Hizkuntzaren tratamendu automatikoa: helburuak eta abiaburuak

Iñaki Alegria, Xabier Artola eta Kepa Sarasola[1]

Sarrera: lengoaia naturalaren prozesamenduaren helburuak

Inprentaren sorkuntzak hizkuntzaren tratamendua eta zabalkuntza irauli zuen moduan, mende honetan sortu den ordenadoreak horren pareko iraultza ekarri du. Testu–prozesaketarako baliabide berria den ordenadoreak erraztasun handiak eskaintzen ditu gaur egun testuak kopiatu, osatu eta zuzentzeko, eta baita mila formatu edo itxura desberdinetan aurkeztu ahal izateko ere. Baina testu–edizioko tresna horiek baino askoz ere laguntza hobeak dira merkatuan eta are laguntza bereziagoak bilatzen dira ikertokietan. Ordenadorearen bitartez hizkuntzaren tratamendua egiten duten aplikazioak eta programak gero eta ugariago dira, ordenadoreekiko komunikazioa egunero erabiltzen dugun hizkuntzaren bitartez egin ahal izatea gero eta normalago izango baita. Beste alde batetik, gizarte eleanitzak hizkuntza diferenteen artean egin behar izaten dituen joan-etorriak leuntzeko ere aparteko lagun izango dugu ordenadorea. Gainera, telekomunikazioetan gertatutako aurrerapen izugarriak eragin duen Internet fenomenoak areagotu egin du hizkuntzaren tratamendu automatikoaren beharra; zeren eta, nahiz eta sarearen bidez informazio kopuru izugarria lortu ahal izan, ez baita erraza bilatzen dugun informazioa aurkitzea, eta informazioa ondo selekzionatzeko tratamendu linguistikoa lagungarria baino ezinbestekoa da.

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari *Lengoaia Naturalaren Prozesamendua* (LNP) esaten diogu, nahiz eta batzuetan, hizkuntzalaritzako ikuspuntua garrantzitsua denean batez ere, *Linguistika Konputazionala* ere esan. Hizkuntzaren industria oso bat sortzen ari da, ordenadoreaz baliatuz hizkuntza tratatzea helburu duena. Hizkuntzaren teknologiaz hitz egiten da dagoeneko. Teknologia horren oinarrian ikerkuntza dago, hizkuntzaren tratamendu automatikoaren arloko ikerkuntza, alegia. Horiek guztiak dira artikulu honen aztergaiak; hasieratik argi utzi behar dugu, ordea, ez garela ariko euskarazko softwareaz, hau da, euskaraz erabil daitezkeen ordenadore-programei buruz[2], ezta orokorrean euskarak informatikaren munduan egun duen leku eskasaz ere[3].

Dena dela, azken urteotako lorpen hauek muga handiak dituzte. Bost urteko edozein ume hitz egiten eta ulertzen ondo moldatzen denez, hizkuntza erabiltzea lan erraza dela pentsatzen dugu, baina hori ez da horrela. Lengoaia sortzea eta ulertzea oso prozesu konplexuak dira eta gaur egungo ordenadoreak urrun ikusten ditu giza adimenaren ahalmen linguistiko orokorrak. Baina horrek ez du esan nahi hizkuntza lantzeko tresna automatikoak utopia direnik, hizkuntzaren oinarrizko ezagutza minimo batekin laguntza interesgarriak eskain daitezke eta. Testu guztiak ez dira zailtasun maila berekoak: ez da berdin ulertzea "Obabakoak", telebistako eguraldi-iragarpena, edo egunkariko zinema-karteldegia. Hirurak euskarazko hizketa izan arren, bakoitzean erabiltzen diren hitzak, esateko erak eta esanahiak maila desberdinekoak dira erabat. Euskaldun alfabetatu batek ederto ulertuko luke hiru kasuotan, baina ordenadore bidez ulertu nahi duen programa batek zailtasun handiagoak izango ditu, lehenengo bi kasuetan behintzat. Emaitza probetxugarriak lortzeko, ordenadorearen lana aztergai espezifiko eta mugatu batean kokatu behar da. Egun aurretiko hitzordua ematen duten sistema gehienek zenbakiak eta astegunen izenak besterik ez dituzte ulertzen, baina hala ere ekonomikoki oso interesgarriak diren aplikazioak antolatu dira horrekin. Etorkizunean, aplikazio mugatuko sistemak bilduz, lor litezke ahalmen handiagoko sistema berriak, baina egun ibili dabiltzan aplikazioek helburu espezifikoak dituzte.

