Morfologia konputazionala eta euskal morfologia

MIRIAM URKIA
(UZEI)

Abstract

This paper constitutes an approximation to the automatic treatment of the Basque morphology. After considering several alternatives, the Two Level Formalism is adopted and adapted to Basque. On the other hand, a morphosyntactic analysis is adopted which is superior to the above mentioned formalism. On the basis of the results of the analysis proposed here, some theoretical problems that the automatization of language, in this case of Basque, raises are also considered: morpheme partitioning vs. piling, ambiguity, lexical delimitation, ellipsis within words, nominalization and categorial changes.

Seguru asko bat baino gehiago harritu egingo zen txosten honen izenburua ikustean. Izan ere, morfologia konputazionala ez dago hedatuegia gure artean eta askorentzat informatikoen kontua ere izango da, morfologia aitzakiatzat hartuta sistema berriak garatzeko arrazoia. Baina ez da horrela.

Hemendik kanpo gehiago landu den zientzia hau duela gutxi gureganatu dugu eta lehen saioak besterik egin ez badira ere, teknologia mailan beste edozein hizkuntza adina bada euskara konputazionalki landua izateko. Hori azaltzen saiatuko gara hurrengo minutuotan.

Berez ez dugu aportazio teorikorik egingo morfologiari dagokionean; aitzitik, morfologia konputazionalaren nondik norakoak azaldu eta euskararako lehen aplikazioaren berri ematen saiatuko gara. Hizkuntzalariek esandakoa jaso eta sistematizatze-saio baten emaitza aurkeztu nahi dugu besterik gabe.

1. Morfologia konputazionalaren tratamendua

1.0. Azken urteotan linguistika konputazionalak indar handia hartu du ikerlarien artean, aurrerakuntza teknologikoak eta ordenadoreen gureganatzeak eraginda neurri handi batean. Garai batean, ez hain urrun askorentzat, zapata-kaxatan gordetzen ziren milaka fitxa haiek datu-base sinpletan gorde ohi ditugu gaur, eta horrek berak informazioaren egituratzera eraman gaitu ezinbestean. Egun informazio-masa handia bildu eta ustiatzen dugu, nahiz informatikarienganako dependentzia handia dugun

oraindik ere, askotan beraiek egindako aplikazioak papereratu eta gorriztatzera mugatuz.

Behar horiek erakutsi digute zeinen beharrezkoa den hizkuntzaren sistematizazioa, are gehiago itzulpen automatikoaz hain aisa hitz egiten den egunotan.

Disziplinarteko gai baten aurrean gaude, beraz. Informatika hizkuntzaren erabileraren eta hizkuntzalaritzaren zerbitzutan batetik, eta hizkuntzalaritza informatikaren zerbitzutan bestetik.

Linguistika konputazionala aipatzen dugunean helburu orokor bat dugu aurrean, hau da, gizakiak hizkuntza prozesatzeko darabilen mekanismoaren analogoa ordenadore bidez lantzea. Baina badira gertuagoko bi helburu nagusi: bat, linguistikaren ezagutza esplizituki formalizatu eta, bi, hipotesi eta teoria linguistikoetarako probabanku bat izan. Hau, betiere, hitzaren forma grafikoa (idatzia) hartuta, ez fonologikoa (ahozkoa).

Linguistika konputazionala ahaleginak egiten ari da hizkuntza naturala maila gehienetan prozesatzeko eta, ezinbestean, hainbat esparru ukitzen ditu: lexikologia, lexikografia, lematizazioa, terminologia, dokumentuen tratamendua, datu-baseen kontsulta, morfologia, sintaxia, semantika, testuen analisi estilistikoak eta estatistikoak, ahozko aplikazioak (bai irakurketarako bai ahotsaren ezagutzarako), egiaztatzaile eta zuzentzaile ortografikoak, tutorialak, itzulpen automatikoa, etab.[1]

Morfologiaren tratamendu automatikoa da gure gaia eta horretara mugatuko dugu aurkezpena. Zenbaitentzat ez da gairik erakargarriena edo aberasgarriena izango beharbada, baina ezin ahantz dezakegu gainerako aplikazioetarako oinarri dela morfologia. Linguistika konputazionalaren helburu orokorrak ikusi ondoren, morfologia konputazionalaren xedea bi ideia nagusitan bil daiteke: *deklaratibotasuna* eta *bidirekzionalitatea*, hots, analisia eta sorkuntza.

Beraz, prozesadore morfologikoak analizatu eta sortu egin behar badu, lematizazioa edo morfemen segmentazioa interesatuko zaigu batetik eta, bestetik, informazio lexikoaren berreskuratzea, segmentaziotik abiatuta.

Prozesadore bat osoa izateko beharrezkoa da osagai sintaktikoekin eta semantikoekin erlazionatuta egotea, hau askotan aplikazioaren araberakoa bada ere. "Morphology is truly a crossroads of linguistics" dio Bauer-ek [Bauer 1983]. Beraz, goian aipatutako beste aplikazioak ezingo ditugu erabat bazterrean utzi.

Hau guztia teorian horrela den arren, praktikan analisia izan da gehien aztertu dena eta sorkuntza, aldiz, apenas ukitu den. Bestalde, flexioari garrantzia eman zaion bezala, ia ez dira hitz-elkarketa eta eratorpena landu.

(1) Euskararako ere garatzen ari dira hainbat proiektu hizkuntza eta informatika uztartu nahian. Analizatzaile morfologikoa oinarri hartuta, Donostiako Informatika Fakultatea (EHU) eta UZEI elkarlanean garatzen ari diren proiektuak dira ondokoak: XUXEN (egiaztatzaile/zuzentzaile ortografikoa), EDBL (Euskararen Datu-Base Lexikala), Murriztapen-Gramatiken euskararako aplikazioa (lehen fasean oraindik), hizkuntzen irakaskuntzarako tutorialak, EUSLEM (Euskararako Lematizatzaile Automatikoa),... Badira Euskal Herrian beste proiektu batzuk ere, analizatzailea ez darabiltenak: LFG (Lexical Functional Grammar) euskarara aplikatua (Joseba Abaituaren tesia), hiztegi-sistema urgazle adimenduna (HIZTSUA), ezagutza semantikorako tresnak, ahotsaren tratamendua (AHOZKA), etab.

1.1. Morfologia konputazionalaren ikuspegi historikoa

1.1.0. Hasieran ikerketa programazioaren ildotik joan zen, irizpide linguistikoak gehiegi hartu gabe kontuan. Honen arrazoi nagusietako bat ingelesa izan daiteke, morfologikoki ez duelako konplexutasun handirik eta sarrera guztiak zerrendatzea ez delako nekosoegia. Horregatik bataiatu zuten "Full listing hypothesis" (FLH) izenarekin orduko prozedura. Baina lekuan lekuko hizkuntzak automatikoki lantzen saiatu orduko ohartu ziren ezin zela guztia zerrendatu eta bide berria urratu zutenak linguista eskandinaboak izan ziren.[2] Poliki azaltzen du Martin Kay-k [Kay 1973] aldaketa honen beharra. Izan ere, karaktere-segidak listatuta, ingelesez esaterako (sinpleenetakoa izan daitekeelako), sarrera-kopurua laukoiztu edo seikoiztu egingo litzateke eta, gainera, lexiko-sorkuntzari ateak itxiko lizkioke. Baina pentsaezina da hau hizkuntza eranskarietan, adibidez. Edo hitz-elkarketa hain emankorra duen alemaniera bezalakoetan. Honen guztiaren berri ez du inongo hiztegik ematen.

Urrutira joan gabe, imajina ezazue zenbat denbora, espazio eta kostua izango litzatekeen euskarazko hiztegi bat horrela sartzea. Batetik, ziur asko ez genituzke forma guztiak zerrendatuko eta, gainera, lexiko-sorkuntzari bideak itxiko genizkioke, morfologiaren alderdi hau hankamotz utziz. Gaur egun ingelesak ere onartzen du analisi morfologikoaren beharra.

