzaile morfologikoak aldaera dialektal batzuk tratatzen dituela, baina prozesaketa hori maila lexikora mugatzen da. Ondorioz, euskara batua ez diren aditzak agertzen direnean, interpretazioak jasotzen dituzte, analizatzaile sendoa delako, okerrak, ordea. EEBS corpuseko euskalkietako testuak modu egokian tratatzeko, lehenengo egiaztatzaile ortografikoa aplikatu zitzaien testuei, eta ontzat ematen ez ziren hitzak UZEIn aztertu ziren, hiztegi berezitua osatuz. Hala ere, agertutako adizkiak besterik ez ziren landu. Gainera, corpus diakronikoak analizatzeko behar hori areagotu egiten da.

Beraz, euskara batuarekin egin zen moduan, euskalkien lanketa zehatza burutu beharko litzateke. Lematizatzaile/etiketatzailea egokitzeko, ordea, horretaz gain, gutxienez hitz ez-estandarren tratamendua eta desanbiguazio-erregelak errepasatu beharko lirateke. Gainera, euskalki bakoitzeko ikasketa-corpus bat biltzeak ere lagunduko luke egokitzapena doitzen.

Hizkuntza espezializatu eta teknikoa tratatzeko ere, bai maila lexikoan bai corpus mailan egokitzapena burutzeak emaitzak hobetzen lagunduko luke. Horretarako, terminologiaren erauzketa automatikoaren emaitzak oso lagungarriak izan daitezke. Horrela, etiketatzailearen aplikazioak elika dezake etiketatzailearen lexiko teknikoa.

VIII.2.3 Etorkizunerako lanak

Etorkizunari begira, bi bide jorratuko dira bereziki. Batetik, beste eredu batzuk aplikatzea interesgarri jotzen dugu; bestetik, EUSLEM oinarri gisa harturik garatuko diren tresnak aipatuko dira.

Desanbiguazio estokastikoaren inplementazio hau aplikatzeak bi arazo nagusi sortu dizkigu. Lehenengoa, anbiguotasun-klase eta etiketa guztiak ikasketa-corpusean ez agertzeak probabilitateen berrikasketa suposatzen du. Hori ekiditeko, *smoothing* edo leuntze-teknikaren bat aplikatu behar da ezinbestean. Horretarako, MULTEXTen inplementazioan leuntze-funtzioak integra ditzakegu edota leuntzea burutzen duen beste inplementazioren bat erabil dezakegu.

Etiketa-sistema diseinatzerakoan, kontuan hartutako ezaugarrietako bat elipsiarena izan da, baina lan honetan aurkeztu diren saiakuntzetan ez da elipsia etiketan islatu. Gure ustez interesgarriago litzateke elipsidun analisiak banatzea, etiketa-sekuentzian izen bat agertzen dela esplizituki agertzeko. Ideia hau turkierarako inplementatu den desanbiguatzaile morfologiko estatistikoan (Hakkani-Tür *et al*. 2000) aplikatu da era orokorragoan, baina gure kasuan lehen hurbilpen gisa elipsiaren tratamendurako erabili nahi dugu.

Hala ere, badirudi Markov-en eredu ezkutuen bidezko desanbiguazioa goi-muga jotzear dagoela, 2. mailan behik behin. Desanbiguazio morfosintaktikoari buruzko atalean, hainbat teknika aztertu dira, metodo horiek saiatzea ikerkuntza-lan honetatik kanpo geratu den arren. Hala ere, horietako batzuk euskararen desanbiguaziorako Markov-en eredu ezkutuak baino egokiagoak izan daitezke, batez ere ikasketa automatikoan oinarritutakoak. Erabaki-zuhaitzetan oinarritutako etiketatzaileak bereziki interesgarriak iruditzen zaizkigu, baina gainerako aukerak baztertu gabe.

EUSLEMen oinarritutako tresnen artean, azaleko analisi sintaktikoa aipatu nahi dugu. Dagoeneko garatzen ari den tresna hau oso garrantzitsua da, analisi sintaktiko osoa izan gabe

ere diseinatu eta inplementatu ahal diren tresnei begira. Horien artean batzuk aipatzekotan informazioaren erauzketaren inguruko atazak, hizketaren sorkuntza, terminologiaren erauzketa eta itzulpen automatikoa ditugu, guztiak ere, garapen-maila handiagoan edo txikiagoan, IXA taldeak esku artean dituen proiektuak.

Bibliografia

(Aberdeen *et al*. 1995) Aberdeen J., Burger J., Day D., Hirschman L., Robinson P., Vilain M. Mitre: description of the Alembic system used for MUC-6. *Proceedings of Message Understanding Conference (MUC-6)*. 1995.

(Abney 1996) Abney S. Part-of Speech Tagging and Partial Parsing. K. Church, S. Young and G. Bloothooft, eds. *Corpus-Based Methods in Language and Speech.* ELSNET book, Kluwer Academic Publishers, Dordrecht. 1996.

(Aduriz 2000) Aduriz I.. *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz*. Tesia. Euskal Herriko Unibertsitatea. 2000.

(Aduriz *et al*. 1993) Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M. A Morphological Analysis Based Method for Spelling Correction. *Proceedings of EACL'93*. 463. 1993.

(Aduriz *et al*. 1994) Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Da Costa A., Diaz de Illarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M. Xuxen-Mac: un corrector ortográfico para textos en euskara. *Proceedings of UNIMAC-94* . 1994.

(Aduriz *et al*. 1994) Aduriz I., Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. Urkia M. EUSLEM: Un lematizador/etiquetador de textos en euskara, *Actas X. SEPLN Cordoba*. 1994.

(Aduriz *et al*. 1995) Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. Different issues in the design of a lemmatizer/tagger for Basque. *Proceeding of ACL SIGDAT Workshop "From texts to tags: Issues in Multilingual Language Analysis", EACL'95*. 1995.

(Aduriz *et al*. 1996-a) Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. "EUSLEM: A lemmatiser/tagger for Basque" *Proceedings of the EURALEX'96*. 1996

(Aduriz *et al*. 1996-b) Aduriz I., Aldezabal I., Alegria I., Ezeiza R., Urizar R. Del analizador morfológico al etiquetador/lematizador: unidades léxicas complejas y desambiguación. *Actas del XII SEPLN. Sevilla*. 90-100. 1996.

(Aduriz *et al*. 1996-c) Aduriz I., Aldezabal I., Artola X., Ezeiza N. and Urizar R. Multiword lexical units in EUSLEM, a lemmatiser-tagger for Basque. *Proceedings of COMPLEX*. 1-8. 1996.

(Aduriz *et al*. 1996-d) Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. A corpus based morphological disambiguation tool. *Actas XII .SEPLN Sevilla*. 41-50. 1996.

(Aduriz *et al*. 1997) Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. Morphosyntactic disambiguation for Basque based on the Constraint Grammar formalism. *Proceedings of RANLP*. 1997.

(Aduriz *et al*. 2000) Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. A Word-grammar based morphological analyzer for agglutinative languages. *Proceedings of the International Conference on Computational Linguistics (COLING-2000)*. 2000.

(Aduriz eta Aldezabal 1995) Aduriz I., Aldezabal I. 1995. Eratorpenak eragindako aldaketak. Barne-txostena.UPV/EHU-LSI-TR ?-95. 1995.

(Agirre *et al*. 1989) Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola K., Urkia M. Aplicación de la morfología de dos niveles al euskara, *SEPLN*, vol. 8, 87-102. 1989.

(Agirre *et al*. 1991) Agirre E., Agirre A., Alegria I., Arregi X., Artola X., Diaz de Illarraza A,.Goenaga P., Maritxalar M., Sarasola K., Urkia M. Bi mailatako morfologiaren euskararako egokitzapena, *Elhuyar*, 17, 6-14. 1991.

(Agirre *et al*. 1992) Agirre E., Alegria I., Arregi X., Artola X., Diaz de Illarraza A,. Maritxalar M., Sarasola K., Urkia M. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology, *Proceedings of ANLP'92*. 119-125. 1992.

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

(Agirre *et al*. 1994-b) Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K. Conceptual Distance and Automatic Spelling Correction. *Proceedings of the Workshop on "Computational Linguistics for Speech and Handwriting Recognition"*. 1994.

(Agirre *et al*. 1995) Agirre E., Arregi X., Arriola J.M., Artola X., Diaz de Illarraza A., Insausti J.M., Sarasola K. Different issues in the design of a general-purpose Lexical Database for Basque. *First workshop on application of Natural Language to Data Bases, NLDB'95*. 1995.

(Aizpurua *et al*. 2000) Aizpurua I., Alegria I., Ezeiza N. GaIn: un buscador Internet/Intranet avanzado para textos en euskera. *Actas del XVI Congreso de la SEPLN* Universidade de Vigo, 26-28 septiembre de 2000.

(Aldezabal *et al*. 1997) Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. Terminologiaren erauzketa automatikoa eta bere aplikazioa euskararako. *Nazioarteko Terminologia biltzarra*, 495-508 orr. Donostia. Azaroaren 12-13-14, 1997.

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

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

(Aldezabal *et al*. 1999-c) Aldezabal I., Gojenola K., Oronoz M. Combining Chart-Parsing and Finite State Parsing. *Proceedings of the European Summer School in Logic, Language and Information (ESSLLI) Student Session 99*, Utrecht, The Netherlands. August 16-20, 1999.

(Aldezabal *et al*. 2002) Aldezabal I., Alegria I., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Hernández G., Mayor A., Oronoz M. Soroa A. *Hizkuntza prozesatzeko tresnen integrazio, SGML erabiliz*. Barne-txostena. UPV/EHU LSI / TR 2-2002. 2002.

(Alegria 1995) Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak.* Doktoretza-tesia. Euskal Herriko Unibertsitatea. 1995.

(Alegria *et al*. 1996) Alegria I., Artola X., Sarasola K., Urkia M. 1996. Automatic morphological analysis of Basque. *Literary & Linguistic Computing*. 11(4). 193-203. Oxford University Press. 1996.

(Alegria *et al*. 1997) Alegria I., Artola X., Sarasola K. Improving a Robust Morphological Analyser using Lexical Transducers. *Recent Advances in Natural Language Processing. Current Issues in Linguistic Theory (CILT) series.* John Benjamins publisher company. Ruslan Mitkov and Nicolas Nicolov editors. Vol. 136. 97-110. 1997.

(Alegria *et al*. 1999) Alegria I., Ezeiza N., Oronoz M., Urizar R. Extracción Automática de Terminología a partir de Etiquetado y Lematización. *VI Simposio Internacional de Comunicación Social.* Santiago de Cuba, 25-28 de Enero de 1999.

(Alegria *et al*. 2001) Alegria I., Aranzabe M., Ezeiza A., Ezeiza N, Urizar R. 2001. Using Finite State Technology in Natural Language Processing of Basque. *Proceedings of the International Conference on Implementations and Applications of Automata.* 2-12. 2001.

(Alegria *et al.* 2002-a) Alegria I., Aranzabe M., Arregi O., Ezeiza A., Ezeiza N., Urizar R., Casillas, A. Trabajos en el área de Recuperación de la Información del grupo IXA de la Universidad del País Vasco. *I Jornadas de Tratamiento y Recuperación de Información, JOTRI.* 2002.

(Alegria *et al.* 2002-b) Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. *Proceedings of LREC-2002 Workshop Customizing knowledge in NLP applications*. 2002.

(Alegria *et al*. 2003) Alegria I., Aranzabe M.J., Ezeiza A., Ezeiza N., Urizar R. Robustez y flexibilidad en un lematizador/etiquetador. *VIII Simposio Internacional de Comunicación Social.* Santiago de Cuba, 21-24 de Enero de 2003. 2003. (*Forthcoming*)

(Antworth 1990) Antworth E.L. *PC-KIMMO: A two-level processor for morphological analysis*. Occasional Publications in Academic Computing, No. 16, Dallas, Texas. 1990.

(Aone *et al.* 1996) Aone C. Hausman K. Unsupervised learning of a rule-based Spanish part of speech tagger. *Proceedings of COLING'96*. 53-58 1996.

(Arévalo 2002) Arévalo M. MICE, un recurso para la resolución de la anáfora. *International Workshop on Computational Linguistics*. *http://www.lsi.upc.es/~nlp/iwcl02*. 2002.

