SARRERA ETA AURKEZPEN OROKORRA

I. Lanaren nondik norakoak eta aurkezpen orokorra.

I.1. Sarrera gisako aurkezpena.

Euskararen prozesaketa automatikoan lehen urrats bat izan nahi du aurkezten dugun *Euskal morfologiaren tratamendu automatikorako tresnak* izeneko lan honek.

Euskararen prozesaketa automatikoa bultzatu eta garatzeko epe luzerako egitasmo zabal batean kokatu behar da lan hau, horretarako hizkuntzalari eta informatikarien artean osaturiko talde bat elkarlanean aritzen garelarik.

Lengoaia Naturalaren Prozesaketak, bere gorabeherak eta guzti, garapen handia izan du azken hamarkadetan, baina garapen eta aplikazio gehienak ingeleserako egin dira. Bada hizkuntz sorta bat garapen honetaz baliatu dena hein txikiago batean izanda ere, batzuetan garapen berri hauetatik ideia eta eredu interesgarri orokorrak sortu direlarik. Azkenik, tratamendu informatikotik at gelditu diren hizkuntzak dauzkagu, normalean hiztunen kopuru txikiarengatik edota ezagutza ofizial ezarengatik merkatu-interesetatik kanpo daudelako.

Euskara azken multzo honetan kokaturik zegoela ikusirik —Abaituarena (1988) zen arlo honetan aipa daitekeen lan bakarra—, eta euskal gizarteak normalizazioaren bidean halako tresnak edukitzea ezinbesteko urratsa zelakoan, Donostiako Informatika Fakultateko Lengoaia eta Sistema Informatikoen Sailean Lengoaia Naturalaren

Prozesaketarako (LNP) talde bat sortzea erabaki genuen. Aipatutako arrazoietan oinarrituz egitasmo bat planteatu genuen ondoko helburu metodologiko hauekin:

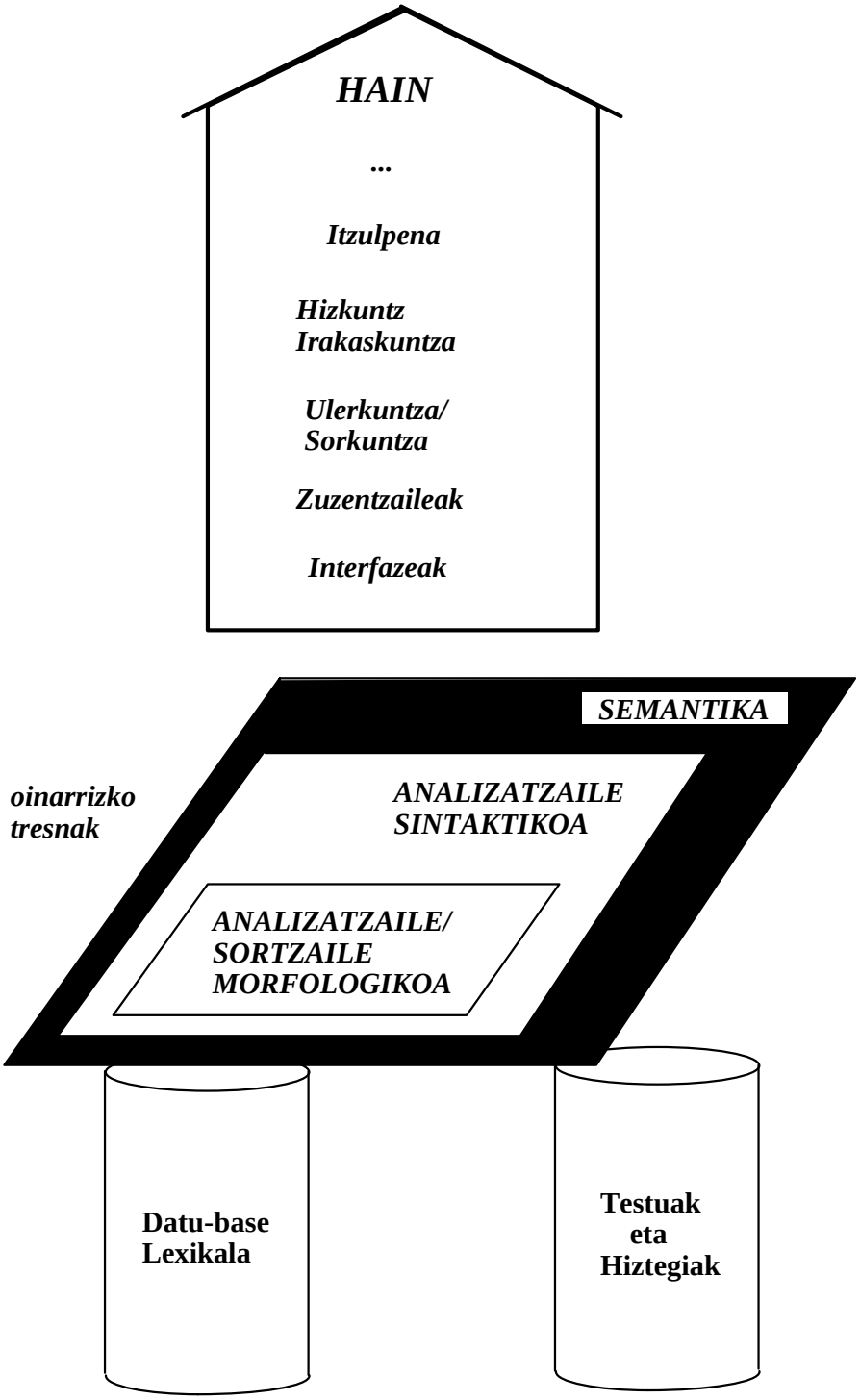
- **LNPren ikerkuntza-esparrua jorratzea**. Oinarrizko tresnetatik hasiz, oinarri sendoa osatzeko asmoz, etorkizunean helburu zabalagoetara heltzea da helburua.
- **Jakintza-arloen arteko elkarlana**. Hizkuntzaren alorra eta hizkuntza zehatz bat uztartzea teknologia informatikoaren eredu eta pentsamoldeekin, helburu bikoitza lortzeko asmoz: hizkuntzaren ezagumendua ustiatzea tresna automatikoak eraikitzeko batetik, eta teoria zein ekarpen linguistikoak egiaztatzea edota frogatzea informatikak eskaintzen dituen tresnak erabiliz. Uztartze honetan UZEI —Unibertsitate-Zerbitzuetarako Euskal Ikastetxea, terminologian eta lexikografian aritzen dena— izan da osagarri egokiena Informatika Fakultate batean sortutako talde honetarako.
- **Aplikazioa**. Garapen teorikoak baztertu gabe aplikazioa da gure lanaren zio 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 aplikazioei erantzuten dieten sistemen eraikuntza da gure helburu nagusia.
- **Berrerabilgarritasuna**. Burutzen diren aplikazioak berrerabilgarriak izan daitezela zentzu bikoitzean: batetik, aplikazio horien gainean eta ondoko urratsetan aplikazio konplexuago eta osotuagoak eraiki ahal izatea, eta bestetik, irekiak izatea aplikazio hauek beste erabiltzaileen esku jarriz.
- **Corpus idatzietan oinarrituta** baina **arauak errespetatuz**, eta corpusekin egiaztatuta eta neurtuta. Hau da, Euskaltzaindiak eta beste batzuek sortutako arauak eta teoriak kontuan hartzen dira, baina sistemen baliagarritasuna hizkuntzaren erabilera errealarekin alderatuz neurtu behar da.