Ikus dezagun adibide arrunt bat, ingelesa, gaztelania eta euskara erkatuz:

Ingelesa: *mountain-mountains (2 forma)*
Gaztelania: *monte-montes (2 forma)*

Euskara: 13 kasu x 4 (MG mugagabea, MS mugatu singularra, MP mugatu plurala, MPH mugatu plural hurbila) + 2 kasu (partitiboa eta prolatiboa mugagabean bakarrik) = 54. Kasu gramatikalak eta mugagabekoak bakarrik kenduta, besteei genitiboak (edutezkoa eta leku-denborazkoa) ere gehi dakizkieke, eta ondoren deklinabide osoa hasiko litzateke berriro. Beraz, teorian *ad infinitum* irits daiteke, normalean genitibo bat ala gehienez bi ezartzen badira ere.

Egia esan, tresna linguistikoak ingeleserako landu dira batez ere, baina bide honetatik hizkuntza nagusiak areago nagusituko dira, minorizatuak gero eta bazterreagoan utziz. Horrelako tresnarik gabe hizkuntza batek ezingo du besteen erabilera-maila lortu, eta zientzia munduan hizkuntzaren heriotzara darama.

(2) Lehen saioa suomierarako sistema izan zen [Broda eta beste 1980]; ondoren sorkuntza fonologikorako formalismoa [Kaplan eta Kay 1981].

	mugagabea	mug. sing.	mug. plur.	pl. hurbila
Absolutua	mendi	mendia	mendiak	mendiok
Partitiboa	mendirik			
Ergatiboa	mendik	mendiak	mendiek	mendiok
Datiboa	mendiri	mendiari	mendiei	mendioi
Inesiboa	menditan	mendian	mendietan	mendiotan
Leku gen.	menditako	mendiko	mendietako	mendiotako
Adl. s.	menditara	mendira	mendietara	mendiotara
H. adlat.	menditarantz	mendirantz	mendietarantz	mendiotarantz
M. adlat.	menditaraino	mendiraino	mendietaraino	mendiotaraino
Ablat.	menditatik	menditik	mendietatik	mendiotatik
Genit.	mendiren	mendiaren	mendien	mendion
Soz.	mendirekin	mendiarekin	mendiekin	mendiokin
Instr.	mendiz	mendiaz	mendiez	mendioz
Mot.	mendigatik	mendiagatik	mendiengatik	mendiongatik
Destin.				
- Gen. g. er.	mendirentzat	mendiarentzat	mendientzat	mendiontzat
- Ines. g. er.	menditako	mendiko	mendietako	mendiotako
- Adl. g. er.	menditarako	mendirako	mendietarako	mendiotarako
- Banatz.	mendiko			
- Prolat.	menditzat			

(*EGLU* I, 431. or.)

1.2. Prozesadoreak

Hainbat prozesadore morfologiko eraiki da, baina aitortu beharra dago ia batek berak ere ez duela eredu teorikorik hartu erreferentziatzat bere pertinentzia frogatzeko. Izan dira hurbilpenak, noski, baina hurbilpen teoriko zurrunegirik ez.

Prozesadore batzuk ikusiko ditugu segidan, aipatu besterik egingo ez baditugu ere:

1.2.1. DECOMP [Allen eta beste 1987]

MITalks-eko DECOMP modulu hau sistema zaharrenetakoa da (duela 25 urte ingurukoa) eta gramatika sortzaileari (fonologiari batez ere) asko zor dio.

Analisia egiten du, ez sorkuntza. 12.000 sarrera erabili ziren Brown (Kucera & Francis 1967) corpusetik hartuta eta ingeleseko 120.000 testu-hitz trata ditzake, % 95 ondo aztertuz.

1.2.2. KEÇI [Hankamer 1986]

Turkierarako landu zen estatu finituzko analizatzaile morfologiko hau ondo atera izana landutako hizkuntzarekin oso lotua dago. Ez dirudi eranskariak ez diren beste hizkuntzetan hain emaitza onik emango zukeenik.

Erro bakoitza kategoria gramatikalarekin listatuta dago eta afixuek zein kategoriari lotu eta zein kategoria lortzen duten esaten digute.

1.2.3. AMPLE [Weber eta beste 1988]

Esplorazio morfologikorako tresna da hau, ketxuerarako egindako parser morfologiko baten ondorio. Infijaziorako eredurik garbiena da kontzeptualki, baina morfemen azaleko okurrentzia guztiak listatu egin behar dira.

1.2.4. Gaztelaniarako analizatzaile morfologikoa [Tzoukermann eta Liberman 1990]

Eredu oso sinplea baina efizientea, ez da oso interesgarria ikuspegi linguistikotik. Benetako selekzio morfologikoak fonologikoki baldintzatutako alternantziekin nahasten dituzte.

1.2.5. MARS (Morphological Analysis for Retrieval Support) [Meya 1987]

SIEMENS etxeko MARSek corpuseko terminoen arteko erlazio morfologikoen azterketa egiten du gaztelaniarako. Flexiorako oso erosoa ez dirudien arren, eratorpena eta hitz-elkarketa lantzeko sistema egokitzat hartzen dute; alemanerako ere prestatu dute.

1.2.6. AM [Martí 1987]

MARS bezala, gaztelaniarako sortutako analizatzaile morfologikoa da, datu-basetik informazioa berreskuratu ahal izateko baldintzez osatutako automata Marcoviarraz baliatzen dena.

1.2.7. ATEF [GETA 1982]

Edozein hizkuntzari aplika dakiokeen analizatzailea da Grenobleko GETA taldeko ATEF tresna. Analisia erraz lot daiteke analisi sintaktikoarekin. Bestalde, azaleko mailan bakarrik mugitzen denez, hiztegia ere horretara behartzen du eta alomorfoak listatu egin behar dira.

Sistema hau baliatuz euskararako saio txiki bat egin da, baina ez da oso malgua morfologiarako, nahiz sintaxirako egokiagoa dirudien [Arregi eta Urkia 1989].

1.2.8. Bi Mailatako Morfologia (Two-Level Morphology) [Koskenniemi 1983]

1981ean Kaplanek eta Kay-k [Kaplan eta Kay 1981] sorkuntza fonologikorako asmaturiko gramatika morfografemikoan oinarritu zen Koskenniemi suomierarako formalismo hau garatzeko. Atal berezia eskainiko diogu ondoren honi, eredurik arrakastatsuena izan baita azken urteotan. Euskararako egindako aplikazioaren berri ere emango dugu.

Badira beste hainbat formalismo ere: Helberg (suedierarako), Karlsson (suomiera), Pisako talde bat eta IBM (gaztelania), EEEko EUROTRA proiektua, Bear, Ritchie eta beste (ingelesa), Trost, etab.

1.3. Zer aportatu dio morfologia konputazionalak morfologia teorikoari?

Arestian aipatu dugun bezala, gaurko beharrei erantzuteaz gain, hizkuntzaren (kasu honetan, morfologiaren) sistematizazioari ateak ireki eta behartu egin du horretara. Are gehiago, normalizazio bidean dauden hizkuntzak, euskara kasu, behartu egiten ditu erabakiak hartu eta guztia finkatzera. Inoiz planteatu gabeko arazoei

aurre egin beharrak hizkuntza bera aztertzera darama. Ikusiko ditugu zenbait adibide euskararako analizatzaileari buruz mintzatzean.[3]

1.4. Bestalde, zer aportatzen dio morfologia teorikoak morfologia konputazionalari?

Oinarrizko materiala, datuak batez ere, informazioa jaso eta datu horien sistematikotasunaren muturrera(ino) eramateko.

Morfologiaren teoria desberdinak ere iturri dira abiapuntu informatikoaz gain teoria horien araberako lan-tresnak prestatzeko.

Morfologia konputazionalak teoriatik datorren informazioa kodetu eta egokitu behar du. Normalean morfologia konkatenatibo soila landu izan da, erakargarriena, bestalde, morfologia konputazionalarentzat.

Baina artizkien tratamenduari, murriztapen sintaktiko eta semantikoei ere erantzun behar die, finean *hitza* aztertzen baita, zurigune arteko letra-segidako osagaiak, alegia.[4]

1.5. Morfologia konputazionalaren arazoak

Hainbat arazo planteatzen du gaur morfologia konputazionalak. Moreno Sandovalek [Moreno Sandoval 1991] hiru alderditatik ikusita sailkatzen ditu:

1. Fenomeno morfologikoen deskribapen linguistiko zehatz baten hutsunea.
2. Hizkuntzaren erabileran fluktuazioak. Hau aplikazioaren araberakoa ere bada.
3. Muga informatikoak: datuak gordetzeko ahalmena, programazio-hizkuntzak,...