(Armstrong *et al*. 1995) Armstrong S., Russel G., Petitpierre D., Robert G. An open architecture for Multilingual Text Processing. *Proceedings of EACL'95*. v1, 101-106. 1995.

(Arrieta *et al*. 2001) Arrieta B., Alegria I. , Arregi X. An Assistant Tool For Verse-Making In Basque Based On Two-Level Morphology. *Literary and Linguistic Computing*, 16(1). Oxford *University press.* 2001.

(Artola *et al*. 2000) Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Soroa A. A proposal for The Integration of NLP Tools using SGML-Tagged documents. *Proceeding of LREC-2000*. 2000.

(Artola *et al*. 2002-a) Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Hernández G., Maritxalar A., Soroa A. *Lengoaia naturalaren prozesamendurako tresnen integraziorako programa-liburutegia.* Barne-txostena. UPV/EHU / LSI / TR 1-2002. 2002.

(Artola *et al*. 2002-b) Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Hernández G., Soroa A. A Class Library for the Integration of NLP Tools: Definition and implementation of an Abstract Data Type Collection for the manipulation of SGML documents in a context of stand-off linguistic annotation. *Proceedings of* LREC-2002. 2002.

(Barton 1985) Barton G.E. On the Conplexity of ID/LP Parsing. *Computational Linguistics,* 11(4). 205-218. 1985.

(Baum 1972) Baum L.E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process. *Inequalities,* 3.1-8. 1972.

(Basili *et al*. 1994) Basili R., Pazienza M.T., Velardi P. A (not-so) shallow parser for collocational analysis. *Proceedings of COLING'94*. 1994.

(Beesley eta Karttunen 2002) Beesley K., Karttunen L. *Finite State Morphology: Xerox Tools and Techniques*. Cambridge University Press. 2002.

(Bikel *et al*. 1997) Bikel D., Miller S., Schwatz R., Weischedel R. Nymble: a High-Performance Learning Name-Finder. *Proceeding of ANLP'97*. 194-201. 1997.

(Bikel *et al*. 1999) Bikel D., Miller S., Schwatz R., Weischedel R. An Algorithm that Learns What's in a Name. *Machine Learning: Special Issue on Natural Language Learning, 34*. 1999.

(Black *et al.* 1987) Black A., Ritchie G., Pulman S and Russel G. Formalisms for morphographemic description. *Proceedings of EACL'87*, 11-16.1987.

(Black *et al.* 1991) Black A., van de Plassche J., Williams B. Analysis of Unknown Words through Morphological Descomposition. *Proceedings of 5th Conference of the EACL*, pp. 101-106. 1991.

(Black *et al*. 1992) Black E., Jelinek F., Lafferty J., Mercer R., Roukos S. Decision Tree Models Applied to the Labeling of Text with Parts-of-Speech. *Proceedings of the DARPA Workshop on Speech and Natural Language Processing*. 1992.

(Black *et al*. 1998) Black W.J., Rinaldi F., Mowatt D. FACILE: Description of the NE System used for MUC-7. *Proceeding of Message Understanding Conference (MUC-7).* 1998.

(Borthwick *et al*. 1998-a) Borthwick A., Sterling J., Agichtein E., Grishman R. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. *Proceeding of Workshop on Very Large Corpora (WVLC'98).* 153-160. 1998.

(Borthwick *et al*. 1998-b) Borthwick A., Sterling J., Agichtein E., Grishman R. NYU: Description of the MENE Named Entity System as Used in MUC-7. *Proceeding of Message Understanding Conference (MUC-7).* 1998.

(Brants 2000) Brants T. TnT—A Statistical Part-of-Speech Tagger. *Proceedings of ANLP-2000.* 2000.

(Breidt *et al.* 1996) Breidt E., Segond F., Valetto G. Local grammars for the description of multi-word lexemes and their automatic recognition in texts. *Proceedings of COMPLEX'96.* 19-28. Budapest 1996.

(Brill 1992) Brill E. A simple rule-based part of speech tagger. *Proceedings of ANLP'92 (ACL),* 152-155. 1992.

(Brill 1995) Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics,* 21(4). 543-565. 1995.

(Brill eta Wu 1998) Brill E., Wu J. Classifier combination for improved lexical disambiguation. *Proceeding of COLING-ACL'98*. 1998.

(Carmona *et al*. 1998) Carmona J., Cervell S., Màrquez L., Martí M.A., Padró L., Placer R., Rodríguez H., Taulé ., Turmo J. An environment for morphosyntactic processing of unrestricted Spanish text. *Proceedings of LREC'98*. 915-922. 1998.

(Carroll 1993) Carroll J.A. *Practical Unification-Based Parsing of Natural Language*. PhD thesis, University of Cambridge, Cambridge, UK, October 1993. Computer Laboratory. Technical Report 314. 1993.

(Cha *et al*. 1998) Cha J., Lee G., Lee J-H. Generalized unknown morpheme guessing for hybrid POS tagging of Korean. *Proceedings of WVLC'98*. 85-93. 1998.

(Chang eta Chen 1993) Chang C.H., Chen C.D. HMM-based part-of-speech tagging for Chinese corpora. *Proceedings of WVLC'93*. 107-120. 1993.

(Chanod 1994) Chanod J.P. *Finite-state composition of french verb morphology*. Xerox MLTT-005. 1994.

(Chanod eta Tapanainen 1994) Chanod J.P., Tapanainen P. *Statistical and constraint-based taggers of French*.Xerox MLTT-016. 1994.

(Chanod eta Tapanainen 1995-a) Chanod J.P., Tapanainen P. Creating a tagset, lexicon and guesser for a French tagger. *Proceeding of ACL SIGDAT Workshop "From texts to tags: Issues in Multilingual Language Analysis", EACL'95*. 1995.

(Chanod eta Tapanainen 1995-b) Chanod J.P., Tapanainen P. Tagging French - comparing a statistical and a constraint-based method. *Proceedings of EACL´95*. 149-156 1995.

(Charniak *et al*. 1993) Charniak E., Hendrickson, Jacobson N., Perkowitz M. Equations for Part-of-Speech Tagging. *Proceedings of National Conference on Artificial Intelligence*. 784-789. 1993.

(Chinchor 1997) Chinchor N. MUC-7 Named Entity Task Definition. Version 3.5. 1997. *http://www.itl.nist.gov/iaui/894.02/related_projects/muc/*

(Church 1988) Church K. W. A stochastic parts program and phrase parser for unrestricted text, *Proceedings of ANLP´88*, 136-143. 1988.

(Church eta Hanks 1989) Church K. W., Hanks P. Word association norm, mutual information and lexicography. *Proceedings of ACL'89*. 1989.

(Church eta Hanks 1990) Church K. W., Hanks P. Word association norm, mutual information and lexicography. *Computational Linguistics*, 16(1). 22-29. 1990.

(Church eta Mercer 1993) Church K. W., Mercer R.L. Introduction to the special issue on Computational Linguistics using large corpora, *Computational Linguistics*, 19(1). 1993.

(Ciravegna *et al*. 1999) Ciravegna F., Lavelli A., Mana N., Matiasek J., Gilardoni L., Mazza S., Ferraro M., Black W.J., Rinaldi F., Mowatt D. FACILE: Classifying Texts Integrating Patter Matching and Information Extraction. *Proceedings of IJCAI99*. 1999.

(Clausen eta Lyly 1994) Clausen U., Lyly E. Criteria for identifying and representing idioms in a phraseological dictionary. *The way words work/combinatorics, EURALEX'94*. 258-262 1994.

(Collins 2002-a) Collins M. Dsicriminative Training Methods for Hidden Markov Models: Theory and Experiments with the Perceptron Algorithm. *Proceedings of EMNLP-2002*. 2002.

(Collins 2002-b) Collins M. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Percetron. *Proceedings of ACL-2002*. 2002.

(Collins eta Duffy 2002) Collins M., Duffy N. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron.

(Collins eta Singer 1999) Collins M., Singer Y. Unsupervised Models for Named Entity Classification. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Workshop on Very Large Corpora (EMNLP-VLC-99)*. 1999.

(Cowie 1995) Cowie J. Description of the CRL/NMSU Systems used for MUC-6. *Proceedings of 6th Message Understanding Conference (MUC-6)*. 1995.

(Cowie *et al.* 1985) Cowie A.P. Mackin R., McCraig I.R. *Oxford Dictionary of Current Idiomatic English.* v2. Oxford University Press. Paperback edition, 1985.

(Cussens 1997) Cussens J. Part-of-Speech Tagging Using Progol. *Proceedings of the International Conference on Inductive Logic Programming ILP'97*. 1997.

(Cutting *et al.* 1992) Cutting D., Kupiec J., Pedersen J., Sibun P. A practical part-of-speech tagger. *Proceeding of ANLP'92*. 133-140. 1992.

(Cutting 1994) Cutting D. Porting a stochastic part-of-speech tagger to Swedish. In R. Eklund (ed.), *Proceedings 9:e Nordiska Datalingvistikdagarna, Sockholm 3-5 June 1993.* Department of Linguistics, Computational Linguistics, Stockholm University, Stockholm. 65-70. 1994.

(Cucchiarelli *et al*. 1998-a) Cucchiarelli A, Luzi D., Velardi P. Automatic Semantic Tagging of Unknown Proper Names. *Proceedings of COLING-ACL'98*. 286-292. 1998.

(Cucchiarelli *et al*. 1998-b) Cucchiarelli A., Luzi D., Velardi P. Using Corpus Evidence for Automatic Gazetteer Extension. *Proceedings of LREC'98*. 1998.

(Daciuk 2000) Daciuk J. Finite State Tools for Natural Language Processing. *Proceedings of the COLING 2000 workshop Using Toolsets and Architectures to Build NLP Systems.* 2000.

(Daciuk 1999) Daciuk J. Treatment of Unknown Words. *Proceedings of the Workshop on Implementing Automata (WIA'99)*. 1999.

(Daciuk *et al*. 1998) Daciuk J., Watson B., Watson R. Incremental Construction of Minimal Acyclic Finite State Automata and Transducers. *Proceedings. of the International Workshop on Finite State Methods in NLP*. 1998.

(Daelemans 1996) Daelemans, W. Abstraction Considered Harmful: Lazy Learning of Language Processing. *Van den Herik, J. and T. Weijters (eds.) Benelearn-96. Proceedings of the 6th Belgian-Dutch Conference on Machine Learning.*3-12. 1996.

(Daelemans *et al*. 1996) Daelemans W., Zavrel J., Berck P., Gillis S. MBT: A Memory-Based Part of Speech Tagger-Generator. *Proceeding of WVLC'96*. 14-27. 1996.

(Darroch eta Ratcliff 1972) Darroch N., Ratcliff D. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics.* 43(5). 1470-1480. 1972.

(de Rose 1988) de Rose S. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1). 31-39. 1988.

(Dermatas eta Kokkinakis 1995) Dermatas E., Kokkinakis G. Automatic Stochastic Tagging of Natural Language Texts. *Computational Linguistics*, 21(2). 1995.

(Dias *et al*. 1999-a) Dias G., Guilloré S, Lopes G. Language independent automatic acquisition of rigid multiword lexical units from unrestricted corpora. *Traitement Automatique des Langues Naturalles*. 12-17. 1999.

(Dias *et al*. 1999-b) Dias G., Guilloré S, Lopes G. Mutual Expectation: a measure for multiword lexical unit extraction. *Venezia per il trattamento automatico delle lingue*. 22-24. 1999.

(Dias *et al*. 2000-a) Dias G., Guilloré S., Lopes G. Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? *Proceedings of Recherche d'Informations Assiste par Ordinateur (RIAO2000)*. 2000.

(Dias *et al*. 2000-b) Dias G., Guilloré S, Lopes G. Normalization of Association Measures for Multiword Lexical Unit Extraction. *Proceedings of the International Conference on Artificial and Computationak Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*. 2000.

(Diaz de Ilarraza *et al*. 2000) Diaz de Ilarraza A., Mayor A., Sarasola K. Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System. *Proceedings of Machine Translation 2000.* 19-22. 2000.

(Dice 1945) Dice L. Measures of the amount of ecologic association between species. *Journal of Ecology*. 1945.

(Douglas eta Dale 1992) Douglas S., Dale R. Towards robust patr. *Proceedings of COLING'92*. 1992.

