ONDORIOAK ETA AURRERA BEGIRAKOAK

VII. Ondorioak eta zabaldutako ikerlerroak.

VII.1. Ondorioak.

Ikerlan honen emaitza gisa bi oinarrizko tresna eraiki dira: euskararen morfologia ezagutzen duen prozesadore sendo eta estaldura handiko bat batetik, eta prozesadore horrek burutzen duen analisi eta sorkuntza morfologikoa erabiliz zuzentzaile ortografiko bat bestetik.

Tresna horiek euskararen prozesaketa automatikorako egitasmo orokor baten barruan kokatzen dira. Bide horretan, eta egitasmo honi oinarria emateko, EDBL izeneko Euskararako Datu-Base Lexikala egituratu eta osatzeaz gain, corpus multzo bat bildu da, horren gainean maiztasunaren araberako hitz- eta trigrama-zerrendak lortu direlarik.

Prozesadore morfologikoa Koskenniemik definitutako bi mailatako morfologian dago oinarriturik. Formalismo honen egokitasuna frogatuta gelditu da, euskal morfologiaren deskribapen dotore, eroso eta malgua bideratzen baitu; eskala errealeko definizioa erraztuz. Egokitasuna eta malgutasuna frogatu da berriro Euskaltzaindiak Leioako 1.994ko kongresuan arau berriak onartu dituenean, oso lan sinplea izan baita sistema egokitzea arau berri hauei.

Bi mailatako formalismoaren barruan, morfemen arteko urruneko menpekotasuna adierazi ahal izateko, morfotaktikaren funtsa diren jarraitze-klaseen deskribapen-

ahalmena handitzen duten “jarraitze-klase hedatuak” izeneko mekanismoa proposatzen da.

Prozesadore morfologikoa hedadura handikoa izan dadin lau modulutan banatzen da; euskara estandarrerako modulua, erabiltzailearen lexikoa edo hiztegi berezituaren kudeaketarakoa, erabilera ez-estandarrak edo aldaerak tratatzeko modulua eta lexikorik gabeko prozesaketa morfologikoa bideratzen duena. Hizkuntza estandarrerako tratamendurako bi mailatako morfologiaren erabilera ohizkoa bada ere, gainontzeko moduluetan erabiltzea proposamen berritzailea da. Eta berritzaileena dena zera da: prozedura guztiak bi mailatako morfologian oinarriturik egotea, sistema homogeno eta trinkoa osatuz.

Bi mailatako formalismoaren gure inplementazio burutzea lan neketsua baina interesgarria izan da; alde batetik formalismoan sakontzeko balio izan duelako, eta bestetik, beraren gainean aldaketak eta hobekuntzak esperimentatzeko aukera eman digulako. Kode hau merkaturatu da Xuxen zuzentzaile ortografikoaren barruan.

Bi mailatako morfologiak, 1983an sortu zenetik, bilakaera garrantzitsua izan du, eta hobekuntza aipagarrienetako bat lexiko-itzultzaileena da, eraginkortasuna eta deskribapen-ahalmena izanik metodo horren abantailarik garrantzitsuenak. Ikerlan honetan metodo hori ebaluatu egin dugu, morfologia banatu dugun lau moduluetan emaitza azpimarragarriak lortuz, erabiltzailearen hiztegian aplikatzeko arazoak aurkitu badira ere.

Xuxen izeneko zuzentzaile ortografikoa izan da prozesadore morfologikoan oinarriturik egin dugun lehen produktu komertziala. Aurretik aipatutako modulu gehienak berrerabiltzen dira zuzentzaile honetan, horrela bi mailatako morfologian oinarritutako zuzentzaile “linguistikoa” lortuz.

Zuzenketa ortografikoaren problematika aztertu ondoren eta euskararen ezaugarriak kontuan hartuz, gaitasun-erroreak tratatzeko beharra da funtsezko ondorioa. Errore horiek detektatzeko aldaeren analisi morfologikorako erabilitako mekanismo bera erabiltzen da eta analisi horretan lortzen den informazioa gorde egiten da, ondoren akatsari dagokion zuzenketa lortzeko, sorkuntza morfologiko estandarra erabiliz helburu horrekin.