2. Bi mailatako morfologia (BMM) [Koskenniemi 1983]

2.0. 1983an Koskenniemi-k bi mailatako morfologiaren eredu konputazionala definitu zuen. Eredu honek harrera bikaina izan du ondorengo urteetan, besteak beste, dituen ezaugarri hauengatik:

—Eredu orokorra da, edozein hizkuntzari aplika dakiokeena.

—Baliagarria da hitzen analisi morfologikorako zein hitz-sorkuntzarako.

—Ezagutza linguistikoa eta algoritmoa bereizi egiten ditu eta, ondorioz, programa berak edozein hizkuntzatarako balio dezake.

—Analizatu edo sortuko den hitzaren azaleko maila eta hiztegiko sisteman (sistema lexikoan) errepresentatzen den maila lexikoa[5] edo sakonekoa argi eta garbi

(3) Morfema hutsak, mugatasun lexikala, morfemen segmentazioa, hauen metaketa eta azken emaitza, hitzaren barruko elipsia, urruneko menpekotasuna, etab.

(4) "Hitza" zer den argitzea beharrezkoa da, konputazionalki nahikoa finkatua dago-eta. Sproat-ek [Sproat 1992] dioenez, hitzaren egitura norberaren begiratuaren araberakoa da. Hitza zer den horren arabera erabakiko da. Ez da berdin izango lexema, hitz lexikografikoa edo hitz ortografikoa aukeratzea. Morfologia konputazionalean input ortografikoa erabiltzen da, hau da, zurigune artekoa. Adiera horrekin erabiliko dugu guk ere aurrerantzean "hitza" terminoa.

(5) Koskenniemiren terminologia erabiliko dugu aurrerantzean. Berak azaleko maila eta maila lexikoa bereizten ditu, Gramatika Sortzailearen azaleko egitura eta sakoneko egitura adierazteko gutxi

bereizten ditu. Hau dela-eta, ez dago aldaketa morfofonologikoengatik sortutako morfema baten forma desberdinak gorde beharrik.

—Fonologia sortzaileko berridazketa-erregelen ordez erregela paraleloak erabiltzen ditu, sistema kontzeptualki zein konputazionalki errazago bihurtuz.

Morfologia honek oinarrizko bi osagai ditu: sistema lexikoa eta erregelak.

2.1. Sistema lexikoa

Sistema lexikoak morfema-multzoa definitzen du, morfemen artean egon daitezkeen kateamenduen arabera sailkapena eginez. Azpilexikoen multzoa eta erroen eta afixuen sekuentzia posibleak erregulatzen dituzten jarraitze-klaseek sistema hau osatzen dute.

Azpilexikoek ezaugarri berdineko elementu lexikoak (atzizkiak, aurrizkiak, lemak...) biltzeko balio dute. Azpilexiko guztiek egitura bera dute, identifikatzen dituen izena eta sarrera-multzoa. Sarrera bakoitzak hiru eremu ditu:

—*Adierazpen lexikoa*, karaktere-sekuentzia bat da. Karaktere hauek azaleko karaktereak edo hautapen-markak izan daitezke. Azkeneko hauei azaleko beste karaktere batzuk egokitu lekizkieke erregelen bitartez.

—Dagokion *jarraitze-klasea*. Zenbait azpilexiko edota beste jarraitze-klase batzuk biltzen dituen identifikadorea da. Jarraitze-klasean biltzen diren osagaiak dira definitutako sarreraren atzetik ager daitezkeen bakarrak.

—Sarrerari dagokion *informazio morfologikoa*.

Beraz, jarraitze-klaseak hitz batean ager daitezkeen morfemen arteko konbinaketa posibleak definitzeko mekanismoaren oinarria dira. Hauen adierazpen-ahalmena txikia da eta, horregatik, kasu batzuetan beharrezkoa izango ez litzatekeen zenbait deskripzio-bikoizketa gertatu ohi da, morfemen arteko distantzia handiko menpekotasuna (*long-distance dependencies*) tratatzerakoan, esaterako.[6]

2.2. Erregelak

Bi mailatako ereduak errepresentazio lexikoa eta azalekoa erabiltzen ditu. Lexikoak erroen eta afixuen errepresentazio morfofonologikoak dauzka.

Bi errepresentazioen artean ez dago tarteko egoerarik, eta hauxe da fonologia sortzailearekiko desberdintasun nagusia [Antworth 1990].[7] Beraz, hitzen azterketa azaleko formari dagozkion errepresentazio lexiko onargarriak aurkitzean datza. Alde-

---

gorabehera. Desberdintasuna zera da: Gramatika Sortzailean transformazio bidez pasatzen da egitura batetik bestera eta BMMan, aldiz, ez dago tarteko egoerarik (Ik. 7. oharra).

(6) Koskenniemiri kritikatu izan zaion puntuetako bat da hau, gaur neurri batean konponduta dagoena Trost-i [Trost 1990] esker. Honek morfologia ez-konkatenatiboa tratatzen du, Umlaut-aren kasua bereziki.

(7) Gramatika Sortzaileko transformazio-erregelen ordez erregela deklaratiboak darabiltza Koskenniemik; t—> c/_i beharrean t:c=>_i bezala errepresentatzen dira, tarteko errepresentazio mailarik gabe. Eta norabide bakarrekoak izan beharrean bidirekzionalak dira. Urrunketa hauen arrazoiak bi dira nagusiki: arrazoi pragmatikoak eta teoria baino gehiago tresna izatea, alegia.

rantziz gertatzen da sorkuntzan, errepresentazio lexiko ezagunetik abiatu eta berari dagozkion azaleko errepresentazioak bilatzen baitira.

Hiru osagai dituzte erregelek:

—*Korrespondentzia*, edo karaktere-bikote bat, lehenengoa lexiko-mailakoa eta bigarrena azaleko mailan aurrekoari dagokiona.

—*Testuingurua*, korrespondentzia gertatzen deneko kasuak mugatzen dituena, aurreko eta ondorengo karaktereen arabera.

—*Eragilea*, testuinguruaren eta korrespondentzian adierazitako bikotearen artean zer-nolako erlazioa dagoen finkatzen duena. Era desberdinetakoa izan daiteke: testuinguru-murriztapena (=>), azaleko koertzioa (<=), biak batera (<=>) edo debeku-ezarpena (/<=).

l:i	=>	b:b	_	e:e
korrespondentzia	eragilea	ezkerreko testuingurua		eskuineko testuingurua

Erregela hauen hasierako sintaxiak aldaketa batzuk izan ditu ([Koskenniemi 1985], [Dalrymple eta beste 1987]) konpiladore bat inplementatu ahal izateko eta, honela, automatetarako itzulpena eskuz egin behar ez izateko.

Koskenniemik analizatzaile morfologikoa egiten duen arren, fonologia konputazionalean ere ekarpen bikaina egin du. Hau da, hain zuzen ere, Moreno Sandovalek eta beste zenbaitzuk egin dioten kritikatako bat, alegia, morfologia baino fonologia gehiago ote den formalismo hau.

3. Euskal morfologia: Bi mailatako ereduaren euskararako egokitzapena

3.0. Euskara edozein azterketa teorikorako erakargarria den bezala, ordenadore bidez lantzea erronka bilakatu da, eta honen lehen emaitza analizatzaile morfologikoa izan da, esan bezala, beste azterketa batzuetarako oinarrizko tresna delako, besteak beste.

Hain esparru geografiko mugatuan hainbeste urtez iraun duen hizkuntzak edozein erakar dezake, eta linguistika konputazionalean ere hala gertatzen da. Etxeparek euskarazko lehen liburuaren inprimatzea ospatu zuen bezala, ordenadore bidez lehen testua analizatzea edo sortzea ospatuko duen Etxepare modernoarentzat bidea eta tresnak prestatzea gero eta gertuago daukagu.

Hizkuntza eranskarietatik gertu dagoen hizkuntza da euskara eta, honenbestez, testu-hitz bakarrean funtzioaren berri ere ematen digu. Hau aipatzea garrantzitsua da, euskal morfologia hutsa baino euskal morfosintaxia ere egin daitekeelako (morfologia konputazionalaren ikuspegitik ari gara, noski).

Aipatu beharrik ez dago elkarren segidan datozela lema, determinatzailea, numeroa eta deklinabide-kasua, ordena honetan eta independenteki. Deklinabide-taula ere bakarra du eta sintagmako azken osagaiak bakarrik hartzen du mugatasuna, numeroa eta kasua.