(Dunning 1993) Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1). 61-74. 1993.

(Džeroski *et al*. 2000) Džeroski S., Erjavec T., Zavrel J. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. *Proceedings of LREC-2000*. 1099-1104. 2000.

(Egunkaria 1992) Euskaldunon Egunkaria. *Estilo Liburua*. Donostia, Egunkaria. 1992.

(Eineborg eta Lindberg 1998) Eineborg M., Lindberg N. Induction of Constraint Grammar-rules using Progol. *Proceedings of the International Conference on Inductive Logic Programming (ILP-98)*. 1998.

(Elhuyar 1990) *Munduko leku-izenak*, Elkar. 1990.

(Elhuyar 1993) *Hiztegi entziklopedikoa*, Elhuyar. 1993.

(Elhuyar 1996) *Elhuyar Hiztegia*, Elhuyar. 1996.

(Elhuyar 2000) *Elhuyar Hiztegia*, 2. edizioa. Elhuyar. 2000.

(Elworthy 1993) Elworthy D. Part-of-speech Tagging: A working paper, Acquilex WP No. 10. 1993.

(Elworthy 1994) Elworthy D. Does Baum-Welch re-estimation help taggers?*Proceedings of ANLP´94*, 53-58. 1994.

(Elworthy 1995) Elworthy D. Tagset Design and Inflected Languages. *Proceeding of ACL SIGDAT Workshop "From texts to tags: Issues in Multilingual Language Analysis", EACL'95*. 1995.

(Etxebarria eta Mujika 1987) Etxebarria, J.M. *Euskararen oinarrizko hiztegia. Maiztasun eta Prestasun Azterketa,* Eusko Jaurlaritza. 1987.

(Euskaltzaindia 1973) *Aditz laguntzaile batua*., Euskaltzaindia. Bilbo 1973.

(Euskaltzaindia 1979-a) *Euskal Aditz Batua*. 1979.

(Euskaltzaindia 1979-b) *Euskal Herriko Udalen Izendegia*.Euskaltzaindia. 1979.

(Euskaltzaindia 1979-c) Euskaltzaindiaren Gomendioak eta erabakiak (I eta II), *Euskera* (Separata). 1979

(Euskaltzaindia 1983) *Euskal Izendegia*.1983.

(Euskaltzaindia 1985) *Euskal Gramatika: Lehen urratsak (I eta II)*. Euskaltzaindia. Bilbo 1985.

(Euskaltzaindia 1986) *Maileguzko hitz berriei buruz Euskaltzaindiaren erabakiak*. 1986.

(Euskaltzaindia 1992) *Hitz elkartuen osaera eta idazkera*. Hitz-elkarketa/4. LEF batzordea. 1992.

(Ezeiza 1997) Ezeiza N. *EUSLEM, euskararako lematizatzaile/etiketatzaile baten diseinua eta inplementazioa*.Tesina. Euskal Herriko Unibertsitatea 1997.

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

(Feldweg 1995) Feldweg H. Implementation and evaluation of a German HMM for POS disambiguation. *Proceeding of ACL SIGDAT Workshop "From texts to tags: Issues in Multilingual Language Analysis", EACL'95*. 1995.

(Fellbaum 1999) Fellbaum C. ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. 1999.

(Fine *et al*. 1998) Fine S., Singer Y., Tishby N. The Hierarchical Hidden Markov Models: Analysis and Applications. *Machine Learning*, 32(1). 41-62. 1998.

(Fontenelle *et al*. 1994) Fontenelle T., Adriaens G., De Braekeleer G. The Lexical Unit in the Metal® MT System. *MT*.The Netherlands. v9. 1-19. 1994.

(Francis eta Kucera 1982) Francis W.N., Kucera F. *Frequency Analysis of English usage. Lexicon and grammar*. Hougton Mifflin, Boston, 1982.

(Frantzi eta Ananiadou 1996) Frantzi K.T., Ananiadou S. Extracting Nested Collocations. *Proceedings of COLING'96*, 41-46. 1996.

(Franz 1996) Franz A. A Model for Part-of-Speech Prediction. In D. Fisher and H.-J. Lenz (eds.) *Learning from Data: AI and Statistics V*. Springer-Verlag. 1996.

(Frazier 1995) Frazier L. Processing discontinuous lexical items, by whatever name. *Cognition 54*. 357-359. 1995.

(Gale 1991) Gale W. Concordances for parallel texts. *Proceedings of the 7th Annual Conference UW Center for the New OED and Text Research, Using Corpora*. 1991.

(Garate 1998) Garate G. *27173 Atsotitzak - Refranes - Proverbes – Probervia*. BBK. 1998.

(Garside *et al.* 1987) Garside R., Leech G., Sampson G. *The Computational Analysis of English: A corpus-based approach*. Longman. London, 1987.

(Gojenola 2000) Gojenola K. *Euskararen Sintaxi Konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan.* Doktoretza-tesia. Euskal Herriko Unibertsitatea. 2000.

(Gojenola eta Sarasola 1994) Gojenola K., Sarasola K. Aplicación de la relajación gradual de restricciones para la detección y corrección de errores sintácticos, *Actas X.SEPLN Cordoba*. 1994.

(Graña *et al.* 2001) Graña J., Barcala M., Vilares J. Etiquetación Robusta del Lenguaje Natural: Preprocesamiento y Segmentación. *Procesamiento del Lenguaje Natural*, 27:173-180, 2001.

(Greene eta Rubin 1971) Greene B., Rubin G. *Automatic Grammatical Tagging of English*. Providence: Brown University. 1971.

(Hajic eta Hladká 1997) Hajic J., Hladká B. Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison. *Proceedings of ANLP'97*. 111-118. 1997.

(Hajic eta Hladká 1998) Hajic J., Hladká B. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. *Proceedings of COLING-ACL'98*. 483-490. 1997.

(Hakkani-Tür *et al*. 2000) Hakkani-Tür D., Oflazer K., Tür G. Statistical Morphological Disambiguation for Agglutinative Languages. *Proceedings of COLING-2000*. 2000.

(Haruro eta Matsumoto 1997) Haruro M., Matsumoto Y. Mistake-Driven Mixture of Hierarchical Tag Context Trees. *Proceedings of ACL-EACL'97*. 230-237. 1997.

(Heid 1994) Heid U. On ways words work together - Topics in Lexical Combinatorics. *The way words work/combinatorics, EURALEX'94.* 226-257 1994.

(Hurskainen 1996) Hurskainen. Disambiguation of Morphological Analysis in Bantu Languages. *Proceedings of COLING'96.* 568-573. 1996.

(Ide eta Veronis 1995) Ide N., Véronis J. *Text Encoding Initiative. Background and Context.* Kluwer Academic: Dordrecht, 1995.

(Izagirre 1981) Izagirre K. *Euskal lokuzioak. Espainolezko eta frantsesezko gida-zerrendarekin.* 1981.

(Jackendoff 1997) Jackendoff R. *The Architecture of the Language Faculty.* Cambridge, MA MIT Press. 1997.

(Jarvinen 1994) Jarvinen T. Annotating 200 million words: the bank of English project. *Proceddings of COLING-94*. 1994.

(Karlsson 1990) Karlsson F. Constraint grammar as a framework for parsing running text. *Proceddings of COLING-90*. 1990.

(Karlsson 1992) Karlsson F. SWETWOL: A comprehensive morphological analyser for Swedish. *Nordic Journal of Linguistics*, 15, 1-45. 1992.

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

(Karttunen 1993) Karttunen L. *Finite-State Lexicon Compiler*. Xerox ISTL-NLTT-1993-04-02. 1993.

(Karttunen 1994) Karttunen L. Constructing Lexical Transducers, *Proceedings of COLING´94*, 406-411. 1994.

(Karttunen 2000) Karttunen L. Applications of Finite-State Transducers in Natural Language Processing. *Proceedings of CIAA-2000.* Lecture Notes in Computer Science. Springer Verlag. 2000.

(Karttunen *et al.* 1993) Karttunen L., Chanod J.P., Grenfenstette G., Schiller A Regular Expressions for Language Engineering. *Natural Language Engineering, 2(4)*: 305-328. 1996.

(Karttunen et al. 1996) Karttunen L., Chanod J.P., Grenfenstette G., Schiller A. Regular Expressions for Language Engineering. *Natural Language Engineering*, 2(4). 305:328. 1996.

(Karttunen eta Beesley 1992) Karttunen L. and Beesley K.R. Two-Level Rule Compiler. Technical Report Xerox ISTL-NLTT-1992-2. 1992.

(Kaplan eta Kay 1994) Kaplan R. M., M. Kay . Regular models of phonological rule systems. *Computational Linguistics*, 20(3). 331-380. 1994.

(Kintana 1984) Kintana X.. *Hiztegia 80*. 1984.

(Kintana 2000) Kintana X.. *Hiztegia 2000*. 2000.

(Koskenniemi 1983) Koskenniemi K. Two-level Morphology: A general Computational Model for Word-Form Recognition and Production. Ph.D. thesis, University of Helsinki. Publications n. 11. 1983.

(Koskenniemi 1985) Koskenniemi K. Compilation of Automata from Morphological Two-level Rules. University of Helsinki, *Publication 15*. 1985.

(Koskenniemi *et al*. 1992) Koskenniemi K., Tapanainen P., Voutilainen A. Compiling and using finite-state syntactic rules. *Proceedings of COLING'92*, 156-162. 1992.

(Koskenniemi eta Church 1988) Koskenniemi K. and Church K.W. Complexity, two-level morphology and Finnish. *Proceedings of COLING'88*, 335-340. 1988.

(Krishnamurthy 1996) Krishnamurthy R. The Data is The Dictionary: Corpus at the Cutting Edge of Lexicography. *Proceedings of COMPLEX'96*. 117-144. Budapest 1996.

(Kucera eta Francis 1967) Kucera F., Francis W.N. *Computational Analysis of Present-day American English.* Providence: Brown University. 1967.

(Kupiec 1989) Kupiec J. Probabilistic models of short and long distance word dependencies in running text. *Proceedings of the DARPA Speech and Natural Language Workshop*. 290-295. 1989.

(Kupiec 1992) Kupiec J. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language* N. 6. 225-242. 1992.

(Lee *et al*. 2000) Lee S.-Z., Tsujii J.-I., Rim H.-C. Part-of-Speech Tagging Based on Hidden Markov Model Assuming Joint Independence. *Proceedings of ACL-2000*. 2000.

(Leech *et al*. 1994) Leech G., Garside R., Bryan M. CLAWS4: The tagging of the British National Corpus. *Proceedings of COLING-94*, 622-628. 1994.

(Lezius *et al*. 1998) Lezius W., Rapp R., Wettler M. A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German. *Porceedings of COLING-ACL'98*. 743-747. 1998.

(Lin *et al*. 1994)Lin Y., Chiang T., Su K. Automatic model refinement: with an aplication to tagging, *Proceedings of COLING-94*, 148-152. 1994.

(Lindberg eta Eineborg 1998) Lindberg N., Eineborg M. Learning Constraint Grammar-style disambiguation rules using Inductive Logic Programming. *Proceedings of COLING-ACL'98*. 775-779. 1998.

(Lindberg eta Eineborg 1999) Lindberg N., Eineborg M. Improving Part of Speech Disambiguation Rules by Adding Linguistic Knowledge. *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP'99).* 1999.

(Longman 1989) Longman. *Dictionary of English Idioms.* 1989.

(López de Ipiña *et al*. 2000) López de Ipiña K., Torres I., Oñederra L., Varona A., Ezeiza N., Hernandez M., Peñagarikano M., Rodriguez L.J. First Approach to the selection of lexical units for continuous speech recognition of Basque. *Proceedings of ICSLP/InterSpeech'2000*. Beijing, China. 2000.

(López de Ipiña *et al*. 2002-a) Lopez de Ipiña K., Ezeiza N., Bordel G. Automatic morphological segmentation for continuous speech recognition of Basque. *Proceedings of LREC-2002*. 2002.

(López de Ipiña *et al*. 2002-b) Lopez de Ipiña K., Ezeiza N., Bordel G., Morphological segmentation for speech processing in Basque. *Proceedings of the IEEE Workshop TTS-2002*. 2002.