LNParen barruan azaltzen diren sistemak eta produktuak hobeto aurkeztearren komeni da bereiztea zein diren aplikagarritasun-maila desberdinak. Lau multzo nagusi egingo ditugu: lehenengoan, linguistikaz edo informatikaz gutxi dakien erabiltzaile arruntarentzat salgai diren **aplikazioak** sartuko ditugu; bigarrenean LNPko ekoizleentzako bakarrik interesgarriak diren **tresnak**, produktu berriak garatzeko baliagarriak; hirugarrenean aztertuko ditugu edozein aplikazio edo tresna garatzeko behar-beharrezkoak izango diren **oinarriak**; eta, azkenik, laugarrenean oraindik aplikazio mailara ailegatu ez diren **ikerketa-gaiak** sartuko ditugu.

Artikulu honetan LNParen egungo egoera bi ikuspuntutatik aztertu dugu. Hasieran merkatuan aurki daitezkeen aplikazioak aurkeztuko ditugu. Horietako gehienak ingelesaren munduan mugitzen dira eta bigarren maila batean frantsesa, alemaniera eta espainiera bezalako hizkuntzak daude; euskararako aplikazio gutxi ditugu oraindik. Artikuluaren bigarren zatian helburu edo aplikazio horietara noizbait heltzeko beharko genituzkeen abiaburuak deskribatuko dira, epe erdi edo luzean euskararen tratamendu automatiko zabala posible egingo duten tresnak eta oinarriak, batik bat Donostiako Informatika Fakultateko IXA taldean azken hamar urte honetan sortu direnak[4].

Aplikazioak

LNParen 40 urteko historian gorabehera handiak izan dira. Helburu liluragarriak lortzear zeudela uste zen une euforikoen ostean, belarriak jaitsi eta helburu apal baina euskuragarriagoetara mugatzeko une pragmatikoak etorri dira birritan edo. Erabateko itzulpen automatikoa ordenadoreen eskutik etorriko zela aurreikusi zuten 1954an Georgetown-eko Unibertsitate inguruan. Alabaina, 1966an itzulpen automatikorako diru-iturri ofizial guztiak itxi egin ziren, ALPAC txosten ezagunak horrela gomendatu eta gero. Aurrerago, 1980 inguruan, adimen artifizialeko teknika berrien eskutik ordenadoreak geure hizkuntzaz —lengoaia naturalean— programatuko genituela agintzen zitzaigun.

---

1 Donostiako Informatika Fakultateko IXA taldekoak.

2 Ikus horretarako: Sarasola K.. Euskarazko softwarearen katalogoa. *Elhuyar, Zientzia eta Teknika*, 117. zk. 60-62 orr. 1997.

3 Ikus horretarako: Artola X.. Informatika eta euskara: gaur egungo arazoak eta aurrera jotzeko bideak. *Uztaro*, 20. zk., 77-92 orr. 1997.

4 Artikulu osoan zehar aipatuko ditugunak talde horren esperientziari eta egungo jardunari lotutakoak izango dira gehienbat. Informazio gehiago nahi duenak jo beza amarauneko helbide honetara: http://www.ji.si.ehu.es/Groups/IXA.

Gaur egun ahaztuta daude horrelako ametsak. Dena dela euforia eta pragmatismoko ziklo horiek bi motako emaitzak utzi dituzte: alde batetik, hobeto baloratu eta ezagutzen dugu hizkuntzaren egitura eta erabilera, eta aitortu behar izan dugu ez direla hasieran uste bezain sinpleak; bestetik, helburu utopiko horiek lortzeko asmotan eraiki diren tresnekin helburu apalagoa duten baina komertzialki bideragarriak diren produktu asko merkaturatu dira. Horrelako zenbait aplikazio arrakastatsu aipatuko ditugu ondoren.