Aipatutako ezaugarri horiek egitasmo osorako pentsaturik badira ere, hemen aurkezten den lanari aplika dakizkioke ere banan banan. Aurkezten dugun lanean bi tresna diseinatu dira prozesadore morfologikoa eta zuzentzaile ortografikoa, eskala errealekoak eta berrerabilgarriak biak, arau eta ezagutza morfologikoan oinarrituak, baina corpusekin egiaztatuak.

Tresnen nondik-norakoak azaldu baino lehen, egitasmo orokorraren barruan duen kokapena eta euskararen ezaugarri garrantzitsuenak azalduko ditugu.

I.2. Hizkuntzaren prozesaketa automatikoaren oinarria eta aplikazioak. Proiektuaren helburuak.

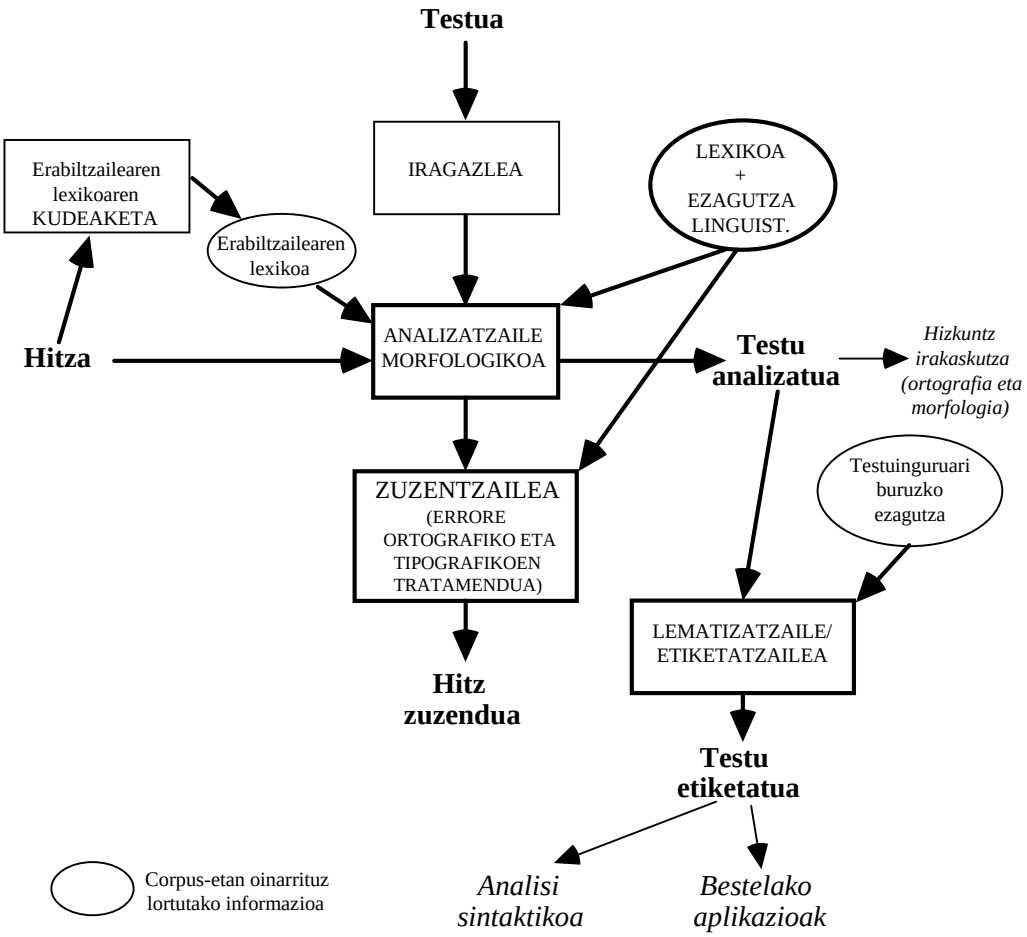
I.1 irudian gure ikertaldean euskara idatziaren prozesaketa automatikorako ezarri ditugun urratsak eta maila desberdinak irudikatzen dira, etxe baten eraikuntza simulatuz.