Sorkuntza lexikoari dagokionez, eratorpenak eta hitz-elkarketak ere baliabide handiak eskaintzen dituzte euskaraz.

Ez dugu ezer berririk esan honekin, baina euskara konputazionalki lantzeko kontuan izan behar ditugu hauek guztiak, hizkuntzaren arabera formalismo sinpleagoa edo konplexuagoa aukera daitekeelako.

Horrez gain, bada beste arazo bat: euskal morfologia konplexua izateaz gain, erabaki asko hartu gabe dago oraindik eta erantzun bat eman behar zaie, ordenadoreak ez baitu ezagutzarik. Halaber, aurretik ez zegoen formalizazio morfologiko konputazionalik euskararako[8] eta zerotik hasi behar izan dugu.

Aipatu bezala, euskararen ezaugarriak kontuan izanda, formalismo ahaltsu bat behar genuen automatizazioari ekiteko. Bi Mailatako formalismoa hainbat hizkuntzari arrakastaz[9] aplikatu zaiola jakitea tentagarria zen, baina suomiera[10] edo arabierarako egokia zela ikustean saio bat egitea erabaki genuen.

3.1. Euskararako egokitzapena

Ondoren datorrena da Bi Mailatako Formalismoaren euskararako aplikazioa eta emaitza. Oraingoz morfologia flexiboa bakarrik dago landua, hitz-elkarketa eta eratorpena prestatzen ari bagara ere.[11]

Formalismoa aipatzean ikusi dugun bezala, hizkuntzak eskaintzen dituen datuak hartu eta morfemak azpilexiko eta jarraitze-klaseetan sailkatu ditugu, gero erregelak finkatzeko.

3.1.1. Sistema lexikoa

Sistema lexikoan hiru osagai nagusi aipatu ditugu:

a) **Azpilexikoak**: hiztegia eta elementu gramatikalak biltzen dituztenak.

1. *Hiztegia*: hiztegi batean sarrera gisa agertzen den orok du hemen lekua: izenak, adjektiboak, aditzak, adberbioak, izenordainak, erakusleak, hitz anitzeko unitate lexikalak,[12] izen bereziak (pertsona-izenak, leku-izenak, erakundeak,...), zenbakiak, laburdurak, lokailuak eta abar.

(8) Bada Joseba Abaituaren tesia [Abaitua 1988], euskararako gramatika lexiko-funtzional bat proposatzen zuena, baina morfologiaren tratamendua oso oinarrizkoa zen, izan ere, sintaxia zuen helburu.

(9) Ingelesa, arabiera, frantsesa, japoniera, errumaniera, turkiera, etab.

(10) Suomiera hizkuntza eranskaria da. Euskaraz unitate lexiko bakoitzak erro bakarra duen bezala, suomieraz lau erro desberdin ere aurki daitezke kasuaren arabera. Ikus, esaterako, Koskenniemiren adibidea [Koskenniemi 1983]:

rakkaus	nom.sg.
rakkauteen	ill.sg.
rakkautta	part.sg.
rakkauksiin	ill.g.

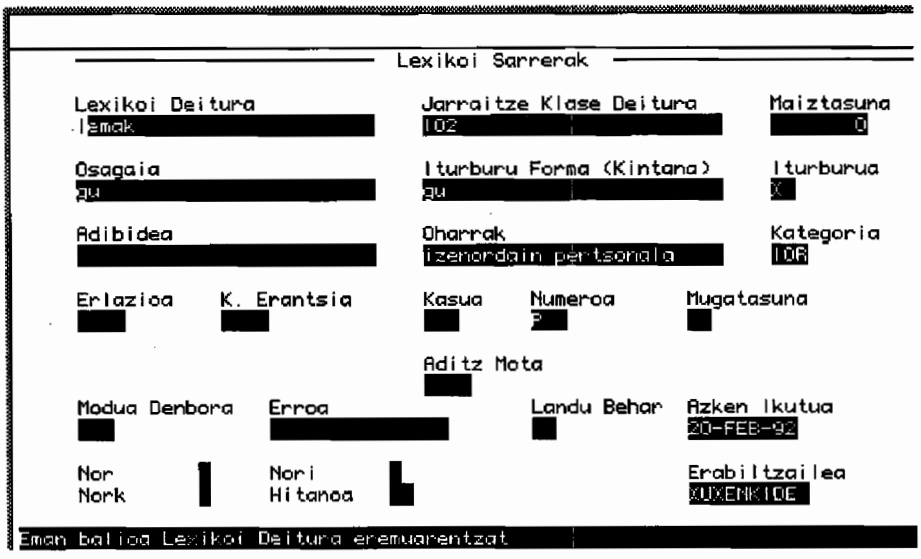
Deklinabideak antzeko funtzionamendua dauka bi hizkuntzetan, hitz-elkarketak eta eratorpenak bezala.

(11) Eratorpenari dagokionez, *ber-/bir-* aurrizkia eta hainbat atzizki (*-pe*, handigarri eta txikigarriak (*-txo*, *-tzar*), aditz faktitiboa) landu ditugu. *Izen + izen* motako hitz elkartuak (zabal jokatuta) ere azter ditzake sistemak.

(12) Tratamendurako oinarri bezala zuriune arteko elementua hartzen dela esan dugu lehen, baina zenbait sarrera beharrezko zela ikusita, azpimarraren bidez landu dira (*sistema_eragile*). Dena den, euskararako lematizatzaile automatiko orokor bat ere prestatzen ari gara (EUSLEM) eta sistematikoki landuko dira hitz anitzeko sarrerak.

Hala ere, kasu gutxi batzuetan alomorfo guztiak sartu behar izan dira morfema-deskonposaketarako arazoak medio (*hau*, baina *hon-*,...).

Datu hauek guztiak jasotzeko datu-base lexikal bat osatu da, EDBL (Euskararako Datu-Base Lexikala) [Agirre eta beste 1994], egun 50.000 sarreratik gora duena.



2. *Adizkiak* [Euskaltzaindia 1973]: formalismoaren filosofiarekin bat ez badator ere, adizkiak zerrendatuta sartu dira lehen fase batean, hau da, aditz laguntzaileak zein adizki trinkoak, hauek hartzen dituzten atzizkiak aparte dauden arren.

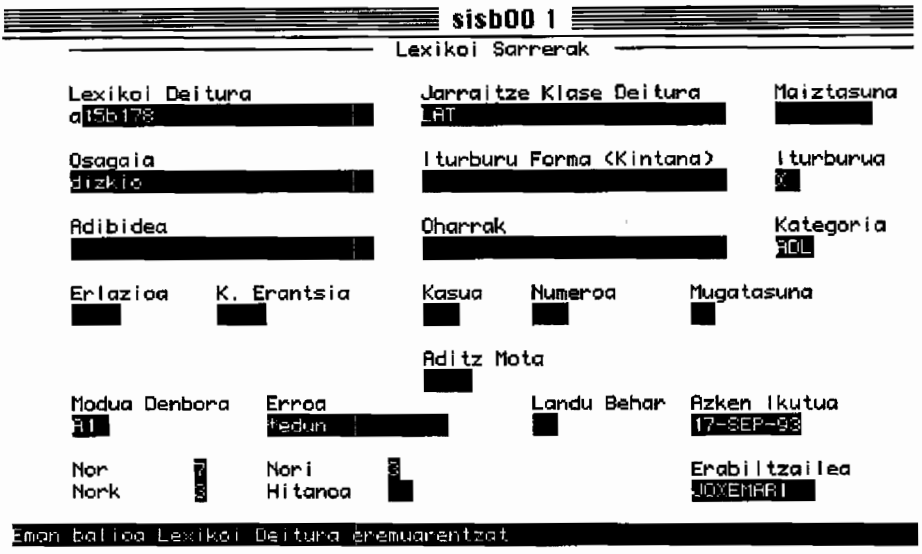
Morfema-banaketa teorikoki landua dago, baina inplementatu gabe oraingoz. Izan ere, oso zatikatuta daude morfemak eta sistema dezente moteltzen du. Morfemen arteko urruneko menpekotasun da arazo nagusia, lehenago ere aipatu dena. Hain zuzen ere, *jarraitze-klase hedatuak* sortu behar izan dira arazo hauek konpondu ahal izateko (ik. jarraitze-klaseak).