(Ma *et al*. 1999-a) Ma Q., Uchimoto K., Murata M., Isahara H. Elastic Neural Networks for Part of Speech Tagging. *Proceedings of IEEE-INNS International Joint Conference on Neural Networks (IJCNN).*1999.

(Ma *et al.* 1999-b) Ma Q., Murata M., Utiyama M.,Uchimoto K., Isahara H. Part of Speech Tagging with Mixed Approaches of Neural Networks and Transforation Rules. *Workshop on Natural Language Processing and Neural Netwotks (NLPNN'99)*.1999.

(Ma eta Isahara 1998) Ma Q., Isahara H. A Multi-Neuro Tagger Using Variable Lengths of Contexts. *Proceedings of COLING-ACL'98*. 802-806. 1998.

(Magerman 1995) Magerman D. Statistical decision tree models for parsing. *Proceedings of ACL'95*. 276-283. 1995.

(Magnini *et al*. 2002) Magnini B., Negri M., Prevete R., Tanev H. A WordNet Approach to Named Entities Recognition. *Proceeding of the Workshop SemaNet'02: Binding and Using Semantic Networks.* 2002.

(Manning eta Schütze 1999) Manning C.D., Schütze H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA. 1999.

(Marcus *et al*. 1993) Marcus M., Santorini B., Marcinkiewicz M.A. Building a large annotated corpus of English: the Penn treebank, *Computational Linguistics*, 19(2). 1993.

(Marques eta Lopes 1996) Marques N.C., Lopes G. Using neural networks for portuguese part-of-speech tagging. *Proceedings of the Fifth International Conference on Cognitive Science and Natural Language Processing*. 1996.

(Marques eta Pereira 2001) Marques N.C., Lopes G. A POS-Tagger Generator for Unknown Languages. *Revista de la SEPLN*, 27.199-206. 2001.

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

(Màrquez eta Padró 1997) Màrquez L., Padró L. A Flexible POS Tagger Using an Automatically Acquired Language Model. *Proceedings of ACL-EACL'97*. 238-245. 1997

(Màrquez eta Rodríguez 1997) Màrquez L., Rodríguez H. Automatically Acquiring a Language Model for POS Tagging Using Decision Trees. *Proceedings of RANLP'97*. 27-34. 1997.

(Màrquez eta Rodríguez 1998) Màrquez L., Rodríguez H. Part-of-speech tagging using decision trees. In C. Nédellec and C. Rouveirol (eds.) *Machine Learning: ECML-98*. Lecture Notes in Artificial Intelligence 1398. Berlin: Springer. 25-36. 1998.

(Màrquez *et al*. 1998) Màrquez L., Padró L., Rodríguez H. Improving Tagging Accuracy by Voting Taggers. *Proceedings of NLP+IA/TAL+AI'98*. 149-155. 1998.

(Màrquez *et al*. 1999) Màrquez L., Rodríguez H., Carmona J., Montolio J. Improving POS Tagging Using Machine-Learning Techniques. *Proceedings of*. EMNLP/VLC'99. 149-155. 1999.

(Maritxalar eta Díaz de Ilarraza 1993) Maritxalar M., Díaz de Ilarraza A. *Integration of Natural Language Techniques in the ICALL System Field: The treatment of incorrect knowledgment*. Barne-txostena. EHU/LSI/TR 993. 1993.

(McDonald 1996) McDonald D. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. *Corpus Processing for Lexical Acquisition. J. Pustejovsky and B. Boguraev Eds.* MIT Press. Cambridge MA. 1996.

(Merialdo 1994) Merialdo B. Tagging English text with a probabilistic model, *Computational Linguistics*, 20(2). 155-172. 1994.

(Mikheev 1996-a) Mikheev, A. Unsupervised Learning of Word-Category Guessing Rules. *Proceedings of ACL'96*. 1996.

(Mikheev 1996-b) Mikheev, A. Learning Part-of-Speech Guessing Rules from Lexicon: Extension to Non-Concatenative Operations. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*. 237-234. 1996.

(Mikheev 1996-c) Mikheev, A. Unsupervised Learning of Part-of-Speech Guessing Rules. *Natural Language Engineering*, 2(2). Cambridge University Press. 1996.

(Mikheev 1997) Mikheev, A. Automatic Rule Induction for Unknown Word Guessing. *Computational Linguistics*, 23(3). 405-423. 1997.

(Mikheev 1999) Mikheev A. A Knowledge-free Method for Capitalized Word Disambiguation. *Proceedings of the ACL'99*. 1999.

(Mikheev 2000-a) Mikheev A. Document Centered Approach to Text Normalization. *Proceedings of SIGIR*. 2000

(Mikheev 2000-b) Mikheev A. Tagging Sentence Boundaries. *Proceedings of NAACL*. 2000.

(Mikheev *et al*. 1998) Mikheev A., Grover C., Moens M. Description of the LTG system used for MUC-7. *Proceeding of Message Understanding Conference (MUC-7)*. 1998.

(Mikheev *et al*. 1999) Mikheev A., Moens M., Grover C. Named Entity Recognition without Gazetteers. *Proceeding of EACL'99*. 1999.

(Mitxelena 1987) Mitxelena, K. *Orotariko Euskal Hiztegia*, Euskaltzaindia. 1987.

(Mohri 1997) Mohri, M. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2). 269-322. 1997.

(Moreno-Torres 1994) Moreno-Torres I. A Morphological Disambiguation Tool (MDT): Aplication to Spanish. Universidad de Málaga. 1994.

(Nagao eta Mori 1994) Nagao M., Mori S. A new Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. *Proceedings of COLING'94*, 611-615. 1994.

(Navas *et al*. 2002) Navas E., Hernaez I., Ezeiza N. Assigning Phrase Breaks Using CARTs for Basque TTS. *Proceeding of Speech Prosody*. 2002.

(Neufeld eta Adams 1996) Neufeld E., Adams G. Part-of-Speech Tagging from "Small" Data Sets. In D. Fisher and H.-J. Lenz (eds.) *Learning from Data: AI and Statistics V*. Springer-Verlag. 1996.

(Nivre 2000) Nivre J. Sparse Data and Smoothing in Statistical Part-of-Speech Tagging. *Journal of Quantitative Lingustics*, 7(1), 1-17. 2000.

(Oflazer 1996) Oflazer K. Error-tolerant Finite State Recognition with Applications to Morphological Analysis and Spelling Correction. *Computational Linguistics*, 22(1), 73-89. 1996.

(Oflazer eta Guzey 1994) Oflazer K, Guzey C. 1994. Spelling Correction in Agglutinative Languages. *Proceedings of ANLP-94*. 1994

(Oflazer eta Kuruöz 1994) Oflazer K., Kuruöz I. Tagging and morphological disambiguation of Turkish Text. *Proceedings of ANLP'94*. 144-149. 1994.

(Oflazer eta Tür 1996) Oflazer K., Tür G. Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation.. *Proceedings of EMNLP'96*. 69-81. 1996.

(Oflazer eta Tür 1997) Oflazer K., Tür G. Morphological Disambiguation by Voting Constraints. *Proceedings of ACL-EACL'97*. 222-229. 1997.

(Orphanos eta Christodoulakis 1999) Orphanos G., Christodoulakis D. POS Disambiguation and Unknown Word Guessing with Decision Trees. *Proceedings of EACL'99*. 134-141. 1999.

(Padró 1996) Padró L. A constraint satisfaction alternative for POS tagging. *Proceeding of NLP+IA/ TAL+AI'96*. 197-203. 1996.

(Padró 1997) Padró L. *A Hybrid Environment for Syntax-Semantic Tagging.* Tesia. Dep. Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. 1997.

(Padró 2001) Padró L. Tendencias en el reconocimiento de entidades con nombre propio. *Confluencias entre procesamiento del lenguaje natural y las tecnologías del habla.* Fundación Duques de Soria. Soria. 2001.

(Padró eta Màrquez 1999) Padró L., Màrquez L. On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpora. *Proceedings of COLING-ACL'98*. 997-1002. 1999.

(Palmer eta Hearst 1997) Palmer D.D., Hearst M.A. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics 23(2).* 241-267. 1997.

(Pereira eta Singer 1999) Pereira F., Singer Y. An Efficient Extension to Mixture Techniques for Prediction and Decision Trees. *Machine Learning*, 36(3).183-199. 1999.

(Pérez-Ortiz eta Forcada 2001) Pérez-Ortiz J.A., Forcada M.L. Part-of-speech tagging with recurrent neural networks. *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2001*.1588-1592. 2001.

(Perurena 2000) Perurena S. *EusLem: desanbiguazio-lanetarako interfazea.* Karrera bukaerako proiektua. Donostiako Informatika Fakultatea. 2000.

(Pla *et al*. 2000) Pla F., Molina A., Prieto N. Tagging and Chunking with Bigrams. *Proceedings of COLING-2000*. 2000.

(Pla *et al*. 2001) Pla F., Molina A., Prieto N. Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para castellano. 215-221. *Actas de SEPLN-2001*. 2001.

(Pla eta Prieta 1998) Pla F., Prieto N. Using Grammatical Inference Methods for Automatic Part-of-Speech Tagging. *Proceedings of LREC'98*. 597-601. 1998.

(Rabiner eta Juang 1986) Rabiner L.R., Juang B.H. An introduction to hidden markov models. *IEEE ASSP Magazine*. 4-16. Urtarrila 1986.

(Ratnaparkhi 1996) Ratnaparkhi A. A Maximum Entropy Part-Of-Speech Tagger. *Proceeding of EMNLP'96*. 1996.

(Reynar eta Ratnaparkhi 1997) Reynar, J.C., Ratnaparkhi, A. A Maximum Entropy Approach to Identifying Sentence Boundaries. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97).* 16-19. 1997.

(Riley 1989) Riley M.D. Some applications of tree-based modelling to speech and language indexing. *Proceedings of the DARPA Speech and Natural Language Workshop*. 339-352. 1989.

(Roche eta Schabes 1995) Roche E. and Schabes Y. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics,* 21(2). 1995.

(Roth eta Zelenko 1998) Roth D., Zelenko D. Part of Speech Tagging Using a Netqork of Linear Separators. *Proceedings of COLING-ACL'98*. 1136-1142. 1998.

(Sag *et al*. 2002) Sag I, Baldwin T., Bond F., Copestake A., Flickinger D. Multiword Expressions: a pain in the neck for NLP. *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2002*. 2002.

(Samuelsson eta Voutilainen 1997) Comparing a Linguistic and a Stochastic Tagger. *Proceedings of ACL-EACL'97*. 246-253.

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

(Sánchez eta Nieto 1995) Sánchez F., Nieto A. Desarrollo de un etiquetador morfosintáctico para el español. 1995.

(Sarasola 1982) Sarasola I. *Gaurko euskara idatziaren maiztasun-hiztegia. (3gn. liburukia)*. GAK, Donostia. 1982.

(Sarasola 1984) Sarasola, I. *Hauta-Lanerako Euskal Hiztegia*, GK. 1984.

(Sarasola 1996) Sarasola, I. *Euskal Hiztegia*. 1984.

(Schabes 1994) Schabes Y. Statistical versus Rule-Based Methods for Text Analysis. Tutorial. *European Summer School on Language and Speech Communication.* Utrecht 1994.

(Schiller 1996) Schiller A. Multilingual Finite-State Noun Phrase Extraction. *Proceedings of the ECAI'96 Workshop Extended Finite State Models of Language.* 1996.

(Schmid 1994-a) Schmid H. Part-of-speech tagging with neural networks, *Proceedings of COLING'94*, 173-176. 1994.

(Schmid 1994-b) Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceeding of the International Conference on New Methods in Language Processing, NeMLaP*. 44-49. 1994.

(Schmid 2000) Schmid H. Unsupervised Learning of Period Disambiguation for Tokenisation. Internal Report, IMS, University of Stuttgart, May. 2000.

(Schone eta Jurafsky 2001-a) Schone P., Jurafsky D. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords A Solved Problem?. *Proceedings of the EMNLP-2001*. 2001.

(Schone eta Jurafsky 2001-b) Schone P., Jurafsky D. Language-Independet Induction of Part of Speech Class Labels Using Only Language Universals. *Proceedings of the IJCAI-2001 Workshop "Machine Learning: Beyond Supervision"*. 2001.

(Schütze 1995) Schütze H. Distributional Part-of-Speech Tagging. *Proceedings of EACL´95,* 141-148. 1995.