I.1 irudia.- LNPrako urratsak eta mailak etxe baten eraikuntza irudikatuz.

Prozesagarriak diren datu-base lexikala, testuak edo corpusak, eta hiztegiak dira etxeko zimenduak. Funtsezko informazio horiek ez baditugu, ezinezkoa izango da tresnak eraikitzen hastea, zeren tresna horiek errealitateari egokituak eta kalitatekoak izan daitezen, ezinbesteko oinarri eta erreferentzia baitira. Zimendu horiek bideratzen dute etxeko zorua eta egitura eraikitzea. Morfologia, sintaxia eta semantika dira garrantzi

handieneko lanbideak LNPrako sistemetan; zuzenean aplikazio komertzializagarriak ez izan arren, tresna komertzial gehienen funtsa baitira. Zimenduak, zorua eta egitura osoa buruturik dagoenean, etxeko paretak eta teilatua diren produktu erabilgarriak egitea lan gogorra da baina beti ziurra, oinarria ondo jarrita baitago. Produktuak aipatzean honako hauek azpimarra daitezke: zuzentzaileak, lematizatzaileak, lengoaia naturalaren bidezko interfazeak, testuen sorrera eta ulermena, ordenadorearen bidezko hizkuntz irakaskuntza, itzultzaile automatikoak edo semiautomatikoak, etab. Etxeko elementu guztiak kudeatzeko eta ustiatzeko ingurune bat ere aurrikusten da, HAIN —Hizkuntz Aplikazioetarako INgurunea— izenekoa.



I.2 irudia.- Aurkezten den lanaren eskema orokorra.

Dena den aurkeztu den egitasmo orokorrean esandakotik ez da ondorioztatu behar edozein produktu egiten hasi baino lehen zimenduek, zoruak eta egiturak erabat bukatuta egon behar dutenik, baina bai produktu bakoitzari dagokion oinarriari merezi duen garrantzia eman behar zaiola.

Lan honetan azaldutako egitasmoaren atal bat aurkezten da, morfologiaren inguruan dagoena hain zuzen ere. Morfologia lantzeko bidean EDBL izeneko datu-base lexikal bat

prestatu da, eta hizkuntzari buruzko ezagumendua lortzeko eta emaitzak ebaluatzeko testu-multzo bat lortu eta erabili ere. Eraikitako prozesadore morfologikoa erabiliz zuzentzaile ortografiko bat egin da, eta lematizatzaile/etiketatzaile bat proposatzen da. Prozesaketaren unitatea hitza denez, token-ezagutzailea edo iragazlea oso elementu garrantzitsua da.

Lanaren eskema orokorra I.2 irudian ikus daiteke.

I.3. Euskararen ezaugarriak modu laburrean.

Euskara da Europako hizkuntzen artean zaharrena, hizkuntza indoeuroparrak iritsi baino lehenago Europan zeudenen artean gelditu den bakarra baita. Teoria desberdinak badaude ere, bere jatorria zeharo egiaztatu gabe dago gaur egun.

Gaur egun hiztunak 650.000 inguru dira, Euskal Herriko populazioaren laurdena baino gutxixeago. Lurraldeei dagokienean, azken mendeetako murrizte-prozesua gelditzeko zorian dago Hego Euskal Herrian, baina ez Iparrean.

Ofizialtasuna du Araban, Bizkaian eta Gipuzkoan eta koofiziala da Nafarroan. Lapurdin Behe Nafarroan eta Zuberoan ez du ofizialtasun-aitorpenik.

Dialektoei dagokienean, oso hizkuntza aberatsa da zortzi euskalki bereizten dira eta. Aberastasun hori eta tradizio idatzi murritza dela eta, orain dela gutxi arte ez dira batasunerako urrats eraginkorrak eman. 1968an, zeuden kezkak eta saioak ikusita, Euskaltzaindiak bultzatu zuen euskara idatziaren batasuna, erabat arrakastatsua gertatu dena eta oraindik osatzen ari dena.

Morfologiaren aldetik honako ezaugarri hauek azpimarra daitezke:

- Oso flexio aberatsa, hamalau kasu desberdinekin, generorik gabe eta singularra eta pluralaz gain mugagabea ere bereiziz. Flexioa amankomuna da izen eta adjektiboetarako, oro har. Hizkuntza eranskaria da, askotan ezaugarri morfologiko bakoitzari morfema edo hizki bat dagokio eta. Horrela zenbait atzizkiren atzean atzizki gehiago metatu daitezke.
- Ergatiboa, kasu hau ez da hizkuntza indoeuroparretan agertzen.
- Aditza oso aberatsa da, aberastasun hau aditz laguntzailean eta trinkoan ere agertzen dela, eta forma bakar batean hiru pertsona-marketaraino irits daitekeela. Euskarak egiten duen genero-banaketa bakarra aditzean aurki daiteke, bigarren pertsona hurbilaren tratamenduan.