Hauek ematen duten informazioa honakoa da: kategoria, oinarri-forma, modua/denbora, nor, nori, nork pertsona, hitanoa.

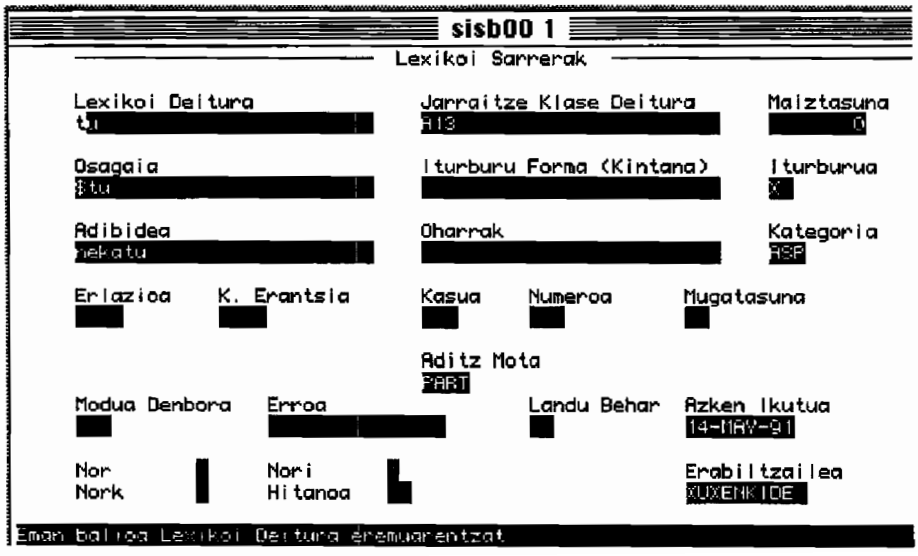
3. *Egitura morfotaktikoak* [Euskaltzaindia 1985, 1991]: morfema ez-independenteak dira hauek: deklinabide-atzizkiak, erlazio-morfemak, aurrizki eta atziki lexikalak,[13] etab.

Batzuetan "0 morfemak" edo morfema hutsak ere sartu dira hemen, hots, lekuzko kasuetan, esaterako, ez da mugatasunaren berri esplizituki ematen, azaletik begiratuta behintzat. *Etxeko*, adibidez, ez dugu *etxe a n go* bezala aztertzen, *etxe 0 ko* bezala baizik. 0-k, hemen, M(ugatu) S(ingularra) adieraziko luke eta horren berri eman.

(13) Esan dugu oraingoz kasu puntualak landu ditugula. Ik. 11. oharra.



Sarrera hauen multzoa azpilexikotan banatuta adierazten da. Azpilexikoetako sarrera guztiek beren jarraitze-klasea eta informazio morfologikoa daramate. Hala ere, zenbait kasutan informazioa morfosintaktikoa ere izango da, aurrerago ikusi ahal izango dugun bezala.



b) **Jarraitze-klaseak:** sarrerek ondoren eraman ditzaketen morfema ez-independenteak biltzen dituzten identifikadoreak dira. Beste jarraitze-klase batzuek eta azpilexikoek osatzen dituzte.

Jk_Deitura: ADJK

I1	*handi, -a, -ek, -agan*
adj_ago	*handiago...*
adj_egi	*handiegi...*
adj_en	*handien...*
grad2	*handitxo...*

Koskenniemiren lehenengo sisteman ez bazetorren ere, urruneko menpekotasunak ('long distance dependencies') konpontzeko *jarraitze-klase hedatuak* sortu ditugu. Horrela debekuak zehazten dira.

(@BAIT, (PERTSONA -LA -N))
**bainaizela*

hau da, *bait-* aurrizkiak ez dezake *-la* edo *-n* atzizkirik eraman ondoren.

Analisi morfologikoa, beraz, elementuen deskonposaketa burutu ahala hiztegitik datuak atera eta informazio hori biltzean datza.

c) **Informazio morfologikoa/morfosintaktikoa**

Eskaintzen den informazioaren berri azpilexikoen sailean eman da, nolabait lotura finkoagoa izateagatik. Hala ere, ez da beti informazio morfologiko soila ematen (Koskenniemiren eredua horretara mugatua badago ere). Gure aplikazioan sintaxira hurbiltzeko saio bat egin da.[14]

Adib.: Erlazioa *du3* *LAT*

Jk_Deitura: LAT

I0
la
lako
larik
n1
n3
na
nean
nentz
nerako
netik
nik

(14) Egia esan, gure morfologia horrelakoa izanda, ez du lan gehiegirik eskatu honek, eremu berri batean funtzioa gehitzea besterik ez da-eta. Deklinabide-atzizkiek bai formaren bai funtzioaren berri ematen digute, baina analisi morfologikoa egitera mugatzen garenean kasuen deskribapen hutsa egiten da, funtzioaren berri eman gabe (ik. aurrerago *egitera* adibidea).

non

Lx_Deitura: la		/ / /o/
2la	I0	
	ERL	
	Laster etorriko dela esan dit.	
2la	I0	
	ERL	
	Prakak urraturik zituela etorri zen.	
2la	I0	
	ERL	
	Autobusean zihoala gogoratu zen.	

diren.

3.1.2. Erregelak

Transformazio morfofonologikoen eraginez sakoneko eta azaleko mailan sortzen diren aldaketak adierazteko erregelak erabiltzen dira. Aipatu ditugu lehen erregela sortzaileekiko desberdintasunak (ik. 7. oharra). Erregela hauek hiru transformazio-mota adierazten dituzte: eransketa, ezabaketa eta trukaketa.

a) *Eransketa*: Bokal eta kontsonante epentetikoen eransketa:

Adib.: 2. erregela. "e" epentetikoa erantsi.

```
E:e <=>  %%: %+: %-: MM _ ;
          [ #: g a u | Konts (%.:) | R: ] (%+: %-:) MM _ ;
         Q: MM _ z ;
          %$: MM _ ;
E:e =>   [ Txis | AlboSud | R: ] %%: MM _ [LehGor | r | Txis | :n] ;
         %/: (%+: %-:) MM _ ;
         %!: MM _ ;
E:e <=   [ Afrik | AlboSud ] %%: MM _ [LehGor | r | Txis | :n] ;
         Txis MM %% _ [ Txis |  :n ] ;
         %!: MM _ \k:g ;
```

Adib.: *usurbil%+Eko:*usurbileko
*eibarR%+Etik:*eibarretik
polit+Ean:politean
nau$+En:nauen
zeQ+Ez:zerez
haQ+Ek: hark
izaN+Etik:izanetik
bi+garren!+Eko:bigarreneko, bigarrengo
duE+Ela:duela
zaituE+En:zaituen

Har dezagun erregela honen ondoko zatia:

E:e => [... | R:] %%: MM _ [...]

Zati honetan aditzera ematen da lexiko-mailako E hautapen-markari azaleko "e" egokitzen zaionean, halabeharrez aurreko karaktereak —ezkerreko testuingurua— R % hautapen-markak direla eta morfema-muga (MM); eskubian, aldiz, leherkari gorra, dardarkaria, txistukaria edo azalean "n" bezala gauzatzen den karaktere bat joango da.

Adibidea: **zuhaitzeko**

Lexiko-maila : *eibaR%	+Eko
Azaleko maila : *eibarr	0eko (R-ren erregela ere aplikatuko da dagokionean).

b) *Ezabaketa*: Karaktere baten desagerketa bi elementu elkartzen direnean:

Adib.: 4. erregela. "a"ren galera.
A:0 => i _ #: ;
%#:0 => _ #: ;
%&:0 <=> _ MM LekKas ;
a:0 <=> :a +:= _ r: a: z ;
a:0 => r a +:= a:0 r:0 _z ;

Adib.: amA+a:ama
*azpeiti&+Era:*azpeitira
gelA+Ean:gelan
filosofiA:filosofi
kultur#:kultur
argitara+araz+i:argitarazi
(kontuz: *i , kultura, hizkuntza, literatura, natura, eliza, burdina bakarrik)

Erregela-atal hau hartuta:

a:0 => r a +:= a:0 r:0 _ z ;

Erregela honek dioenez, lexiko-mailan "a" badator, azaleko mailan hutsa dagokio (Ø) baldin eta aurretik lexiko-mailan "ra" eta morfema-marka (+) badatoz, azalean gauzatzen ez diren "a" eta "r"rekin batera, eta ondoren "z".