(Segond eta Breidt 1995) Segond F., Breidt E. IDAREX: Formal description of German and French Multi-Word Expressions with Finite State Technology. Barne-txostena *MLTT-22 Grenoble.* 1995.

(Segond eta Tapanainen 1995) Segond F., Tapanainen P. Using a finite-state formalism to identify and generate multiword expressions. Barne-txostena *MLTT-19 Grenoble.* 1995.

(Sekine *et al*. 1998) Sekine S., Grishman R., Shinnou H. A Decision Tree Method for Finding and Classifying Names in Japanese Texts. *Proceeding of Workshop on Very Large Corpora (WVLC'98).* 171-177. 1998.

(Silva *et al*. 1999) Silva J., Dias G., Guilloré S, Lopes G. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Proceedings of 9th Portuguese Conference in Artificial Inteligence*. 21-24. 1999.

(Smadja 1993) Smadja F. Retrieving Collocations from Text: Xtract. *Computational Linguistics,* 19(1). 143-177. 1993.

(Sperberg-McQueen eta Burnand 1994) Sperberg-McQueen C.M., Burnand L. *Guidelines for Electronic Text Encoding and Interchange.* Chicago & Oxford, 1994.

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

(Spyrs 1994) Spyrs P. A robust category guesser for Dutch medical language. *Proceedings of ANLP'94*, 150-155 1994.

(Startvik eta Eeg-Olofsson 1982) Svartvik J., Eeg-Olofsson M. Tagging the London-Lund Corpus of Spoken English. In Johanson (1982), 85-109. 1982.

(Stede 1992) Stede M. 1992. The Search of Robustness in Natural Language Understanding. *Artificial Intelligence Review 6*, 383-414.

(Swish-E 2000) http://sunsite.berkeley.edu/SWISH-E. 2000.

(Tapanainen 1996) Tapanainen P. *The Constraint Grammar Parser CG-2*. Department of General Linguistics. University of Helsinki. 1996.

(Tapanainen eta Jarvinen 1994) Tapanainen P., Jarvinen T. Syntactic analysis of natural language using linguistic rules and corpus-based patterns. *Proceedings of COLING'94*. 1994.

(Tapanainen eta Voutilainen 1994) Tapanainen P., Voutilainen A. Tagging Accurately - Don´t guess if you know. *Proceedings of ANLP´94*, 47-52. 1994.

(Thede 1998) Thede S. Predicting Part-of-Speech Information about Unknown Words using Statistical Methods. *Proceedings of COLING-ACL'98*. 1505-1507. 1998.

(Tutin 1996) Tutin A. The formalization of collocations for Natural Language Processing: the Syntagmatic Lexical Functions model. *Proceedings of COMPLEX'96*. 243-256. Budapest 1996.

(Tufis 1999) Tufis D. Tiered Tagging and Combined Language Models Classifiers. *Proceedings of the Workshop on Text, Speech and Dialog (TSD99)*. 1999.

(Tufis eta Mason 1998) Tufis D., Mason O. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. *Proceedings of LREC'98*. 589-596. 1998.

(Tür eta Oflazer 1997) Tür G., Oflazer K. Tagging English by Voting Constraints. *Proceedings of COLING-ACL'98*. 1277-1281. 1998.

(Uchimoto *et al*. 2000) Uchimoto K., Ma Q., Murata M., Ozaku H., Isahara H. Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules. *Proceedings of ACL-2000*. 2000.

(Urizar *et al*. 2000) Urizar R., Ezeiza N., Alegria I. Morphosyntactic structure of terms in Basque for automatic terminology extraction. *Proceedings of the ninth EURALEX International congress*. 8-12 August 2000. Stuttgart.

(Urkia 1997) Urkia M. *Euskal morfologiaren analisi automatikorantz*. Doktoretza-Tesia. 1997.

(Urkia eta Sagarna 1991) Urkia M, Sagarna A. Terminología y Lexicografía asistida por ordenador. La experiencia de UZEI, *SEPLN*, vol 8. 1991.

(UZEI 1988) UZEI. *Laburtzapenen Gidaliburua. Siglak, ikurrak, laburdurak*, Elkar. 1988.

(van der Linden eta Kraaij 1990) van der Linden E., Kraaij W. Ambiguity resolution and the retrieval of idioms: two approaches, *COLING-90*, vol 2, 245-249. 1990.

(van Halteren *(ed.)* 1999) van Halteren H. (*ed*.) *Syntactic Wordclass Tagging*. University of Nijmegen, Kluwer Academic Publishers. Dordrech/Boston/London. 1999.

(van Halteren *et al*. 1998) van Halteren H., Zavrel J., Daelemans W. Improving Data Driven Wordclass Tagging by System Combination. *Proceedings of COLING-ACL'98*. 491-497. 1998.

(van Halteren *et al.* 2001) van Halteren H., Zavrel J., Daelemans W. Improving accuracy in word class tagging through combination of machine learning systems. *Computational Linguistics,* 27(2). 199-230. 2001.

(Voutilainen 1994) Voutilainen A. *Three studies of grammar-based surface parsing of unrestricted English text*. Publications n. 24. Dept. of General Linguistics. University of Helsinki. 1994.

(Voutilainen 1995) Voutilainen A. A syntax-based part-of-speech analyser. *Proceedings of EACL´95,* 157-164 1995.

(Voutilainen *et al.* 1992) Voutilainen A., Heikkilä J., Anttila A. *Constraint Grammar of English. A Performance-Oriented Introduction*. Publication No. 21. Department of General Linguistics. Helsinki: University of Helsinki.

(Voutilainen eta Järvinen 1995) Voutilainen A. and Järvinen T. Specifying a shallow grammatical representation for parsing purposes. *Proceedings of EACL'95*. 1995.

(Voutilainen eta Tapanainen 1993) Voutilainen A. and Tapanainen P. Ambiguity resolution in a reductionistic parser. *Proceedings of EACL'93*. 1993.

(Weischedel *et al.* 1993) Weischedel R., Meteer M., Schwartz R., Ramshaw L., Palmuzzi J. Coping with Ambiguity and unknown words through probabilistic models. *Computational Linguistics,* 19(2). 1993.

(Yamaguchi *et al*. 1998) Yamaguchi M., Kojima T., Inui N., Kotani Y., Nisimura H. Combination of an Automatic and an Interactive Disambiguation Method. *Proceedings of COLING-ACL'98*. 1423-1427. 1998.

(Zajac 1998) Zajac R.. Reuse and Integration of NLP Components in the Calypso Architecture. Workshop on Distributing and Accessing Language Resources. *Proceedings of LREC'98*. 1998.

(Zajac *et al*. 1997) Zajac R., Casper M., Sharples N.. An Open Distributed Architecture for Reuse and Integration of Heterogeneous NLP Components. Proceedings of the 5th Conference on Applied Natural Language Processing. Washington, DC, 1997.

Eranskinak

A Eranskina: kategoria sistema

A.1 Kategoria Lexikalak

A.1.1 Kategoria Nagusiak

- **IZE** **IZENAK**
  - **ARR** **ARRUNTAK** *(zuhaitz)*
  - **IZB** **PERTSONA-IZEN BEREZIAK** *(Mikel)*
  - **LIB** **LEKU-IZEN BEREZIAK** *(Donostia)*
- **ADJ** **ADJEKTIBOAK**
  - **IZO** **IZENONDOAK** *(handi)*
  - **IZL** **IZENLAGUNAK** *(benetako)*
- **ADI** **ADITZAK**
  - **SIN** **SINPLEAK** *(ekarri)*
  - **ADK** **KONPOSATUAK** *(lo egin)*

ADP **PERIFRASTIKOAK** *(ahal izan)*

FAK **FAKTITIBOAK** (*etorrarazi*)

- **ADB** **ADBERBIOAK**

ADO *ADITZONDOAK*

ADOARR **ARRUNTAK** *(gaur)*

ADOGAL **GALDETZAILEAK** *(noiz)*

ALG *ADITZLAGUNAK*

ALGARR **ARRUNTAK** *(abian)*

ALGGAL **GALDETZAILEAK** *(nondik)*

- **DET** **DETERMINATZAILEAK**

ERK *ERAKUSLEAK*

ERKARR **ARRUNTAK** *(hau)*

ERKIND **INDARTUAK** *(berori)*

NOL *NOLAKOTZAILEAK*

NOLARR **ARRUNTAK** *(edozein)*

NOLGAL **GALDETZAILEAK** *(zein)*

ZNB *ZENBATZAILEAK*

DZH **ZEHAZTUAK** (*bi*)

BAN **BANATZAILEAK** (*bina*)

ORD **ORDINALAK** (*bigarren*)

DZG **ZEHAZTUGABEAK** *(zenbait)*

ORO **OROKORRAK** *(guzti)*

- **IOR** **IZENORDAINAK**

PER *PERTSONALAK*

PERARR **ARRUNTAK** *(ni)*

PERIND **INDARTUAK** *(neu)*

IZG *ZEHAZTUGABEAK*

IZGMGB **MUGAGABEAK** *(norbait)*

IZGGAL **GALDETZAILEAK** (*nor*)

ELK **ELKARKARIAK** *(elkar)*

- **LOT** **LOTURAZKOAK**

  LOK **LOKAILUAK** (*hala ere*)

  JNT **JUNTAGAILUAK** (*edo*)

- **PRT** **PARTIKULAK** (*omen, ote, ...*)
- **ITJ** **INTERJEKZIOAK** (*alajaina!*)
- **BST** **BESTELAKOAK** (*baldin*)

A.1.2 Kategoria lagungarriak

- **ADL** **ADITZ LAGUNTZAILEAK** (*du*)
- **ADT** **ADITZ SINTETIKOAK** (*dator*)
- **HAOS** **HAUL OSAGAIA** (*hogeita*)
- **LAB** **LABURDURAK** (*etab.*)
- **SIG** **SIGLAK** (*EHU*)
- **SNB** **SINBOLOAK** (*a, b, c,...*)

A.2 Kategoria Morfologikoak

- **AMM** **ADITZ-MOTA MORFEMAK** *(-tu, -t(z)e,...)*

- **ASP** **ASPEKTU-MORFEMAK (Ø, -ko,...)**
- **ATZ** **ATZIZKIAK** (*-pe*)
- **AUR** **AURRIZKIAK** (*ber-*)
- **DEK** **DEKLINABIDE MORFEMAK** (*-aren*)
- **ELI** **ELIPSIA** (*Ø*)
- **ERL** **ERLAZIO ATZIZKIAK** (*-(e)la*)
- **GRA** **GRADUATZAILEAK** (*-ago*)
- **MAR** **MARRA** (-)

A.3 Kategoria Lagungarriak

- **PUNT** **PUNTUAZIO IKURRAK** (*puntu, koma, galde-ikurra...*)
- **BEREIZ** **BESTELAKO BEREIZGARRIAK** (*parentesi, komatxo*)
- **BESTE** **BESTELAKOAK** (*karaktere arrotzak dituzten hitzak)*
- **ID** **IDENTIFIKATZAILE-KODEKETAK**