Testuen edizioa eta gestioa

Ordenadorea kalkulu ugari eta konplexuak egiten dituen makina dela esan daiteke. Programak idazteko erabiliko zirela uste genuen, ez ordea prosa idazteko. Baina zer dela-eta zabaldu zaie atea etxe eta bulego gehienetan? Etxeko jaun-andreek programatzen ikasi dutelako? Ez, ezpada ordenadorea testuak idazteko idazmakina azkarra delako, edo Interneten bidez hainbat informazio eskuratzeko tresna ona delako. Edonork ulertu eta idazten du prosa baina gutxik programak. Testu-ediziorako eta testu-gestiorako tresnen garapena guztiz lotuta dago azken urteotan konputagailuen erabilera masiboarekin. Honez gero testu-edizioa ez da tekleatze hutsa, edo testu baten bertsio berri bat lortzea aurreko bat kopiatu, moldatu edo osatuz; ezta hamaika formatu edo itxura desberdinetan aurkeztu ahal izatea bakarrik ere. Egun badira testu-egileari eskaintzen zaizkion laguntza bereziak. Ikus ditzagun orain zein diren garrantzitsuenak.

Ortografia-zuzentzaileek bete dituzte urte batzuk merkatuan, eta gaur egun hizkuntza askotarako aurki daitezke. Zuzentzaile hauek testuko hitz bakoitza aztertzen dute ea hitz posiblea den egiaztatzeko, baina gehienetan testuingurua kontuan hartu gabe. Euskara bezalako hizkuntzen kasuan hitzak kasu desberdinetan deklinatuta agertzen direnez askoz lan konplexuagoa da hitz zuzenak eta okerrak bereiztea, analisi morfologikoa egin behar baita. Hala ere, 1994tik dago dendetan XUXEN euskararako egiaztatatzaile/zuzentzaile ortografikoa. Zenbait hizkuntzarentzat *idazkera- eta sintaxi-zuzentzaileak* ere merkaturatu dira; hauek testuingurua kontuan hartzen dute eta, adibidez, “nik joan naiz” esaldia prozesatuz gero, ortografia-zuzentzaileak ez luke errorerik salatuko, hiru hitzok isolatuta posibleak baitira, baina sintaxi-zuzentzaileak testuinguru horretan “nik” hitza gaizki dagoela salatuko luke eta “ni” izan beharko lukeela proposatu. Nahiz eta errore guztiak harrapatu ez, laguntza ederra eskaintzen diote eskutitzak edo bestelako txostenak idazten dituenari. *Laguntza lexikaletan* edozein hitzen sinonimo edo antonimoak lor daitezke testu-prozesaketako programatik atera gabe, baita taxonomikoki konkretuagoak edo orokorragoak diren antzeko hitzak ere (adibidez: *intsektu* hitzetik orokorragoa den *animalia* edo konkretuagoak diren *inurri*, *euli*, ...), *thesaurusa* kontsultatuz.

Testu eleanitzak lantzeko adibide gisa Siemens-en Eurolang Optimizer, IBM-ren TranslationManager/2 eta Trados-en Translation Workbench programak aipatu behar dira. Prozesadore zabalduenetan integratzen diren programa hauek glosategi, hiztegi eta itzulpenen berrerabilpenerako laguntzak eskaintzen dituzte. Itzultzaileek arazo franko izaten dute terminologiarekin testuko gaiarekin ohituta ez daudenean. Horrelakoetan terminologi akatsak (hitz ezegokia ematea edo termino bera orrialde berean desberdin itzultzea) askoz itsusiagoak dira errore ortografikoak baino. Glosategi, hiztegi orokor edota hiztegi berezituen on-line moduko erabilerak errore horiek urritzeko baliabide eraginkorrak dira. Itzulpenen berrerabilpenerako laguntzek itzultzaileari lana errazten diote testuen bertsio berriak egiterakoan, aldatu dena itzuli beharko baitu eta ez testu osoa. Automatikoki sortzen dute bertsio itzuli berria, osatu behar diren hutsuneak bereizita agertzen direla. Siemens-en EuroLang Optimizer programak Metal itzulpen-sistemarekin batera ere lan egin dezake.