Egin dugun lanak ekarpen bat izan nahi du batasunaren bide horretan; helburu horrekin analizatzaile morfologiko estandarrak ez ditu ezagutuko forma estandartzat hartu ez diren hitz asko. Beraz, batasunaren aldeko apustu horrek oinarrizko analizatzailearen estalduran —ezagutzen eta analizatzen diren hitzen portzentaia— ondorio negatiboak ekarriko ditu. Hala ere analizatzaile estandarra baino harantzago doan analizatzaile hedatuaren bidez (ikus laugarren kapitulua), euskarazko hitz ez-estandar asko ezagutzeko aukera dago. Oinarri-lan honen fruitu gisa sortutako eta merkaturatutako ordenadore pertsonaletarako zuzentzaile ortografikoa batasunerako oso tresna baliagarria delakoan gaude.

Gure lanari ekiteko garaian izan ditugun oztopo nagusiak bi izan dira: batetik morfologiari buruzko lan sistematikoen falta, eta bestetik, batasunarekin loturiko gatazkak, irizpide finkoak ez zeudelako edo aldatuz joan direlako. Izan ere, azken urteetan euskararen inguruan egindako hainbat lan —gramatikak, hiztegiak, ikerketa-lanak, etab.— behar-beharrezko laguntza izan dira gure proiektuan.

I.4. Zimenduak: Euskararako Datu-Base Lexikala eta corpusak.

Proiektu hau bideratzeko, eta etorkizuneko aplikazioetarako datuak biltzeko funtsezko osagaiak dira bi hauek.

I.4.1. Euskararako Datu-Base Lexikala (EDBL).

Euskararako Datu-Base Lexikala (EDBL) funtsezko ezagumendu-oinarria da Lengoaia Naturaleko Prozesaketaren arlo askotan, eta bereziki morfologiaren alorrean. Eskala errealeko proiektu erreal bati ekitean pentsaezina da dimentsio errealeko informazioa testu arruntetan edo fitxategi konbentzionaletan biltegiratzea, eta datu-basea da dudarik gabe dagokion errepresentazio-sistema. EDBLk euskararen tratamendu automatikorako datu-base lexikal orokor bat izan nahi du, eta horrexegatik bertan mota guztietako informazioak biltzen dira, morfologikoak, sintaktikoak eta semantikoak.

Hasiera batean morfologiarekin lotutako proiekturako erabili denez, informazio morfosintaktikoari bultzada handia eman zaio, semantikari buruzko datuak gerorako utziz. Jakintza-arloen arteko talde-lana izatean, datu-basearen eguneratzea, zuzenketa eta mantenua linguisten zeregina izan den bitartean, informatikariena izan da datu-basearen eta intefazearen diseinua, esportaziorako prozeduren idazketa eta integritate- zein osotasun-egiaztapenerako murriztapenen definizioa.

Datu-basean informazio anitz metatzen da, baina inportanteena informazio lexikala dugu. Hirurogei mila sarreratik gertu daude erabat landuta, eta beste asko guztira osatu gabe. Sarrera bakoitzeari dagokion informazioa honako multzo hauetan bil daiteke:

- forma kanonikoa
- bi mailatako forma (morfologian erabiliko den ereduari egokitua)
- itsats dakizkiokeen morfemei buruzko informazioa
- erabilpenaren adibide bat
- kategoria, azpikategoria eta aditz-mota
- flexioari buruzko informazioa: kasua, zenbakia, mugatasuna, erlazioa, modu/denbora, pertsona
- kategoria erantsia
- iturburua, oharrak eta zalantzak
- maiztasuna (Sarasolaren maiztasun-hiztegiaren arabera, 1982)
- eguneratze-data eta berau egin duen hizkuntzalaria

Datu-basearen diseinuan eredu erlazionalari jarraitzen zaio, baina etorkizunerako, objektuei zuzendutako diseinu berri baten gainean ari gara lanean, gorde behar den informazioak duen konplexutasunari ondo erantzun ahal izateko, bertan lokuzioak, hitz anitzeko terminoak etab. biltegiratu nahi ditugu eta. Eredu berri horretan saihestuko da gaur egun datu-baseak bi mailatako morfologiarekin duen menpekotasuna, morfologia-formalismotik independentea bihurtuz. Datu-base honi buruz zehaztasun gehiago azaltzen dira hirugarren kapituluan.

I.4.2. Corpusak.

Testuek edo corpusek ematen dute benetan erabiltzen den hizkuntza idatziaren neurria. Gaur egun beren erabilpena areagotu egin da arrazoi horregatik, baita erregela-sistemen aurrean corpusetan oinarritutakoek duten eraginkortasuna eta sendotasunagatik ere. Leech-ek esaten duen bezala (Garside *et al.* 87: 3), corpusetan oinarritutako hurbilpenek eta erregeletan oinarrituek ezaugarri osagarriak dituzte, eta beraz osagarriak dira:

> ... The strength of the corpus-based approach is that, through probabilistic predictions, it is able to deal with any kind of English language text which is presented to it: it is eminently robust. Its weakness is that the very reliance on probability admits the possibility of error. The probabilistic system makes the best “guess” available to it, based on textual material that has been analysed in the past.
>
> This combination of strength and weakness is the exact opposite of the AI-based system which assumes (...) that 100% successful processing is possible, but which falls short of the ability to deal with uncensored, unrestricted text.
>
> We would argue that the two approaches are complementary: ...