B Eranskina: adibideak

B.1 "*CARTIER-BRESSONen*" hitzaren interpretazioak analisi morfologikoaren irteeran

```
/<CARTIER-BRESSONen>/<DEN_MAI_(DEK)>/
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + DEK GEN MG)
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + DEK GEN NUMP + DEK ABS MG)
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK ABS MG)
   ("CARTIER-BRESS"  SIG + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
   ("CARTIER-BRESSA"  SIG + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
   ("CARTIER-BRESSA"  SIG + DEK GEN PH MUGM + DEK GEN MG)
   ("CARTIER-BRESSA"  SIG + DEK GEN PH MUGM + DEK GEN NUMP + DEK ABS MG)
   ("CARTIER-BRESSA"  SIG + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
   ("CARTIER-BRESSA"   SIG + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK ABS
                       MG)
   ("CARTIER-BRESSA"  SIG + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
   ("CARTIER-BRESSON"  SIG + DEK GEN MG + DEK ABS MG)
   ("CARTIER-BRESSON"  SIG + DEK GEN MG LG)
   ("CARTIER-BRESSON"  SIG + DEK GEN NUMP MUGM + DEK ABS MG)
   ("CARTIER-BRESSON"  SIG + DEK GEN NUMP MUGM)
   ("CARTIER-BRESSON"  SIG + DEK INE)
   ("CARTIER-BRESSONA"  SIG + DEK GEN NUMP MUGM + DEK ABS MG)
   ("CARTIER-BRESSONA"  SIG + DEK GEN NUMP MUGM)
   ("CARTIER-BRESSONE"  SIG + DEK INE)
   ("CARTIER-BRESSONEN."  SIG)
   ("CARTIER-BRESSONEN"  SIG + DEK ABS MG)
   ("CARTIER-BRESSONEN"  SIG)
   ("Cartier-Bresson"  IZE IZB PLU- + DEK GEN MG + DEK ABS MG)
C  ("Cartier-Bresson"  IZE IZB PLU- + DEK GEN MG)
   ("Cartier-Bresson"  IZE LIB PLU- + DEK GEN MG + DEK ABS MG)
   ("Cartier-Bresson"  IZE LIB PLU- + DEK GEN MG)
   ("Cartier-Bresson"  IZE LIB PLU- + DEK NUMS MUGM + DEK INE)
   ("Cartier-bressone"  IZE LIB PLU- + DEK NUMS MUGM + DEK INE)
   ("Cartier-bressonen"  IZE IZB PLU- + DEK ABS MG)
   ("Cartier-bressonen"  IZE IZB PLU-)
   ("Cartier-bressonen"  IZE LIB PLU- + DEK ABS MG)
   ("Cartier-bressonen"  IZE LIB PLU-)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + DEK GEN MG)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + DEK GEN NUMP MUGM + DEK ABS MG)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK ABS
                     MG)
   ("cartier-bress"  ADJ IZO + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
   ("cartier-bress"  IZE ARR + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
   ("cartier-bress"  IZE ARR + DEK GEN PH MUGM + DEK GEN MG)
   ("cartier-bress"  IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP MUGM + DEK ABS MG)
   ("cartier-bress"  IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
```

```
("cartier-bress"  IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK ABS
                  MG)
("cartier-bress"  IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
("cartier-bress"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN MG + DEK
                  ABS MG)
("cartier-bress"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN MG)
("cartier-bress"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP MUGM
                  + DEK ABS MG)
("cartier-bress"   IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP
                  MUGM)
("cartier-bress"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN
                  NUMP MUGM + DEK ABS MG)
("cartier-bress"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN
                  NUMP MUGM)
("cartier-bressa"  ADJ IZO + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
("cartier-bressa"  ADJ IZO + DEK GEN PH MUGM + DEK GEN MG)
("cartier-bressa"  ADJ IZO + DEK GEN PH MUGM + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bressa"  ADJ IZO + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
("cartier-bressa"   ADJ IZO + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK
                  ABS MG)
("cartier-bressa"  ADJ IZO + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
("cartier-bressa"  IZE ARR + DEK GEN PH MUGM + DEK GEN MG + DEK ABS MG)
("cartier-bressa"  IZE ARR + DEK GEN PH MUGM + DEK GEN MG)
("cartier-bressa"  IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bressa"  IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP MUGM)
("cartier-bressa"   IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM + DEK
                  ABS MG)
("cartier-bressa"  IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN NUMP MUGM)
("cartier-bressa"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN MG + DEK
                  ABS MG)
("cartier-bressa"  IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN MG)
("cartier-bressa"  IZE ARR + MAR + IZE ARR + DEK GEN PH + DEK GEN NUMP MUGM +
                  DEK ABS MG)
("cartier-bressa"   IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + DEK GEN NUMP
                  MUGM)
("cartier-bressa"   IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN
                  NUMP MUGM + DEK ABS MG)
("cartier-bressa"   IZE ARR + MAR + IZE ARR + DEK GEN PH MUGM + ELI + DEK GEN
                  NUMP MUGM)
("cartier-bresson"  ADJ IZO + DEK GEN MG + DEK ABS MG)
("cartier-bresson"  ADJ IZO + DEK GEN MG)
("cartier-bresson"  ADJ IZO + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bresson"  ADJ IZO + DEK GEN NUMP MUGM)
("cartier-bresson"  ADJ IZO + GRA SUP + DEK ABS MG)
("cartier-bresson"  ADJ IZO + GRA SUP)
("cartier-bresson"  IZE ARR + DEK GEN MG + DEK ABS MG)
("cartier-bresson"  IZE ARR + DEK GEN MG)
("cartier-bresson"  IZE ARR + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bresson"  IZE ARR + DEK GEN NUMP MUGM)
("cartier-bresson"  IZE ARR + MAR + IZE ARR + DEK GEN MG + DEK ABS MG)
("cartier-bresson"  IZE ARR + MAR + IZE ARR + DEK GEN MG)
("cartier-bresson"  IZE ARR + MAR + IZE ARR + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bresson"  IZE ARR + MAR + IZE ARR + DEK GEN NUMP MUGM)
("cartier-bressona"  ADJ IZO + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bressona"  ADJ IZO + DEK GEN NUMP MUGM)
("cartier-bressona"  ADJ IZO + GRA SUP + DEK ABS MG)
("cartier-bressona"  ADJ IZO + GRA SUP)
("cartier-bressona"  IZE ARR + DEK GEN NUMP MUGM + DEK ABS)
("cartier-bressona"  IZE ARR + DEK GEN NUMP MUGM)
("cartier-bressona"  IZE ARR + MAR + IZE ARR + DEK GEN NUMP MUGM + DEK ABS MG)
("cartier-bressona"  IZE ARR + MAR + IZE ARR + DEK GEN NUMP MUGM)
```

```
  ("cartier-bressonen"  ADJ IZO + DEK ABS MG)
  ("cartier-bressonen"  ADJ IZO)
  ("cartier-bressonen"  IZE ARR + DEK ABS MG)
  ("cartier-bressonen"  IZE ARR + MAR + IZE ARR + DEK ABS MG)
  ("cartier-bressonen"  IZE ARR + MAR + IZE ARR)
  ("cartier-bressonen"  IZE ARR)
  ("cartier-bressonendu"  ADI SIN + AMM ADOIN)
  ("cartier-bressonendu"  IZE ARR + MAR + ADI SIN + AMM ADOIN)
```

B.2 "*CARTIER-BRESSONen*" hitzaren interpretazioak desanbiguazio tipografikoaren irteeran

```
/<CARTIER-BRESSONen>/<DEN_MAI_(DEK)>/
  ("CARTIER-BRESSON"  SIG + DEK GEN MG + DEK ABS MG)
  ("CARTIER-BRESSON"  SIG + DEK GEN MG LG)
  ("CARTIER-BRESSON"  SIG + DEK GEN NUMP MUGM + DEK ABS MG)
  ("CARTIER-BRESSON"  SIG + DEK GEN NUMP MUGM)
  ("CARTIER-BRESSON"  SIG + DEK INE)
  ("Cartier-Bresson"  IZE IZB PLU- + DEK GEN MG + DEK ABS MG)
C ("Cartier-Bresson"  IZE IZB PLU- + DEK GEN MG)
  ("Cartier-Bresson"  IZE LIB PLU- + DEK GEN MG + DEK ABS MG)
  ("Cartier-Bresson"  IZE LIB PLU- + DEK GEN MG)
  ("Cartier-Bresson"  IZE LIB PLU- + DEK NUMS MUGM + DEK INE)
```

B.3 "*Valentine*" hitzaren interpretazioak analisi morfologikoaren irteeran

```
/<Valentine>/<HAS_MAI>/
   ("VALENTINE"  SIG + DEK ABS MG @OBJ @SUBJ)
   ("VALENTINE"  SIG)
C  ("Valentine"  IZE IZB PLU- + DEK ABS MG @OBJ @SUBJ)
   ("Valentine"  IZE IZB PLU-)
   ("Valentine"  IZE LIB PLU- + DEK ABS MG @OBJ @SUBJ)
   ("Valentine"  IZE LIB PLU-)
   ("valentine"  ADJ IZO + DEK ABS MG @OBJ @SUBJ)
   ("valentine"  ADJ IZO)
   ("valentine"  IZE ARR + DEK ABS MG @OBJ @SUBJ)
   ("valentine"  IZE ARR)
   ("valentinetu"  ADI SIN + AMM ADOIN)
```

B.4 "*Valentine*" hitzaren interpretazioak desanbiguazio tipografikoaren irteeran

```
/<Valentine>/<HAS_MAI>/
C  ("Valentine"  IZE IZB PLU- + DEK ABS MG @OBJ @SUBJ)
   ("Valentine"  IZE IZB PLU-)
   ("Valentine"  IZE LIB PLU- + DEK ABS MG @OBJ @SUBJ)
   ("Valentine"  IZE LIB PLU-)
```

B.5 "*hala eta guztiz ere*" hitzen interpretazioak hitz anitzeko unitateen tratamenduaren aurretik

```
/<hala>/
C ("hala"  ADB ADOARR)

/<eta>/
C ("eta"  LOT JNT EMEN)
  ("eta"  LOT MEN KAUS)

/<guztiz>/
C ("guztiz"  ADB ALGARR)

/<ere>/
C ("ere"  LOT LOK EMEN)
```

B.6 "*hala eta guztiz ere*" hitzen interpretazioak hitz anitzeko unitateen tratamenduaren ondoren

```
/<hala>/
C ("hala_eta_guztiz_ere"  LOT LOK AURK 1/4)

/<eta>/
C ("hala_eta_guztiz_ere"  LOT LOK AURK 2/4)

/<guztiz>/
C ("hala_eta_guztiz_ere"  LOT LOK AURK 3/4)

/<ere>/
C ("hala_eta_guztiz_ere"  LOT LOK AURK 4/4)
```

B.7 Hitz-elkarketen interpretazioak

B.7.1 Loturik idatzitako hitz elkartuen analisiak

```
/<plazagizon>/
  ("plazagizon"  IZE ARR + DEK ABS MG)
  ("plazagizon"  IZE ARR)

/<idazmakina>/
  ("idazmakina"  IZE ARR + DEK ABS MG)
  ("idazmakina"  IZE ARR + DEK ABS NUMS MUGM)
  ("idazmakina"  IZE ARR)
```

B.7.2 Marratxoz bereizirik idatzitako hitz elkartuen analisiak

```
/<begi-nini>/
  ("begi-nini"  IZE ARR + DEK ABS MG )
  ("begi-nini"  IZE ARR)
  ("begi-nini"  IZE ARR + MAR + IZE ARR + DEK ABS MG)
  ("begi-nini"  IZE ARR + MAR + IZE ARR)

/<botoi-zulo>/
  ("botoi-zulo"  IZE ARR + DEK ABS MG)
  ("botoi-zulo"  IZE ARR)
  ("botoi-zulo"  IZE ARR + MAR + IZE ARR + DEK ABS MG)
  ("botoi-zulo"  IZE ARR + MAR + IZE ARR)

/<mahai-hanka>/
  ("mahai-hanka"  IZE ARR + MAR + IZE ARR + DEK ABS MG)
  ("mahai-hanka"  IZE ARR + MAR + IZE ARR + DEK ABS NUMS MUGM)
  ("mahai-hanka"  IZE ARR + MAR + IZE ARR)

/<hauteskunde-emaitzak>/
  ("hauteskunde-emaitza"  IZE ARR + MAR + IZE ARR + DEK ABS NUMP MUGM)
  ("hauteskunde-emaitza"  IZE ARR + MAR + IZE ARR + DEK ERG MG)
  ("hauteskunde-emaitza"  IZE ARR + MAR + IZE ARR + DEK ERG NUMS MUGM)
```

B.7.3 Bereiz idatzitako hitz elkartuen analisiak, lehen osagaia aldatua

```
/<itsas>/
C ("itsas"  IZE ARR)

/<armada>/
  ("armada"  IZE ARR + DEK ABS MG)
C ("armada"  IZE ARR + DEK ABS NUMS MUGM)
  ("armada"  IZE ARR)
```

```
/<euskal>/
C  ("euskal"  IZE ARR)

/<etxea>/
C  ("etxe"  IZE ARR + DEK ABS NUMS MUGM @OBJ @SUBJ @PRED)
```

B.7.4 Bereiz idatzitako hitz elkartuen analisiak, osagaiak aldatu gabe

```
/<hauteskunde>/
   ("hauteskunde"  IZE ARR + DEK ABS MG)
C  ("hauteskunde"  IZE ARR)

/<emaitzak>/
C  ("emaitza"  IZE ARR + DEK ABS NUMP MUGM)
   ("emaitza"  IZE ARR + DEK ERG MG)
   ("emaitza"  IZE ARR + DEK ERG NUMS MUGM)
```

C Eranskina: Emaitzak

C.1 Hitz ez-estandarren inguruko hainbat saiakuntza

IV. kapituluan aipatu den bezala, hitz ez-estandarren tratamendurako diseinatutako prozeduren konbinaketa posible gehienak probatu dira. Saiakuntza horien ondorioz, prozeduren hasierako anbiguotasuna murrizten den heinean emaitzak hobetzen direla ikusi da. Ondoren, interesgarriak diren batzuk aurkezten dira.