Adibidea: **argitararazi**

Lexiko-maila:	argitara	+araz	+i
Azaleko maila:	argitara	0000z	0i

c) *Trukaketa*: Kontsonante gorra ozen bilakatzea sudurkari baten eraginez, etab.; bokalen arteko aldaketak.

Adib.: 7. erregela: k-ren ozenketa sudurkariaren ondoren.
k:g <=> [AlboSud %%: | :n | %!:] MM (E:0) _ o;

Adib.: *usurbil%+Eko:*usurbilgo
*irun%+Eko:*irungo
egiN+ko:egingo
hemen+ko:hemengo
bi+garren!+Eko:bigarrengo

Adibidea: **egingo**
L.M.: egiN +ko
A.M.: egin 0go

24 erregela ditugu gure sisteman, baina erregela bakoitza, berez, erregela-multzo bat da. Hain zuzen ere, karaktere-aldaketaren arabera biltzen dira, fenomeno arras desberdinak bilduz.[15]

3.1.3. Datu hauek egituratu ondoren, sistema prest dago sistema lexikoan dagoen edozein formaren azterketa egiteko, hau da, bai forma baten analisia, bai lematik abiatutako sorkuntza.

```
Adib.: Analisia
   ( (forma "etorrarazten")
     ( (anal 1)
       ( (lema "etoR")((KAT ADI)))
          ( (morf "araz")((KAT ADF)(MDN INF)(ERR araz)))
          ( (morf "ten")((KAT ASP)(ADM EZBU))))
   )
```

Adib.: Sorkuntza

1	etxe ->	etxe	20	etxe+Ez ->	etxez
2	etxe+a ->	etxea	21	etxe+Ek ->	etxek
3	etxe+az ->	etxeaz	22	etxe+Ri ->	etxeri
4	etxe+ari ->	etxeari	23	etxe+Rik ->	etxerik
5	etxe+ak ->	etxeak	24	etxe+tzat ->	etxetzat
6	etxe+areM ->	etxearen	25	etxe+ReM ->	etxeren
7	etxe+oM ->	etxeon	26	etxe+pe ->	etxepe
8	etxe+oz ->	etxeoz	27	etxe+tzaR ->	etxetzar
9	etxe+otaz ->	etxeotaz	28	etxe+txo ->	etxetxo
10	etxe+ok ->	etxeok			
11	etxe+ok ->	etxeok	6	etxe++Etik ->	etxetik
12	etxe+oi ->	etxeoi	7	etxe++Era ->	etxera
13	etxe+ez ->	etxeez	8	etxe++Ean ->	etxean
14	etxe+etaz ->	etxeetaz	9	etxe++Eko ->	etxeko
15	etxe+ek ->	etxeek	10	etxe+areM ->	etxearen
16	etxe+ei ->	etxeei	11	etxe+areM+a ->	etxearena
17	etxe+ak ->	etxeak	12	etxe+areM+az ->	etxearenaz
18	etxe+eM ->	etxeen	13	etxe+areM+ari ->	etxearenari
19	etxe+ ->	etxe	14	etxe+areM+ak ->	etxearenak
			31	etxe+areM+Rik ->	etxearenik

(15) Alternantzia fonologiko, morfologiko, morfonologiko eta hainbatetan ortografiko hutsak ere biltzen dira guk erregela deitutako multzoetan. Fenomeno desberdinen sailkapen bat egiten saiatu gara, Varela [Varela 1990] eta Matthews-en [Matthews 1980] irizpideen arabera, eremu irristakorra bada ere.

3.2. Analisiaren emaitza

Oraingoz sorkuntza alde batera utzita, osagaien metaketaren ondorioz lortzen dugun analisiaren emaitza azter dezakegu eta azaleratzen diren hainbat arazo aurkeztu, batzuetan guk eman diegun irtenbidearekin. Soluzioak baino gehiago arazoak planteatu nahi ditugu atal honetan, hauen teorizaziorik apenas aurkitu dugu-eta.

Zein da osagaien segmentazioaren eta metaketaren ondorioz lortzen dugun azken emaitza?

1. Ondoko erabileraren araberakoa izango da, hots, analizatzaile morfologikoak forma osatzen duten osagaien berri eman beharko du, azken emaitza nahikoa izanez. Hala ere, semantika lantzen denerako interesgarria dateke, eratorpenaren kasuan batez ere, azpikategorizazioa mantendu eta datuak goratzea azken emaitzarekin batera. Baina gatozen lehenengo aukerara oraingoz.

Normalean kategoria, mugatasuna, pluraltasuna eta kasua emango digu analisiak, edo lehen aipatutako adizkien informazioa, etab.

mendi IZE
mendia IZE ABS M S

Gogora dezagun berriro karaktere-segida baten analisia ari garela egiten, eta *mendi* IZE dela besterik ez dugu esango.[16]

2. Zenbaitetan analisi bat baino gehiago izango dugu, noski:

```
Adib.:    amak
        I Z E N A : amak
            ((forma "amak")
             ((anal 1)
             ((lema "amA")((KAT IZE)))
              ((morf "ak")((KAT DEK)(KAS ABS)(NUM P)(MUG M))))
              ((anal 2)
              ((lema "amA")((KAT IZE)))
              ((morf "ak")((KAT DEK)(KAS ERG)(NUM S)(MUG M))))
             ((anal 3)
              ((lema "amA")((KAT IZE)))
              ((morf "Ek")((KAT DEK)(KAS ERG)(MUG MG))))
    )
    — hitz zuzena.
```

(16) Baina, independenteki, sintaxia lantzen hasi garen orduko ikusi ditugu arazoak:

mendi	IZE
itsasoa *mendi* bilakatu zen	IZE ABS MG
zenbait *mendi*	IZE ABS MG

```
Adib.:    egitera
        I Z E N A : egitera
        ((forma "egitera")
         ((anal 1)
          ((lema "egiN")((KAT ADI)))
          ((morf "tera")((KAT ERL))))
         ((anal 2)
          ((lema "egite")((KAT IZE)))
          ((morf "0")((KAT DEK)(NUM S)(MUG M)))
          ((morf "Era")((KAT DEK)(KAS ALA))))
         ((anal 3)
          ((lema "egiN")((KAT ADI)))
          ((lema "te")((KAT ASP)(KER IZE)))
          ((morf "0")((KAT DEK)(NUM S)(MUG M)))
          ((morf "Era")((KAT DEK)(KAS ALA))))
    )
    — hitz zuzena.
```

segun formari ala funtzioari begiratzen diogun.[17]

3. *Mugatasuna* (guk *mugatasun lexikala* deitu duguna): izenordainek, erakusleek, pertsona- eta leku-izen bereziek, siglek, etab.ek, barruan, leman daramate mugatasunaren informazioa. Baina eransten zaizkien atzizkiak ez datoz beti bat horrekin. Adibidez, izen propioek mugagabearen taula hartzen dute. Beraz,

```
Adib.1:  AitorrI *(IZB DAT MG)
         I Z E N A : *aitorri
         ((forma "*aitorri")
          ((anal 1)
           ((lema "*aitoR")((KAT IZB)))
           ((morf "Ri")((KAT DEK)(KAS DAT)(MUG MG))))
     )
    — hitz zuzena.

Adib.2:  guk *(IOR ERG MG)
         I Z E N A : guk
         ((forma "guk")
          ((anal 1)
           ((lema "gu")((KAT IOR)))
           ((morf "Ek")((KAT DEK)(KAS ERG)(MUG MG))))
     )
    — hitz zuzena.
```

mugatasunaren berri leman bertan ematea litzateke irteera bat, eta erregela bidez konpondu gero mugagabea hartzen duenean:

(17) Ikus 14. oharra.

Aitorrek *Aitor (IZB M S) + ri (DAT MG)*
Mug1 + Mug2 => Mug1

Kasu hauetan ezezik, ez dago mugatasun- edo pluraltasun-metaketarik, elipsirik ez dagoenean.