Errore tipografikoen tratamendua arras konbentzionala denez —bibliografia-azterketan azaldutako zailtasunen aurrean bigarren mailako lehentasuna baitzuen gai honek gure sisteman— zehaztasuna/eraginkortasuna faktoreen arteko orekan zentratu da gure lana, proposamenak sortzeko azkartze-metodo desberdinak aztertuz.

Azpimarratu behar da egindako tresnen izaera: eskala errealeko produktu bukatuak baitira eta ez prototipoak edo maketak. Prozesadore morfologikoa euskararen gaineko

edozein aplikaziotarako oinarrizko tresna den bitartean, zuzentzaile ortografikoa salgai dago eta oso harrera ona izan du.

VII.2. Zabaldutako ikerlerroak eta perspektibak.

Lan honetan zabaldutako ikerlerroak anitz dira. Alde batetik daude lanaren alde ahulenak hobetzeko bideak eta lanean zehar hausnartutako etorkizuneko hobekuntza posibleak. Beste aldetik daude proiektu zabalago batean integratuta egotetik datozen ikerlerroak, egindako tresnak beste urratsetan oinarri-tresna 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 lau multzotan banatu ditugu etorkizuneko lan hauek: lehenengo bietan ikerlan honekin zuzenean lotutako bi gaiak hartzen dira mintzagai, prozesaketa morfologikoa eta zuzenketa hain zuzen; hirugarrenean, epe laburrean burutu nahi dugun tresna, EUSLEM lematizatzaile/etiketatzailea, aurkezten da; eta, azkenik, egindako tresnak oinarritzat hartuko dituzten bestelako aplikazioak aipatzen dira.

VII.2.1 Prozesaketa morfologikoa hobetzen.

Prozesaketa morfologikoan lan sakona egin den arren, alde ahulena eraginkortasunarena dugu. Aipatu den bezala ahulezia hau konpontzeko aurrekonpilazioa da bide nagusia, Karttunen-ek (1994) proposaturiko lexiko-itzultzaileen bidea edo Carter-ek (1995) proposatutakoa interesgarrienak izanik. Lexiko-itzultzaileak erabili ditugu eta eraginkortasunaren zein deskribapen-ahalmenaren aldetik oso emaitza onak eskaintzen dituzten arren ez dira nahiko malguak erabiltzailearen hiztegiak integratzeko. Carter-en ekarpena interesgarria da malgutasunaren aldetik, azpilexiko itxiak baino ez baititu aurrekonpilatzen baina arazoak ditu lemen konposaketan. Beste aldetik, sistema hauetan ezin da jorratutako erregela morfofonologikoei buruz informaziorik lortu, eta hau interesgarria da aldaeren tratamendurako zein OLIren arloko aplikazioetarako. Azkenik, lexiko-itzultzaileetan morfotaktikak Koskenniemiren hasierako definizioari jarraitzen dio, urruneko menpekotasuna adierazteko deskribapen-ahalmenik gabe jarraituz. Ezaugarri horiek guztiak integratuko lituzkeen eredu berri bat definitzea irekitako ikerlerro bat da.

Euskararako aplikazioari dagokionean berriz, bi lan nagusi gelditu dira formalki landu gabe: aditz laguntzailea eta trinkoa eta eratorpena. Honez gain, datu-basea eguneratzen jarraitzea eta sistema martxan dagoen hizkuntzaren batze-prozesuari egokitzen joatea da etengabe burutzen ari den lana.

Aditz laguntzailea eta trinkoa hitzez hitz landu da, eta honen arrazoia bikoitza da; batetik eraginkortasuna, morfemak oso motzak eta aldaketak ugariak baitira, eta bestetik barneko morfemen arteko urruneko menpekotasuna. Arazo hauek konpondu ondoren aditz laguntzailearen deskribapen “formala” egingarri bihurtzen da.