Horren adibide, C.1 taulan ikus daitezke. Emaitzei erreparatuz gero, zenbat eta informazio gehiago erabili, emaitzak hainbat eta zuzenagoak direla ikus daiteke, IV.6 taulan agertzen zirenak baino zuzenagoak, alegia[1]. Gainera, anbiguotasuna gehiago murrizten da, 2. mailan batezbesteko interpretazio kopurua 0,1 eta 1. mailan 0,6 inguru txikiagoak izanik banan bana aplikatuta lortutakoarekin konparatuta. Horregatik, desanbiguatzeko parametrizaturik ematen den mailan, 3. mailatik hasita emaitzak birfinduz lortuko dira MORFEUSen irteeran mantendu beharreko interpretazioak. Parametroa zehaztu ezean, 2. maila erabiliko da prozedura honetan.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%100	18,09	18,09	**%98,33**
im+estatistika 3	%100	12,56	12,56	**%96,35**
im+estatistika 3+2	%100	8,78	8,78	**%94,06**
im+estatistika 3+2+1	%100	8,16	8,16	**%86,46**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
im+estatistika 3	%100	14,35	14,35	**%94,04**
im+estatistika 3+2	%100	10,92	10,92	**%87,16**
im+estatistika 3+2+1	%100	9,94	9,94	**%77,52**

C.1 taula.- Informazio morfologikoa eta estatistikak hiru mailetan erabilita.

Bigarren saiakuntza batean anbiguotasuna gehien jaisten duen desanbiguazio tipografikoa lehenengoz aplikatu da. C.2 taulan ikus daiteke informazio morfologikoa era estatistika

[1] IV.6 taulan maila bakoitzeko emaitzak independenteki kalkulatuta lortutako emaitzak aurkezten dira.

erabiltzen duenaren aurretik aplikatuta, lau interpretazio gehiago baztertzen direla errore txikiagoa eginez.

(a)	**AR**	**I/A**	**I/T**	**R**
aurretik	%100	18,09	18,09	**%98,33**
tipografikoa	%99,79	8,23	8,22	**%96,67**
im+estatistika 3	%99,79	5,83	5,82	**%94,79**
im+estatistika 3+2	%99,79	4,06	4,05	**%93,02**
im+estatistika 3+2+1	%94,90	3,99	3,84	**%90,83**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
tipografikoa	%100	7,17	7,17	**%99,08**
im+estatistika 3	%100	5,87	5,87	**%93,58**
im+estatistika 3+2	%100	4,46	4,46	**%87,16**
im+estatistika 3+2+1	%91,28	4,41	4,11	**%81,19**

C.2 taula.- Anbiguotasun neurriak prozedura aurretik eta ondoren.

Ondoren, anbiguotasuna gutxi jaitsi arren, doitasun handia duen eratorrien prozedura gehituko da, C.3 taulako emaitzak lortu direlarik. Ikus daitekeenez, eratorrien prozedurak oraingoan ez du errorerik egiten, aurretik aplikatutakoak siglen kasuetan eratorpen bidezko interpretazioa baztertu duelako. Erreferentzia-corpuseko hitzen %5,65 tratatzen ditu 11,44 interpretaziotik 4,22ra eta egiaztapen-corpusean 16tik 7ra pasatuz.

(a)	AR	I/A	I/T	R
aurretik	%100	18,09	18,09	**%98,33**
tipografikoa	%99,79	8,23	8,22	**%96,67**
eratorpena	%99,79	7,99	7,97	**%96,67**
im+estatistika 3	%99,69	5,71	5,70	**%94,79**
im+estatistika 3+2	%99,69	4,04	4,03	**%93,12**
im+estatistika 3+2+1	%94,79	3,96	3,81	**%90,94**
(b)				
aurretik	%100	19,42	19,42	**%99,54**
tipografikoa	%100	7,17	7,17	**%99,08**
eratorpena	%100	7,10	7,10	**%99,08**
im+estatistika 3	%100	5,82	5,82	**%93,58**
im+estatistika 3+2	%100	4,44	4,44	**%86,70**
im+estatistika 3+2+1	%91,28	4,38	4,09	**%81,19**

C.3 taula.- Anbiguotasun neurriak prozedura aurretik eta ondoren.

C.2 Hitz ez-estandarren tratamenduaren emaitzak

IV. kapituluan modu grafikoan azaltzen dira hitz ez-estandarren emaitzak analizatzaile inkrementala eta hedatua erabilita modu argiagoan konparatu ahal izateko. Hala ere, emaitza osoak zehatz-mehatz azaltzen dira jarraian, datu guztiak kontsultatu ahal izateko nahi izanez gero.

Lehenengo analizatzaile inkrementalari dagozkion emaitzak aurkezten dira, C.4 taulan 4. mailako emaitza xehetasun osoz eta C.5 taulan mailakako batezbesteko neurriak emanez.

(a)	AR	I/A	I/T	R	P	F
estandar	%80,72	3,81	3,27	**%99,73**	%30,53	46,75
aldaerak	%81,83	4,47	3,84	**%92,26**	%24,05	38,16
ezezagunak	%100	18,09	18,09	**%98,33**	%5,44	10,30
testu-hitzak	%81,37	4,39	3,75	**%99,52**	%26,50	41,86
batez beste	%66,96	4,39	3,27	**%99,61**	%30,49	46,68
(b)						
estandar	%81,13	3,82	3,29	**%99,77**	%30,32	46,50
aldaerak	%74,00	4,14	3,32	**%70,00**	%21,08	32,41
ezezagunak	%100	19,42	19,42	**%99,54**	%5,13	9,75
testu-hitzak	%81,76	4,53	3,89	**%99,51**	%25,61	40,73
batez beste	%67,45	4,53	3,38	**%99,60**	%29,46	45,47

C.4 taula.- Analizatzaile morfologiko inkrementalaren emaitzak.

(a)	AR	I/A	I/T	R	P	F
1. maila	%39,79	2,42	1,57	**%99,86**	%63,80	77,86
2. maila	%43,71	2,48	1,65	**%99,66**	%60,55	75,33
3. maila	%64,77	3,31	2,49	**%99,64**	%39,94	57,03
4. maila	%66,96	4,39	3,27	**%99,61**	%30,49	46,68
(b)						
1. maila	%39,50	2,45	1,57	**%99,90**	%63,49	77,64
2. maila	%43,00	2,56	1,67	**%99,62**	%59,58	74,56
3. maila	%65,65	3,48	2,63	**%99,62**	%37,95	54,96
4. maila	%67,45	4,53	3,38	**%99,60**	%29,46	45,47

C.5 taula.- Anbiguotasun neurriak prozesuaren aurretik lau etiketatze-mailetan.

Ondoren, C.6 eta C.7 tauletan hitz ez-estandarren tratamenduaren aplikazioari dagozkion emaitzak azaltzen dira, berriz ere hitz-motaren arabera zehaztuta lehenengoan eta maila bakoitzeko batezbesteko emaitzak bigarrenean.

(a)	AR	I/A	I/T	R	P	F
estandar	%80,71	3,81	3,27	**%99,73**	%30,53	46,75
aldaerak	%75,36	2,97	2,48	**%90,36**	%36,41	51,91
ezezagunak	%83,12	4,11	3,58	**%92,92**	%25,92	40,54
testu-hitzak	%80,68	3,80	3,26	**%99,31**	%30,47	46,63
batez beste	%66,40	3,80	2,86	**%99,43**	%34,77	51,52
(b)						
estandar	%81,13	3,82	3,29	**%99,77**	%30,32	46,50
aldaerak	%60,00	2,80	2,08	**%70,00**	%33,65	45,45
ezezagunak	%92,20	3,98	3,75	**%88,07**	%23,50	37,10
testu-hitzak	%81,36	3,82	3,30	**%99,09**	%30,05	46,11
batez beste	%67,12	3,82	2,90	**%99,25**	%34,28	50,96

C.6 taula.- Analizatzaile inkrementalaren ondoren prozedurak aplikatzearen emaitzak.

(a)	AR	I/A	I/T	R	P	F
1. maila	%37,91	2,32	1,50	**%99,81**	%66,48	79,81
2. maila	%42,54	2,31	1,56	**%99,56**	%63,93	77,86
3. maila	%64,05	2,99	2,27	**%99,51**	%43,76	60,79
4. maila	%66,40	3,80	2,86	**%99,43**	%34,77	51,52
(b)						
1. maila	%37,18	2,33	1,49	**%99,89**	%66,88	80,11
2. maila	%42,14	2,33	1,56	**%99,54**	%63,87	77,81
3. maila	%65,25	3,08	2,35	**%99,47**	%42,25	59,31
4. maila	%67,12	3,82	2,90	**%99,25**	%34,28	50,96

C.7 taula.- Anbiguotasun neurriak prozesuaren ondoren lau etiketatze-mailetan.

Jarraian, analizatzaile hedatuari dagozkion taula parekoak aurkezten dira: C.8 eta C.9 tauletan tratamendua aplikatu aurretiko emaitzak azaltzen dira eta C.10 eta C.11 tauletan, berriz, tratamenduaren ondorengoak.

(a)	AR	I/A	I/T	R	P	F
estandar	%81,03	3,88	3,33	**%99,88**	%29,99	46,13
aldaerak	%83,39	4,65	4,04	**%96,84**	%23,94	38,39
ezezagunak	%99,80	18,25	18,21	**%98,41**	%5,40	10,24
testu-hitzak	%81,71	4,48	3,85	**%99,77**	%25,95	41,18
batez beste	%67,25	4,48	3,34	**%99,81**	%29,87	45,98
(b)						
estandar	%81,54	3,92	3,38	**%99,91**	%29,55	45,61
aldaerak	%72,92	4,83	3,79	**%91,67**	%24,18	38,26
ezezagunak	%100	19,46	19,46	**%99,56**	%5,12	9,73
testu-hitzak	%82,19	4,66	4,00	**%99,83**	%24,93	39,90
batez beste	%67,80	4,66	3,48	**%99,86**	%28,71	44,60

C.8 taula.- Analizatzaile morfologiko hedatuaren emaitzak.

(a)	AR	I/A	I/T	R	P	F
1. maila	%39,99	2,42	1,57	**%99,91**	%63,66	77,77
2. maila	%44,62	2,50	1,67	**%99,87**	%59,75	74,77
3. maila	%65,11	3,39	2,56	**%99,85**	%39,08	56,18
4. maila	%67,25	4,48	3,34	**%99,81**	%29,87	45,98
(b)						
1. maila	%39,79	2,46	1,58	**%99,96**	%63,29	77,51
2. maila	%44,35	2,60	1,71	**%99,90**	%58,47	73,77
3. maila	%66,03	3,59	2,71	**%99,90**	%36,87	53,87
4. maila	%67,80	4,66	3,48	**%99,86**	%28,71	44,60

C.9 taula.- Analizatzaile hedatuaren emaitzak lau etiketatze-mailetan.

(a)	AR	I/A	I/T	R	P	F
estandar	%81,03	3,86	3,31	**%99,88**	%30,14	46,31
aldaerak	%76,74	3,03	2,55	**%94,68**	%37,06	53,27
ezezagunak	%82,49	4,08	3,54	**%93,03**	%26,25	40,95
testu-hitzak	%80,99	3,85	3,31	**%99,54**	%30,11	46,23
batez beste	%66,66	3,85	2,90	**%99,63**	%34,38	51,11
(b)						
estandar	%81,54	3,90	3,36	**%99,82**	%29,69	45,76
aldaerak	%60,42	3,41	2,46	**%91,67**	%37,29	53,01
ezezagunak	%92,11	3,98	3,74	**%88,16**	%23,56	37,19
testu-hitzak	%81,78	3,90	3,37	**%99,31**	%29,47	45,45
batez beste	%67,47	3,90	2,96	**%99,43**	%33,65	50,28

C.10 taula.- Analizatzaile hedatuaren ondoren prozedurak aplikatzearen emaitzak[2].