4. *Elipsia*: Noiz gertatzen da elipsia? Genitibo leku-denborazko edo edutezko baten ondoren kasu gramatikalak, beste genitibo bat, instrumentala, inesiboa (-biz), ablatiboa (-biz), adlatiboa (-biz), partitiboa edo prolatiboa datozenean.

x-ko (x)ren (x)a = liburuko (pertsonaia)ren (aurpegi)a
I Z E N A : liburukoarena
((forma "liburukoarena")
((anal 1)
((lema "liburu")((KAT IZE)))
((morf "0")((KAT DEK)(NUM S)(MUG M)))
((morf "Eko")((KAT DEK)(KAS GEL)))
((morf "areM")((KAT DEK)(KAS GEN)(NUM S)(MUG M)))
((morf "a")((KAT DEK)(KAS NOM)(NUM S)(MUG M))))
)
— hitz zuzena.

Gaur honela emango genuke analisia, atalak gordeta:

```
(  (forma "liburukoarena")
   (  (anal 1)
      (  (osagai 1)
         (  (oina  (lema "liburu")
                   (twol "liburu")
            )
            ( (KAT IZE) )
         )
         (  (morf  (lema "0")
                   (twol "0")
            )
            ( (KAT DEK)(NUM S)(MUG M) )
            )
            (morf  (lema "ko")
                   (twol "+ko")
            )
            ( (KAT DEK)(KAS GEN) )
         )
      )
      (  (osagai 2)
         (  (oina  (lema "0")
                   (twol "0")
            )
```

```
                ( (KAT ELIPSIA) )
            )
            (   (morf (lema "aren")
                    (twol "+areM")
                )
                ( (KAT DEK)(KAS GEN)(NUM S)(MUG M) )
            )
        )
        (   (osagai 3)
            (   (oina  (lema "0")
                    (twol "0")
                )
                ( (KAT ELIPSIA) )
            )
            (   (morf (lema "a")
                        (twol "+a")
                )
                ( (KAT DEK)(KAS NOM)(NUM S)(MUG M) )
            )
        )
        ( ( (KAT IZE_EIZ)(KAS GEN)(NUM S)(MUG M) )
          ( (KAT ELIPSIA)(KAS GEN)(NUM S)(MUG M) )
          ( (KAT ELIPSIA)(KAS NOM)(NUM S)(MUG M) ) )
    )
)
```

Aldiz, ondoren soziatiboa, destinatiboa, motibatiboa, inesiboa (+biz), ablatiboa (+biz) edo adlatiboa (+biz) datozenean, ez da elipsirik gertatzen hauek banatu ezkero, bai baitago atzizki bakar bat bezala emateko aukera, alegia, *-rekin*, *-rentzat*,...

semearekin = seme a ren kin

Elipsidun kasuen berri emateak laguntza eskaintzen du analisi sintaktikoari eta semantikoari begira. Adizkiekin ere gauza bera gertatzen da:

```
Adib.: duena  du n 0 a
   I Z E N A : duena
   (forma "duena")
    ((anal 1)
     ((lema "duE")((KAT ADL)(MDN A1)(NOR 3)(NRK 3)(ERR *edun)))
     (morf "EnA")((KAT ERL)(ERL KONP))))
    ((anal 2)
     ((lema "duE")((KAT ADL)(MDN A1)(NOR 3)(NRK 3)(ERR *edun)))
     ((morf "En")((KAT ERL)(ERL ERLT)))
     ((morf "a")((KAT DEK)(KAS NOM)(NUM S)(MUG M))))
    ((anal 3)
```

```
((lema "duE")((KAT ADT)(MDN A1)(NOR 3)(NRK 3)(ERR ukan)))
(morf "EnA")((KAT ERL)(ERL KONP))))
((anal 4)
((lema "duE")((KAT ADT)(MDN A1)(NOR 3)(NRK 3)(ERR ukan)))
((morf "En")((KAT ERL)(ERL ERLT)))
((morf "a")((KAT DEK)(KAS NOM)(NUM S)(MUG M))))
)
— hitz zuzena.
```

4. analisia bakarrik metatuta:

```
(  (forma "duena")
   (  (anal 4)
      (  (osagai 1)
         (  (oina  (lema "ukan")
                   (twol "duE")
            )
            ( (KAT ADT)(MDN A1)(NOR 3)(NRK 3) )
         )
         (  (morf (lema "en")
                  (twol "+En")
            )
            ( (KAT ERL)(ERL ERLT) )
         )
      )
      (  (osagai 2)
         (  (oina  (lema "e")
                   (twol "e")
            )
            ( (KAT ELIPSIA) )
         )
         (  (morf (lema "a")
                  (twol "+a")
            )
            ( (KAT DEK)(KAS ABS)(NUM S)(MUG M) )
         )
      )
      ( (KAT AD_EIZ) ) (ERL) (ERLT)
        (ELIPSIA) (KAS ABS) (NUM S)(MUG M) )
   )
)
```

Etor daitezke elipsia eta komunztadura batera ere:

```
Adib.: *dizkiedanari
   I Z E N A : dizkiedanari
```

```
((forma "dizkiedanari")
 ((anal 1)
  ((lema "dizkiet^")((KAT ADL)(MDN A1)(NOR 7)(NRI 7)(NRK 1)(ERR *edun)))
  ((morf "En")((KAT ERL)(ERL ERLT)))
  ((morf "ari")((KAT DEK)(KAS DAT)(NUM S)(MUG M))))
 )
— hitz zuzena.
```

Hemen ez dago komunztadurarik. Murriztapenak ezarri behar dira adizki bakoitzean pertsonak kontrolatuta; bitartean, gure analisiak ontzat ematen ditu kasu guztiak.

5. *Nominalizazioa*: aditz bat nominalizatzean IZEn kategoria eman beharko litzateke besterik gabe edo azpian aditza dagoela aipatu:

```
Adib.: egite
   I Z E N A : egite
   ((forma "egite")
    ((anal 1)
     ((lema "egite")((KAT IZE))))
    ((anal 2)
     ((lema "egiN")((KAT ADI)))
     ((lema "te")((KAT ASP)(KER IZE))))
   )
   — hitz zuzena.
```

Zenbait hiztegitan sarrera ematen zaie guztiz nominalizatuta daudenei, baina teorian edozein aditz nominaliza daitekeenez, guk guztiei jarraitze-klasean kategoria aldatzeko aukera eman diegu (KER markak kategoria erantsia adierazten du).

```
Jk_Deitura:     A14
        I0      abes
        araz    abesaraz
        te      abeste
        ten     abesten
        tu      abestu
Lx_Deitura: te / / /o/
        $te     I1
                ASP   KER IZE
                      abeste
```

6. *Kategori aldaketak*: zabal jokatuta, edozein izen aditz bilaka daitekeela esango genuke (*gizon* => *gizondu*). Baina, hizkuntzak inoiz erabili ez dituen formak sortzeko aukerak ematen dizkigun arren, oraingoz ez dugu erregela hau orokortu, berez izenen jarraitze-klasean osagai bat gehitzea besterik ez bada ere. Honek ere azterketa sakonagoa eskatzen duelakoan gaude, eta semantika ere tartean egongo da dudarik gabe (*?mahaitu, ?liburutu*).

Beste arazo bat ere planteatzen da hemen: izen/adjektibo, adjektibo/adberbio eta horrelakoekin, askotan testuinguruak bakarrik zehatz dezake kategoria zein den, baina ez da kategori aldaketarik gertatzen besteen mailan; gehienez ere, hiztegian bina sarrera edo bi kategori informazio beharko lukete.

Arazo hauetako askoren teorizaziorik aurkitu ez dugunez, gure irtenbide praktikoa ematen saiatu gara, eguneraketak eta zuzenketak beti dira posible-eta. Printzipioz kategoriarik arruntena ezarri zaie eta sintaxirako IZE/ADJ eman, desanbiguaketa--erregelek testuinguruaren arabera kasuan kasuan bereiz ditzaten.

Hasierako helburua ahaztu gabe, hau da, gure formalismoak normalizazioari laguntzeko tresna izan nahi duenez, euskara batukoa da erabili den materiala eta Euskaltzaindiaren gomendio/erabakiak [Euskaltzaindia 1991], esaterako, zurrun erabiltzen dira, beste erabileretan abisua emanez. Lan honek ez du preskriptiboa izan nahi, baina deskribapen hutsa ere ez du helburu. Normalizazioa eskuratu nahi bada, gutxienez, Akademiak esandakoa "bete" behar da, bestela pertsona bakoitzeko gramatika bat eta hiztegi bat izango genuke. Analizatzaileak bere egin ditu gomendioak (*arazi*, *bait*,...).