Sistemen zehaztasuna neurtzeko erabilpenaz gain, beste zereginetan ere erabiltzen dira; sistemak azkartzea helburua duten maiztasun handieneko osagaien *buffer*-ak eta etiketatze-lanetan erabiltzen diren Bayes-en ereduak eta eredu markoviarrak dira corpusen erabilpen ezagunenak ezagumendu-iturri gisa. Izan ere, gaur egun corpus eleanitzak ere proposatzen dira haien hasierako eremutik kanpo ziruditen aplikazioetarako ere, haien artean itzulpenarena azpimarra daitekeela.

Corpusen artean ondoko sailkapen sinplea egin daiteke:

- **Orekatuak/ez-orekatuak**. Orekatuetan testu-moten artean halako oreka bat bilatzen da, testu-mota berezituei dagozkien ezaugarri partikularretatik aldenduz. Horretarako, iturburu desberdinetatik testu-zati txiki samar anitz, esanguratsuak eta aberasgarriak biltzen dira, teknika estatistikoak erabiliz. Corpus orekatuak ezinbestekoak dira ezagumendu-iturri gisa; besteek, ez-orekatuek hain zuzen ere, zehaztasuna neurtzeko bakarrik balio dute, helburu bereziturako sistemen eraikuntzan ez bada behinik behin.
- **Etiketatuak/etiketatu gabeak**. Gehienak etiketatu gabeak badira ere, ugaltzen ari da corpus etiketatuen eskaintza, ingeleserako behintzat. Etiketatuetan, aurreprozesaketa batez —eskuzkoa, semiautomatikoa edo automatikoa izan daitekeena— testuak zuen informazioaz gain beste datu batzuk gehitzen dira, zenbait erabilera erraztearren.

Testuak	**Ezaug.**	**hitzak**	**hitz kopurua**	**agerpen kopurua**[1]
1.- Argia aldizkaria (zatiak)	ez-orek.	4.864	2.607	1,86
2.- Filosofiari buruzko artikulua	ez-orek.	2.343	1.429	1,64
3.- EEBSko azken urteak	orekatua	23.364	9.313	2,51
4.- EEBS estandarra	orekatua	396.840	67.816	5,85

I.3 irudia.- Lanean zehar erabilitako zenbait testuren neurriak.

Gure kasuan, testu ez-orekatu batzuez gain, UZEIrekin izandako elkarlanari esker, EEBS proiektutik (Urkia & Sagarna, 91) banandutako corpus orekatu bat eskuratu dugu

[1] Forma bakoitzeko batez-beste agerpen kopurua.

euskarri prozesagarrian, honek lana izugarri erraztu digularik. I.3 irudian erabili ditugun corpus batzuen neurriak agertzen dira.

Corpus orekatu orokorrari “EEBS estandarra” deituko diogu lan honetan zehar, eta bere ustiapenerako zenbait arazo egon dira. Arazo garrantzitsuena hauxe izan da: euskara estandarrerako corpus orekatua zen helburua, baina EEBSn hogeigarren mendeko euskara idatziaren mota guztietako laginak daude; eta data, testu-mota eta euskalkiaren arabera sailkaturik egon arren, euskara batuaren garaiko testu neutroak —euskalkiaren aldetik— aukeratu arren, erabilpen ez-estandarrak agertzen dira maiz, euskara estandarraren arauak eta irizpideak aldatzen ari baitira. Honen ondorioz aukeratutako corpus orekatuan forma ez-estandar anitz agertzen dira, haien arteko batzuk maiztasun handiz. Euskara batua/estandarra bultzatzeko tresnak eraiki nahian, ezin izan diogu eman corpus orekatu honi beste hizkuntza normalizatuagoetan ematen zaion garrantzia; finkatzen ari diren irizpide batzuk corpusetan agertzen diren datuekin kontraesanean daudenean, irizpide horiei lehentasuna eman baitiegu.

Corpus orekatuan oinarriturik bi taula garrantzitsu lortu dira: maiztasun handieneko hitzena, eta maiztasun handieneko trigramena —hiru karaktereko multzo gainjarriak aurreko eta ondorengo zuriuneak kontuan harturik—.

I.5. Egiturazko tresna: prozesadore morfologiko automatikoa.

Prozesadore morfologiko baten eraikuntza eta beraren erabilpena beste tresnak diseinatzeko izan da lan honen muina. Konputagailuaren bidezko morfologiari ekin aurretik eredu desberdinak aztertu dira eta zenbait proba egin ere. Bide horretan, eta burutzapenaren lehen fase batean, bi “maketa” eraiki ziren bi formalismo desberdinen arabera: bi mailatako formalismoari (Koskenniemi, 83) jarraituz bat, eta ATEF sistema (GETA, 82) erabiliz bestea. Bibliografiatik ateratako ondorioak eta esperientzia praktikoetatik ateratakoak bat etorri ziren, eta bi mailatako morfologia izan zen aukeratu genuen eredu konputazionala.

Euskararako prozesadore morfologikoaren eraikuntza bi fasetan izan da burutua: euskara estandarrerako prozesadore morfologikoa batetik, eta aurreko prozesadore morfologikoak ezagutzen duen hitz-multzoa —*coverage* edo estaldura-tasa— handitzen duen “analizatzaile sendoa” bestetik.