Eratorpena gai korapilatsua da, irregularra izanik ondo aztertu gabe dagoelako. Taldeko linguistak egiten ari diren lan teorikoaren emaitzaz, emankortasun handiko morfema sinpleetan oinarritutako eratorpena landuko da, orain arte landutako eratorpen lexikalizatua osatuz.

VII.2.2 Zuzenketa.

Zuzenketan egindako lana aldaerak deitu ditugun gaitasun-erroreen aldetik oso interesgarria den bitartean, errore tipografikoen trataera bi aldetatik hobe liteke: bateko edizio-distantzia baino handiago duten hitzen zuzenketari ekinez batetik, eta bestetik, testuingurua kontuan hartuz, proposamenen artean zuzenketa ondo aukeratzeko zein “benetako hitzaren erroreak” detektatzeko asmoz.

Lehen ekimenean oso abiapuntu interesgarria da Oflazer eta Guzey-rena (1994), emaitza onak lortzeko oraindik eragiketa asko burutu behar badira ere, honek eraginkortasunean duen ondorio kaltegarriarekin. Ekarpen hori hobe daitekeelakoan gaude, lexikorik gabeko analisiak horretan lagun dezakeelarik.

Proposamenen artean zuzena zein den erabakitzea oso zaila gertatzen da testuingurua kontuan hartzen ez bada. Hori egiaztatzeko eskuz egitea baino ez da egin behar, testuingurua ikusi gabe pertsonek ere zalantza handiak dituztelako hitzak zuzentzeko. Testuingurua kontuan hartu gabe hobekuntza batzuk oraindik egin badaitezke ere —adib. hitzen maiztasunak kontuan hartzea, teklatuaren arabera zein aztertutako erroreen arabera aldaketei pisuak esleitzea— mugatik nahikoa gertu gaude eta testuingurua kontuan hartzea ezinbestekoa da emaitza horiek nabarmen hobetzeko, batez ere OCR eta hizketaren prozesaketarako erabili nahi bada zuzenketarako tresna hau. Gainera, testuingurua kontuan hartzen bada, benetako hitzaren erroreak detektatu eta hitz-mugaren gaineko erroreak zuzendu daitezke, bigarren belaunaldiko zuzentzailea sortuz. Testuingurua kontuan hartzeko lau bide bereizten dira nagusiki: corpus-en gaineko estatistikak, sintaxia, semantika —hitzen arteko erlazioak lortuz—, eta soluzio partikularrak. Metodo hauek konbina daitezke baina lehen hiruetarako oinarrizko lanen garapena behar da aurretik: analizatzaile sintaktiko osoa edo partziala, esanahien bilketa eta egituraketa datu-basean eta haien arteko distantzia kalkulatzeko metodoa semantikarako, eta corpusak ustiatzeko tresnak. Oinarrizko tresna hauetan ari gara lanean epe erdian zuzenketan aplikatzeko asmoz.

VII.2.3 EUSLEM.

Garatzen ari garen euskararako oinarrizko lematizatzaile/etiketatzailea da EUSLEM. Hitzaren esparrua gainditzen duenez, lan honetatik kanpo utzi dugu baina aurreratu samarra dago, aurkeztutako analisi morfologikoan oinarriturik baitago. Bere diseinurako aztertutako bibliografia azaltzen da eranskinetan, hemen azalduko dena bertako ideietan oinarritzen da eta. Tresna hau oinarrizko lanabesa izango da beste aplikazioetarako, adibidez analisi sintaktikoa, corpus-en ustiapena, dokumentuen datu-baseen indexazioa, lexikografia etab.

Hasierako lan garrantzitsu bezain korapilatsua etiketen definizioa eta analisi morfologikoarekiko egokitzapena izan da. Maila desberdinetan eratutako etiketa-sistema bat diseinatu da, oraindik ebaluatu gabe dagoena. Euskaran gertatzen den elipsiaren arazoa ere aztertu dugu bere tratamendurako proposamen bat eskainiz (Aduriz *et al*., 95).