Honek beste ideia batera garamatza: testu errealetatik abiatuz analisia egiten saiatzerakoan, hainbat aldiz emaitza ematea ezinezkoa zitzaigun, aldaera askotxo azaltzen zelako, hots, euskalki desberdinetako aldaerak, ezjakintasunez (edo nahita) Euskaltzaindiaren arauak betetzen ez zirelako, etab.

Sistema bideragarriagoa egiteko "Errore tipikoak" izeneko analizatzailea sortu genuen, desbiderapenak tratatzen dituena.

LX_OSAGAIA	ET_OSAGAIA	ET_LX_DEITURA	E_KOD	ADIBIDEA
$tu	$u	et_tu	ATZ	asmau
$araz	$eraz	et_araz	A_FAK	bileraz
$gatik	$gaitik	et_gan_tik	DE_DI	zugaitik
ahots	abots	et_izenak	LD_FO	
arno	ardo	et_izenak	DIAL	
handi	haundi	et_adjektiboak	LD_FO	
guzi	guzti	et_lemak	LD_FO	
katalisi	katalisis	et_izenak	LD_MA	
independente	independiente	et_adjektiboak	LD_MA	
proiektu	projektu	et_izenak	LD_MA	
sistema	sistima	et_izenak	LD_MA	

Erregelen sistema berria ere eraiki zen. Horrela, kasu hauetako bat azaltzen denean, adierazi egiten da zein den gaurko euskara batuko erabilera estandarra. Honen aplikazioa oso baliagarria izan da egiaztatzaile/zuzentzaile ortografikorako.[18]

(18) XUXEN (euskararako egiaztatzaile/zuzentzaile ortografikoa) [Agirre eta beste 1992], UZEIk eta Donostiako Informatika Fakultateak elkarlanean egindakoa, plazaratzear dagoena.

3.3. Zeintzuk dira sistemaren abantailak ikuspegi linguistikotik?

Koskenniemiren eredua lexiko-mailan oinarritzen da eta bi mailen arteko korrespondentzia definitzeko erregelez baliatzen da. Beraz, azpilexikoetan formak oso-ososrik, aldaketarik gabe, gordetzen dira, beste sistemetan ez bezala.

Adibidez, "ama" unitate lexikoa da, "a" organikoa du, baina datibo plurala "amei" forma da; eta horregatik sistema askotan "ama" erabili ordez "am" erabili beharko litzateke (GETAko ATEFen, esaterako). BMMan, lexiko eta azaleko maila argi eta garbi bereizten direnez, unitatea oso-ososrik eta alferriko bikoizketarik gabe manten dezakegu. Linguistikaren ikuspegitik garrantzi handikoa deritzogu analisi morfologikoaren planteamendu honek dakarren unitate lexikoarekiko errespetoari.

3.4. Aurrera begira

Orain arte azaldutakoa morfologia flexiboari dagokio batez ere, baina lexiko-sorkuntza haratago doa eta hori ere tratatu nahi genuke.

Eratorpenean, sistematizazio-saio bat egiten ari da lantaldea, bitartean eratorririk arruntenak hiztegiratuta dauden arren. Sintaxia eta semantikako hainbat murriztapen gehitu beharko da, bide batez formalismoa aberastuz.

Hitz-elkarketari dagokionez, maizenik agertzen diren konposatuak sartuta daude lehendik, lotuta idazten direnak nagusiki, eta marradunak ere ezagutzen ditu formalismoak, baina, besteak bezala, sistematikoki landu nahi da. Berriro, semantikaren beharra azaltzen zaigu eta azpikategorizazio finarena. Bestalde, printzipioz onartuko ez genituzkeen hainbat konposatu posible izango litzateke zabalegi jokatuta.

Morfologia gaindituaz, Murriztapen-Gramatika (Constraint Grammar) oinarritzat hartuta, euskararako analizatzaile sintaktikoa eraiki nahian ari gara, oinarri bezala BMM baliagarria baita. Baina ondoko lanen aurkezpena beste hitzaldi baterako gaia litzateke.

Bibliografia

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

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Sarasola, K., Urkia, M., 1989, "Aplicación de la morfología de dos niveles al euskara", *SEPLN* 8. 87-102.

Agirre, A., Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Goenaga, P., Maritxalar, M., Sarasola, K., Urkia, M., 1991, "Bi mailatako morfologiaren euskararako egokitzapena", *Elhuyar* 17. 6-14.

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K., Urkia, M., 1992, "XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology", *Proc. of the Third ANLP*, 119-125. or.

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

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

Allen, J., Hunnicutt, M. S. eta Klatt, D., 1987, *From text to Speech: The MITal System*. Cambridge University Press.

Antworth, E. L., 1990, PC-KIMMO: *A Two-level processor for morphological analysis*. Computational Linguistics.

Arregi, X. eta Urkia, M., 1989, *ATEF eta ROBRAren euskararako egokitzapena*. Barne-txostena. Argitaragabea.

Barton, E., 1985, *Computational Complexity in two-level Morphology*, MIT, Cambridge, Massachussets.

Bauer, I., 1983, *English Word-Formation*. Cambridge University Press.

Brodda, B. eta Karlsson F., 1980, *An experiment with Automatic Morphological Analysis of Finnish*. Papers from the Institute of Linguistics, University of Stockholm. Publication 40.

Dalrymple, M., Kaplan, R.M., Karttunen, L., Koskenniemi, K., Shaio S. eta Wescoat, M., 1987, *Tools for Morphological Analysis*. Report N. CSLI-87-108.

Euskaltzaindia, 1973, *Aditz laguntzaile batua*. Euskaltzaindia, Bilbo.

———, 1985-90 *Euskal Gramatika: Lehen urratsak (I, II eta III)*. Euskaltzaindia, Bilbo.

———, 1992, *Euskaltzaindiaren Gomendioak eta Erabakiak (I eta II)*, Bilbo.

G.E.T.A., 1982, *Le point sur ARIANE-78 debut 1982*. Vol.1, partie 1: *Le logiciel*. 28-75. or.

Guilbaud, J. P., 1980, *Analise morphologique de l'allemand en vue de la traduction par ordinateur de textes techniques specialisés*, Sorbonne, Paris.

Hankamer, J., 1986, "Finite state morphology and left to right phonology", *Proceedings of the West Coast Conference on Formal Linguistics*, Vol. 5 (Stanford Linguistic Association).

Kaplan, R. M. eta Kay, M., 1981, *Phonological rules and finite-state transducers*. Paper read at the annual meeting of the Linguistic Society of America in New York City.

Karttunen, L., 1983, *KIMMO: A two-level Morphological Analyzer*. Texas Linguistic Forum, vol. 22, 165-186.

Kay, M., 1973, *Morphological Analysis*. A. Zampolli & N. Calzolari eds. (1980). Proc. of the Int. Conference on Computational Linguistics (Pisa).

Koskenniemi, K., 1983, *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics. Publications 11.

———, 1985, *Compilation of Automata from Morphological Two-level Rules*, 143-149. or. Publication 15. University of Helsinki.

Martí, M. A., 1987, "Un sistema de análisis morfológico por ordenador", *SEPLN* 4.

Matthews, P., 1980 *Morfología. Introducción a la teoría de la estructura de la palabra*, Paraninfo, Madrid.

Meya, M., 1987, "Análisis morfológico como ayuda a la recuperación de información". *SEPLN* 4.

Moreno Sandoval, A., 1991, *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español* (tesia, Univ. Autónoma de Madrid).

Sarasola, I., 1982, *Gaurko euskara idatziaren maiztasun-hiztegia*. (3gn. liburukia), GAK, Donostia.

Sproat, R., 1992, *Morphology and Computation*, MIT Press, Cambridge, Massachusetts.

Trost, H., 1990, "The application of two-level morphology to non-concatenative German morphology", *COLING*-90, Helsinki, vol. 2.

Tzoukermann, E. eta Liberman, M., 1990, "A finite-state morphological processor for Spanish", *COLING*-90, vol. 3.

Varela, S., 1990, *Fundamentos de Morfología*, Ed. Síntesis, Madrid.

Weber, D., Black H. A. eta McConnel, S., 1988, *AMPLE: A tool for Exploring Morphology*, Summer Institute of Linguistics, Dallas.

Winograd, T., 1983, *Language as a cognitive process*. Vol. 1: *Syntax*, 544-549. Addison-Wesley.