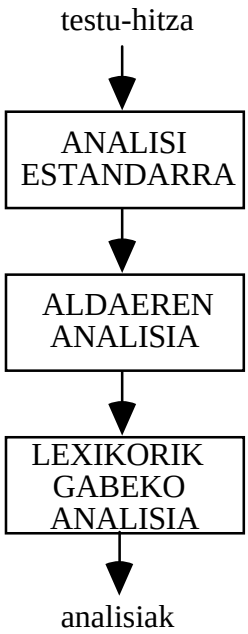
Bi faseetan erabili diren teknikak bi mailatako morfologian (Koskenniemi, 83) daude oinarrituta, eta horri esker sistema osoa homogenoa da irtenbide partikularretatik aldenduz. Hiru hobekuntza burutu dira bi mailatako formalismoaren inguruan: lehenengoz erabiltzaileen lexikoen erabilera bideratu da, bigarrenik bi mailatako paradigmaren

erabilpen “berri” bat egin da, aldaera deitu ditugun forma ez-estandarren tratamendurako; eta azkenik fonologiarako bakarrik erabilia zen “lexikorik gabeko analisia” testuen analisirako izan da hedatua.

Bi mailatako morfologiari jarraituz, euskara estandarraren morfologia deskribatzeko definitu dira oinarrizko bi osagaiak: morfemak eta haien arteko loturak zehazten dituen lexikoa batetik, eta morfemak biltzen direnean gertatzen diren aldaketa (morfo)fonologikoak deskribatzen dituzten erregelak.

Bi mailatako morfologiaren eredu klasikoa ez da nahikoa morfotaktikaren barruan kokatzen den gertakizun bat, urruneko menpekotasuna deitutakoa hain zuzen ere, modu egokian adierazteko. Euskararen morfologian urruneko menpekotasuneko kasu arrunt batzuk daude, eta horiek modu egokiagoan adierazteko proposamen bat egin dugu: jarraitze-klase hedatuak.

Dugun lexikoarekin, eta normalizatzen ari den hizkuntza batean hizkuntza estandarrera mugatzearen ondorioz, probatu diren testuetako %90 hitz inguru ezagutzen dira lehen hurbilpenaren bidez. Honen aurrean, eta emaitza hauek osatzearren, lehen aipatu diren hobekuntzak proposatzen dira prozesadore morfologikoa sendotzeko: erabiltzailearen lexikoen edo lexiko berezituen kudeaketa, forma ez-estandarrei dagozkien aldaeren tratamendua, eta analisia lema lexikoan egon gabe.



I.4 irudia.- Analisi morfologikoaren urrats desberdinak.

Lexiko orokorrean ez dauden lemak erabiltzailearen lexikoetan gorde daitezke; horretarako azpilexiko irekien eta itxien artean bereizketa egin delarik. Honen helburua

zera da: ezagututako hitzen kopurua handitzea, askotan termino teknikoak edo pertsona-zein leku-izenak ez baitaude jasoak lexiko orokorrean. Erabiltzaileak, horrela, aberats dezake lexikoa bere beharretara egokituz.

Aldaeren tratamendua funtsezkoa da hain batze-bide laburra duen hizkuntza baterako. Aldaerak, bi mailatako morfologiaz kudeatzen direnez, bi multzotan banatu ditugu: oso orokorrak direlako erregela morfofonologikoen bitartez adieraz daitezkeenak batetik, eta morfema zehatzei dagozkielako lexikoan adierazten direnak bestetik. Tratamendu honen bidez analizatzailearen estaldura-tasa hobetzeaz gain, forma ez-estandarrei dagozkien estandarrak lor daitezke, prozedura hau zuzenketan eta ordenadorez lagunduriko irakaskuntzan aplikazio zuzenekoa izanik.

Aurreko metodoez hitz bat analizatzerik ez dagoenean, analizatzaile morfologiko sendo bat lortzeko behinik behin, analisia lortzeko bideren bat bilatu behar da. Gure ebazpideak, bi mailatako formalismo barruan kokatzen denak, lemarik gabeko lexiko txiki bat erabiltzen du fonologiarako erabilitako metodo bati (Black et al., 91) jarraituz. Prozesu honi “lexikorik gabeko analisia” deitu diogu eta aipaturiko azpilexikoaz gain bi mailatako erregela berezi pare bat erabiltzen du.

Tratamendu-multzo horrekin aberasturiko analizatzaileak honako ezaugarriak ditu:

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

Deskribatutako prozesadore morfologiko hau oinarria da eraiki dugun Xuxen izeneko egiaztatzaile-zuzentzaile ortografikorako, garatzen ari garen EUSLEM izeneko euskararako lematizatzaile/etiketatzaile orokorrerako eta etorkizun hurbilerako helburu dugun analizatzaile sintaktikorako.

I.6. Prozesaketa morfologikoa hobetzen: Lexiko-itzultzaileak.

Bi mailatako morfologiaren arrakasta izugarria izan da, eta gure proiektua aurrera joan den bitartean beste talde batzuk haren inguruan hobekuntzak burutzen joan dira.

Hobekuntza horien artean azpimarratzekoa da *lexiko-itzultzaile* izenarekin ezagutarazi dena, Xerox-en garatua izan dena. Oinarri teorikoa (Karttunen *et al*., 92) eta aplikazio

praktikoa (Karttunen, 94) azken bi urteetan eman dira aditzera eta ekarri dituzten hobekuntzak bi arlotan bana daitezke:

- Eraginkortasunaren aldetik, lexikoa eta erregelak automata bakar batean biltzean, automata horren optimizazioari esker lortzen da abiadura handitzea oso modu garrantzitsuan.
- Deskribapen-ahalmenaren aldetik, bi mailatako morfologiaren erregela paraleloen abantailak mantendu arren, erregela paraleloen multzo desberdinen arteko konposaketa sekuentziala bideratzen dute, deskribapen ahalmena handitu eta deskribapena bera erraztuz.