Diseinatutako tresna hau (Aldezabal *et al*., 94) honako elementuek osatzen dute funtsean:

- Hitzak, puntuazio-karaktereak, zenbakiak etab. identifikatzen dituen aurreprozesadorea. Analisi morfologikorako egindakoan zenbait aldaketa eginez lortzen da.

- Analizatzaile morfologikoa, hitzei dagozkien lema eta etiketa posibleak zehazteko. Analizatzaile estandarra erabiltzen da lan horretan, ondoren analisiaren arabera etiketak egokitzeko prozedura burutuz.

- Analizatzaile morfologikoak ezagutzen ez dituen hitzen lema eta etiketa hipotetikoak lortzen dituen hitz ezezagunen etiketatzailea edo *guesser*-a. Horretarako, egindako aldaeren analisia eta lexikorik gabeko analisia erabili ondoren, laugarren kapituluan deskribatutako desanbiguazio lokala eta aurreko puntuan aipatutako etiketen egokitzapena burutzen da.

- Hitz anitzeko terminoen identifikazioa, horien artean lokuzioak, hitz-elkarketa eta bestelako kasu asko sartzen direlarik. Flexio handiko hizkuntzetan tratamendu honek ere berezitasunak ditu. Momentu honetan hitz anitzeko terminoekin datu-basea ari da osatzen, terminoei lau ezaugarri egokituz: (1) segurua/anbiguoa, lehen kasuan hitz banatuen analisiak baztertu ahal izateko; (2) finkoa/deklinagarria, finkoetan terminoaren identifikazioa hitzen arabera egingo den bitartean deklinagarrietan lemak identifikatu behar dira; (3) jarraian/ez-jarraian, bigarren kasuan terminoaren osagaiak sakabanatuak egon daitezke eta; (4) ordenan/ez-ordenan, azken hauen identifikazioa korapilatsuago baita.

Ezaugarrietatik ondorioztatzen denez gero, hitz anitzeko terminoen identifikazio-prozesua konplexua da oso.

- Testuinguruan oinarritutako desanbiguazioa, interpretazio morfologiko anitz duten hitz/terminoei lema/etiketa bakarra esleitzeko asmoz. Horretarako metodo desberdinak erabil daitezke: estokastikoak (Garside *et al*., 87), (De Rose, 88), (Cutting *et al*., 92), (Elworthy, 93), linguistikoak (Karlsson *et al*., 92), (Voutilainen, 94) (Tapanainen, 94) edo bion konbinaketa direnak (Leech *et al*., 94). Metodo estokastikoen bidez emaitza hobeak lortzen zirela iritzi zabalduaren aurrean, bestelako iritziak ugaldu egin dira azken urteotan (Brill, 92), (Chanod & Tapanainen, 95). Lanean ari gara bi bideak jorratzeko, tresna linguistikoaren garapenean sintaxirako ere erabiltzen den Murriztapen-Gramatika erabiliz (Karlsson, 95).

Proiektu honen oinarrian EEBS —Egungo Euskararen Bilketa Sistematikoa— (Urkia & Sagarna, 91) dago, eta bertatik lortu ditugu EUSLEMen erabiltzen diren corpus-ak. Momentu honetan testu-zati bat ari gara desanbiguatzen eskuz, geroago ezagutza-iturri gisa zein emaitzen ebaluaziorako erabiliko dena.

VII.2.4 Beste aplikazioak.

Ikusi den bezala, bi mailatako eredua oso aplikagarria da fonologian ere; beraz, hizketaren tratamenduan aplikazio zuzena izan dezake. Hizketaren tratamenduaren inguruan ari den beste ikertalde batekin elkarlanean aztertzen ari gara gure lanaren integrazioa hizketaren sorkuntzarako sistema batean.

Beste aldetik hemen azaldutako tresnak oinarrizkoak dira beste askoren eraikuntzan (ikus §l.9 atala). Analisi sintaktikoa eta itzulpenerako tresna lagungarriak daude taldeko helburu hurbilen artean.