(a)	AR	I/A	I/T	R	P	F
1. maila	%38,01	2,32	1,50	**%99,86**	%66,46	79,81
2. maila	%43,39	2,33	1,70	**%99,77**	%63,33	77,48
3. maila	%64,36	3,04	2,31	**%99,71**	%43,17	60,25
4. maila	%66,66	3,85	2,90	**%99,63**	%34,38	51,11
(b)						
1. maila	%37,36	2,33	1,50	**%99,94**	%66,83	80,10
2. maila	%43,48	2,35	1,59	**%99,75**	%62,88	73,56
3. maila	%65,61	3,15	2,41	**%99,68**	%41,34	58,44
4. maila	%67,47	3,90	2,96	**%99,43**	%33,65	50,28

C.11 taula.- Analizatzaile hedatuaren ondoren prozedurak aplikatzearen emaitzak lau etiketatze-mailetan.

C.3 HABILen emaitzak

V. kapituluan aurkeztutako emaitzak erreferentzia-corpusari dagozkio soilik. Jarraian, bi corpusetako emaitzak aurkezten dira, analizatzaile hedatua eta hitz ez-estandarren

[2] Hitz estandarren kasuan tratamendu berezirik diseinatu ez den arren, lema bat baino gehiago duten maiuskulaz idatzitako zenbait hitzen kasuan lema horietako batzuk baztertzen dira hitz ezezagunetarako diseinatutako prozedura berdina aplikatuz, baina izen berezi ez diren gainerako interpretazio guztiak mantenduz.

tratamendua aplikatuz lortutakoak. Gainera, hurrengo urratsari, hau da, desanbiguazioari begira, maila bakoitzean lortutako anbiguotasun- eta zuzentasun-neurriak ere ematen dira.

(a)	AR	I/A	I/T	R	P	F
estandar	%79,56	3,85	3,27	**%99,88**	%30,57	46,82
aldaerak	%76,69	2,99	2,53	**%94,93**	%37,57	53,83
ezezagunak	%82,49	4,08	3,54	**%93,03**	%26,25	40,95
testu-hitzak	%79,60	3,84	3,26	**%99,54**	%30,52	46,72
batez beste	%66,51	3,84	2,86	**%99,63**	%34,82	51,61
(b)						
estandar	%79,51	3,87	3,28	**%99,82**	%30,40	46,60
aldaerak	%59,57	3,46	2,47	**%91,49**	%37,07	52,76
ezezagunak	%92,11	3,98	3,74	**%88,16**	%23,56	37,19
testu-hitzak	%79,83	3,87	3,30	**%99,31**	%30,14	46,24
batez beste	%65,86	3,87	2,89	**%99,43**	%34,36	51,07

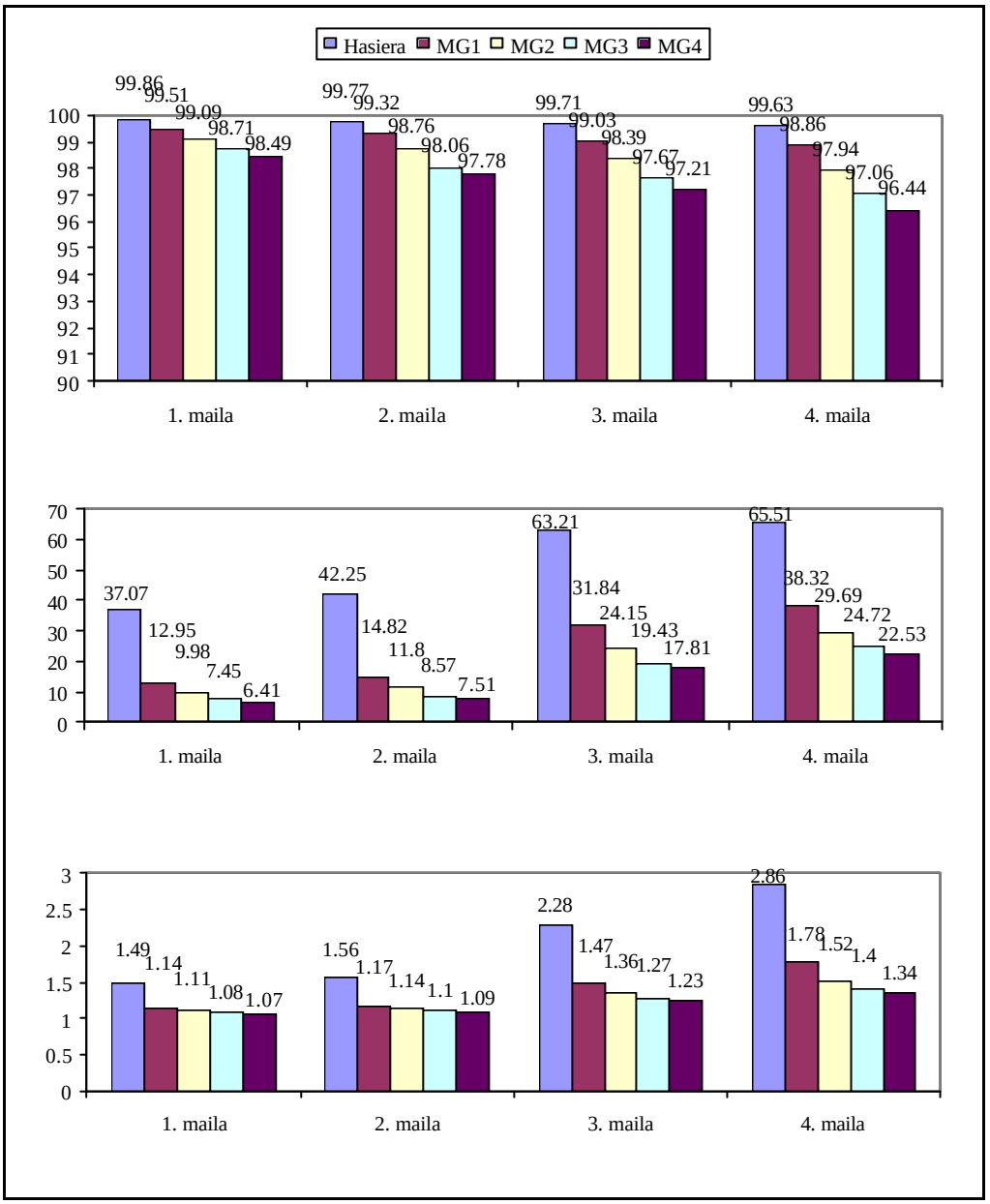
C.12 taula.- HABILen emaitzak analisi hedatuaren gainean.

(a)	AR	I/A	I/T	R	P	F
1. maila	%37,07	2,32	1,49	**%99,86**	%67,09	80,26
2. maila	%42,25	2,32	1,56	**%99,77**	%64,02	77,99
3. maila	%63,21	3,03	2,28	**%99,71**	%43,76	60,79
4. maila	%66,51	3,84	2,86	**%99,63**	%34,82	51,61
(b)						
1. maila	%36,05	2,32	1,48	**%99,94**	%67,67	80,70
2. maila	%42,02	2,34	1,56	**%99,75**	%63,79	77,82
3. maila	%64,01	3,13	2,36	**%99,68**	%42,22	59,32
4. maila	%65,86	3,87	2,89	**%99,43**	%34,36	51,07

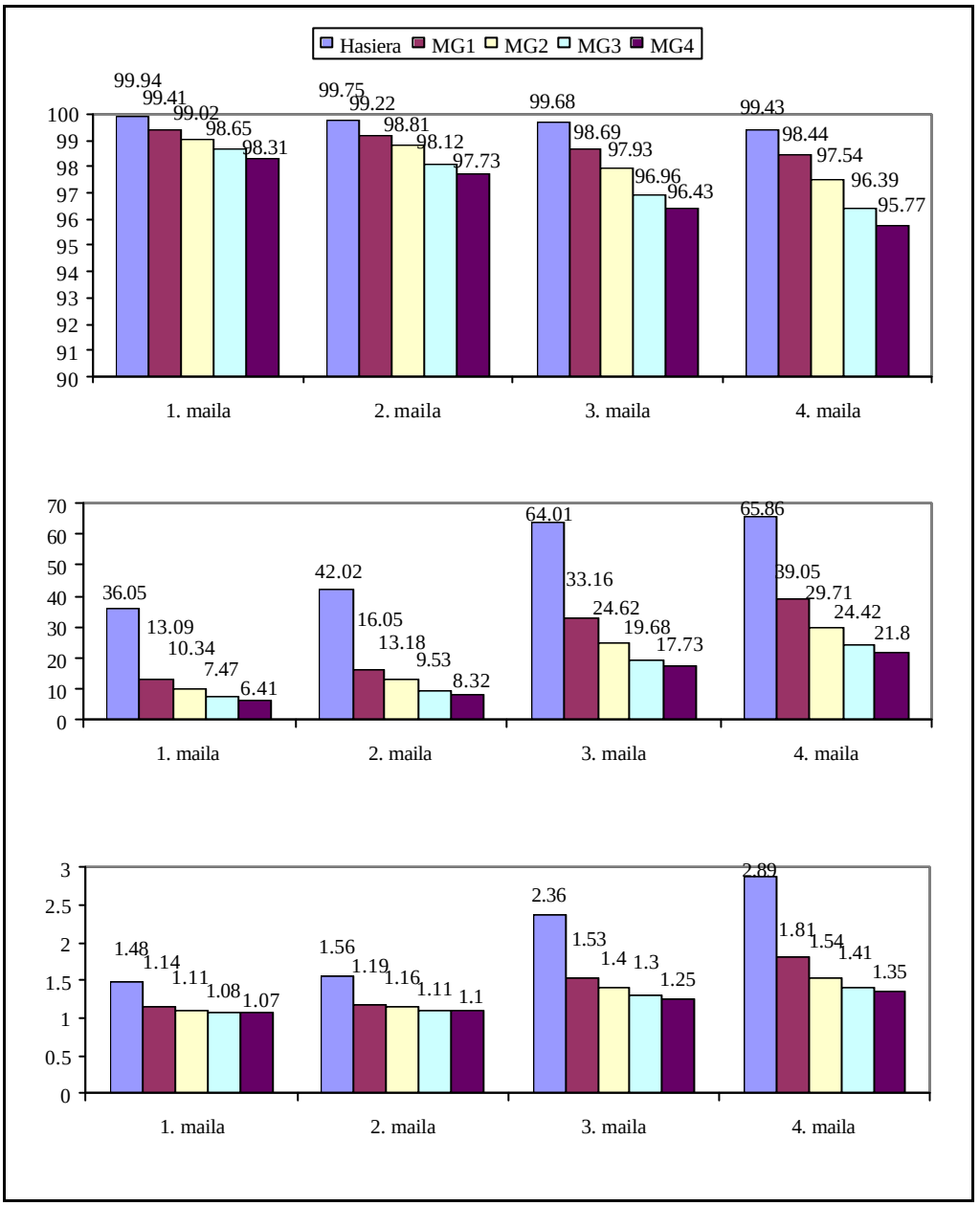
C.13 taula.- HABILen emaitzak analisi hedatuaren gainean lau etiketatze-maileta n.

C.4 Murriztapen gramatikaren bidezko desanbiguazioaren emaitzak

VI. kapituluan, murriztapen-gramatikari buruzko atalean, egiaztapen-corpusaren zuzentasuna eta anbiguotasun-tasaren emaitzak aurkezten dira soilik. Ondoren, erreferentzia- eta egiaztapen-corpusen emaitza osoak aurkezten dira (zuzentasuna, anbiguotasun-tasa eta batezbesteko analisi kopurua.



C.1 irudia.- MGren emaitzak etiketatze-mailaren arabera erreferentzia-corpusean.



C.2 irudia.- MGren emaitzak etiketatze-mailaren arabera egiaztapen-corpusean.