Aldaketa garrantzitsu honen aurrean, eta diseinatutako tresnak erabiltzeko eman diguten aukeraz baliatuz, tresna berri hauen aplikazioa eta baliagarritasuna ebaluatu dugu, baina ez bakarrik analisi estandarrerako, baizik eta bi mailatako morfologiari buruz guk egindako aldaketen eta proposamenen gainean ere lexiko-itzultzaileek joka dezaketen papera ikertu dugu. Horretarako gure inplementaziorako geneuzkan datuak egokitu ditugu eta beraiekin euskara bezalako hizkuntza eranskari bati dagokion sistema erreal baterako aplikazioa aztertu dugu. Tratamendu gehienetarako hobekuntzak besterik ekartzen ez badituzte ere, zenbait muga igarri ditugu beraiengan, hirugarren eta laugarren kapituluetan ikus daitekeenez.

I.7. Produktu komertziala: Xuxen zuzentzaile ortografikoa.

Euskararako zuzentzaile ortografiko bat burutzeko ideia taldearen helburu nagusien artean zegoen hasiera hasieratik. Arrazoi nagusiak honako hauek ziren:

- Inguruko beste hizkuntzetarako eskuragarri zen aipaturiko produktu hori, baina ez euskararako.
- Helburu nagusien artean aipatu diren *aplikazioa* eta *eskala erreala* irizpideekin bat zetorren bete-betean.
- Euskarak bizi duen batasun-prozesurako are garrantzi handiago du halako tresna batek. Zentzu berean, gertatzen diren idazketa-erroreetan batasuna erabat finkatu gabe egoteak problematika berria eta aberatsa dakar, ikergai erakargarria bihurtuz.

Euskararen ezaugarriek, flexio aberatsa eta eranskaria edukitzeak, zuzentzailea morfologian oinarritzera eraman gintuen ezinbestean, hitzak onartuz banaketa morfologiko posiblea baldin badute. Bibliografian agertzen ziren erreferentzia gehienak ez

dira oso baliagarriak, euskara bezalako hizkuntza eranskarietan zuzenketa-prozesua korapilatsuagoa da eta.

Aipatutako testuinguruan, zuzenketari ekin aurretik ondoko erizpide hauek geneuzkan buruan:

1) Testuingurua kontuan hartzen duen zuzentzailearena —bigarren belaunaldiko zuzentzaileak edo estilo-zuzentzaileak ere deituak— proiektu erakargarria izan arren, lehen urrats batean zuzentzaile konbentzional batera murriztu ginen, hartarako behar ziren beste oinarriak —analisi sintaktikoa etab.— egin gabe daudelako.

2) Batasunarekiko zalantzak sortutako erroreak —orokorrean gaitasun-erroreak edo aldaerak deituko ditugunak— ziren tratatzeko lehentasuna zutenak, gainontzekoen tratamendua —errore tipografiko deitutakoena— bigarren maila batean utziz.

Ondorioz, zuzenketa bi modulu osagarriren bidez burutzen da, gaitasun-erroreena batetik eta errore tipografikoena bestetik; eta lehen motako erroreak tratatzeko bi mailatako morfologian oinarritutako metodo berritzaile batera iritsi garen bitartean, errore tipografikoak tratatzeko prozedura klasiko bat erabiltzen dugu, azkartzeko bideetan zenbait ekarpen egin badira ere.

Zuzenketa-aplikazioetan ohizko diren beste moduluez gain, iragazlea adibidez, erabiltzailearen hiztegiarekin ekimen berezi bat egin da, hizkuntzalari ez den erabiltzaile bati informazio morfosintaktikoa eskatzeko modua sakonean aztertu da, informazio horrekin sistemak hitz berri baten flexio guztiak ezagutuko baititu.

I.8. Hurrengo urratsa: EUSLEM.

Aurkezten den lanean hitza da tratamendu-unitatea. Tesi hau mugatzeko orduan, testuingurua kontuan hartzen duen oro kanpoan utzi dugu, arlo horretan lanean ari bagara ere.

EUSLEM diseinatu den eta gauzatze-bidean dagoen lematizatzaile/etiketatzailea da, euskararako eta bi mailatako analisi morfologikoan oinarrituta. Diseinu-filosofia orain arte azaldutako bera da: eskala erreala, aplikagarritasuna, garatutako beste tresnen berrerabilpena garrantzitsuenak izanik. Horrez gain lematizatzaile/etiketatzailea aplikazio konkretutik independente diseinatu da, helburu desberdinekin erabili ahal izateko. Tresna honen oinarrizko osagaiak hauek dira:

- *Token*-ezagutzailea deitzen den aurreprozesadorea, hitzak, puntuazio-karaktereak, zenbakiak etab. identifikatzeko. Analisi morfologikorako egindakoa erabiliko da aldaketa gutxi batzuekin.
- Analizatzaile morfologikoa, hitzei dagozkien lema eta etiketa posibleak zehazteko. Egindakoa berrerabiliko da.
- Hitz ezezagunen etiketatzailea edo *guesser*-a, analizatzaile morfologikoak ezagutzen ez dituen hitzen lema eta etiketa hipotetikoak lortzeko. Egindako aldaeren analisia eta lexikorik gabeko analisia erabil daitezke helburu horrekin.
- Etiketen definizioa eta analisi morfologikoarekiko egokitzapena.
- Hitz anitzeko terminoen identifikazioa, horien tartean lokuzioak, hitz-elkarketa eta bestelako kasu asko sartzen direlarik.
- Testuinguruan oinarritutako desanbiguazio, metodo estokastiko, linguistiko edo bion konbinaketaren bidez egina.

Esan bezala, hitza baino harantzago doazen tratamenduak lan honen esparrutik kanpo gelditzen dira; beraz, aplikazio hau irekitako ikerlerro gisa aurkezten da.

I.9. Egindakoaren aplikazio berri posibleak.

Lan honetan aurkezten diren tresnek morfologia dute oinarritzat; hala ere prozesadore morfologiko batean oinarriturik egin daitezkeen aplikazioak askoz gehiago dira. Honako hauek dira garrantzitsuenak:

- Esan den bezala lematizazioa analisi morfologikoan oinarritzen da, eta lematizazioa funtsezko tresna da lan lexikografikoetan nahiz informazioa biltegiratzen eta berreskuratzen laguntzen duten sistemetan, dokumentuen datu-baseak adibidez.
- Hizkuntz aplikazio sakonagoetarako —sintaxia, itzulpen automatikoa, etab.— lehen urrats gisa.
- Hizketaren sintesia edo testu-sorkuntza lortzeko sorkuntza morfologikoa funtsezko osagarria da. Hizketaren kasuan, ahoskatzeko testua eduki arren, ahoskatzeko orduan informazio morfologikoa garrantzizkoa izan daiteke.
- Itzulpenerako laguntza-tresnak. Hiztegi elebidun baten laguntzaz iturburu-testu batetik abiatzen bagara, hiztegian adierak bilatzeko jatorrizko testuaren lema posibleak lortu behar dira hiztegian bilatu ahal izateko. Gaur egun gorantz doan

ikerlerroa den testu-parekatzean analisi morfologikoak ere funtsezko funtzioa eduki dezake.

- Beste aplikazioak. Zaila izango litzateke aplikazio guztiak banan-banan zerrendatzea. Hona hemen gure sistemarekin buruturiko bat: koherentzi egiaztatzea. Entziklopedia-hiztegi batean zenbait sarrera kendu behar ziren baina hauek kentzean testuak kontsistentzia gal zezakeen, definizioetan sarrera horiek edo berauen flexioak ager baitzitezkeen. Honen aurrean, eta lexiko-sarrerak arruntak ez zirenez, sarrera horiek eta flexio-hizkiek osatutako lexiko bat osatu genuen, eta testua analizatzean analisirik lortzen zuten hitzak baztertzekoak ziren, kendutako sarreren forma flexionatuak baitziren.

Zuzentzaile ortografikoarenak, berriz, hauexek lirateke testu-edizioaz gain:

- Ordenadorez Lagunduriko Irakaskuntzako sistemetan (OLI), morfologia eta ortografia irakatsi eta zuzentzeko.
- OCR dispositiboen bidez jasotako euskarazko testuen zuzenketa.
- Hizketaren analisiaren emaitza zuzentzea. Zuzentzailea egokitu beharko litzateke, testu idatzian eta hizketan agertzen diren hizkuntzen ezaugarriak desberdinak dira eta. Gainera hizketaren kasuan, OCRan bezala, zuzenketa automatikoa behar da.
- Pertsona-makina elkarrekintza aplikazio orotarako, pertsonak barneratzen duen informazioan akatsak egon daitezkeela suposatzen bada behintzat. Datu-baseen zein entziklopedien kontsulta-sistema lokalak zein sarearen bidezkoak, eta elbarritu edo behar bereziak dituzten pertsonekiko komunikazio-sistemak sartzen dira multzo honetan.

I.10. Txostenaren eskema.

Ondoan azaltzen den txostena lau partetan dago banaturik.

Lehendabizikoan euskararako prozesadore morfologikoaren diseinua azaltzen da hiru kapitulutan banatuta. II. kapituluan morfologiaren oinarri minimoak azaldu eta gero, morfologiaren tratamendurako eredu konputazionalen azterketa egiten da, egoera finituko ereduetan eta, batez ere, bi mailatako morfologian sakonduz. Urruneko menpekotasunak ebazteko gure ekarpena den “jarraitze-klase hedatuak” izeneko mekanismoa ere azaltzen da bertan. III. kapituluan bi mailatako morfologiaren aplikazioa den euskara estandarrerako prozesadore morfologikoaren diseinu eta gauzatzea azaltzen da, lexiko-

itzultzaileen bidez ere egiten dena. IV. kapituluan azkenik, aurreko analizatzaile morfologikoa estaldura handiko eta sendo bihurtzeko urratsak azaltzen dira, bertan bi mailatako paradigmari jarraitzen dioten erabiltzailearen lexikoen kudeaketa, aldaeren ezagutza eta lexikorik gabeko analisia adieraziz.

Bigarren parteak zuzenketa du oinarritzat eta bi kapitulutan dago banaturik. V. kapituluan zuzenketaren nondik-norakoak azaltzen dira flexio handiko hizkuntzak eta hizkuntza eranskariak zuzentzeko dauden arazoetan sakonduz. Era berean ez-jakiteak bultzatutako erroreak, gaitasun-erroreak deitutakoak, tratatzeko garrantzia ere azpimarratzen da. VI. kapituluan morfologian oinarritutako euskararako zuzentzaile ortografiko baten diseinu eta gauzatzea deskribatzen da, morfologiarako garatutako hainbat tresna berrerabiltzen direlarik.

Hirugarren partean berriz, egindako lanaren ondorioak eta etorkizuna ditugu hizpide. Bertan azalduko dira egindako lanaren alde azpimarragarrienak, EUSLEM izeneko lematizatzaile/etiketatzailean egindako lanak duen tokia, eta lan honen ondorioz aurkitutako ikergai interesgarrienak.

Azkenik, erabilitako bibliografia gaika azalduta, eta testuan zehar proposatutako eranskinak gehitzen dira.