Testu-masa handiak tratatzeko edo gestionatzeko aplikazio nagusiak lau dira: kontzeptu-bilaketa, kategorizazioa, informazio-erauzketa eta testu-sorkuntza automatikoa.

On-line moduko *kontzeptu-bilatzaileen* inguruan mila milioi dolarreko industria antolatuta zegoen 1994an Estatu Batuetan. Orain arte erabilitako teknikak oso sinpleak ziren (hitz gakoen konbinazio boolear hutsa); gaur egun lematizazioa, perpausen bukaeren detekzioa, akronimoen zabaltzea eta kalkulu estatistikoak ia sistema guztietan egiten dira; Clarit, Conquest eta ConText (Oracle) produktuen egileek, etorkizuneko bidea erakutsiz, beren ekarpenaren iturria LNPko teknika sofistikatuenetan kokatzen dute. Euskararako ere bada berriki Ametzagaina taldeak kaleratutako Kapsula softwarea, euskarazko dokumentu-baseen gestiora zuzendua.

Kategorizazio-sistemak oso baliagarriak dira makina bat dokumentu (adibidez: telefonoetako matxura-parteak, albisteak, hildako militarren parteak, marketineko datuak, ...) kategoria multzo txiki baten arabera sailkatu behar izanez gero. Esate baterako, Carnegie Group enpresaren Construe sistemak Reuter informazio-agentziaren artikuluak automatikoki sailkatzen ditu, eta urtez urte agentziari 750.000 dolarreko aurrezpena ekarri dio 1990 urteaz geroztik. ATT telefono-konpaniak daukan sistemak matxura-parteak automatikoki bideratzen ditu konponketaz arduratu beharko den bulegoraino. Zenbait aplikaziotan, nahiz eta dokumentu guzti-guztiak ezin sailkatu, nahikoa da dokumentuen gaineko estatistika orokorrak lortzea. Esate baterako, matxura-parteen estatistika orokorrak produkzio-kate baten osagai ahulenak zein diren jakiteko balio dezakete. Kategorizazio-sistema batzuk haratago doaz eta saiatzen dira ezagutzen zein diren "elementu agertu berriak", hau da, nahiz eta sarritan azaldu, behin eta berriz sailkatu gabe geratzen direnak beraientzako kategoriarik definitu ez delako. Dokumentu-kategorizazioko teknikak bi motakoak dira: estatistikoak eta ezagutza-ingeniaritzakoak. Ezagutzaren ingeniaritzako teknikak baliatzen dituztenak zehatzagoak dira, kalitate hobea lortzen dute, baina oso garestiak dira, eta ez dira errentagarriak dokumentu kopurua benetan erraldoia ez bada (horrela aitortzen zuten Carnegie Group enpresakoek). Dena dela, tresna estatistikoak erabili behar badira ere, aldez aurretik testu luzeak etiketatu behar dira sistemak horietatik ikas dezan.

Informazio-erauzketako sistemek lengoaia naturalez idatziriko testuetatik datu-base egituratu bat osatzen dute. Azken helburua albiste multzo handi batetik abiatuz fitxa konkretuak betetzea litzateke nork-nori-zer egin dion jakiteko. Dena dela, helburu apalagoak baldin badituzte ere, badira produktu asko merkatuan etekin handiak ateratzen dituztenak. Gehienek dokumentuak aztertzen dituzte enpresa, pertsona, hitzordu-data, telefono edo zerbitzuen erreferentzia hutsen bila. Adibidez Westlaw eta Lexis-Nexis-ek enpresen aipamenak bilatzeko programak saltzen dituzte, enpresaren aipamena modu askotara azal daitekeelarik: esate baterako, IBM, I.B.M., International Bussiness Machine, e.a.

Testu-sorkuntza automatikoa informazio-erauzketaren kontrakoa da. Kasu honetan ordenadore barruan dauden datu konplexuetatik abiatuz (inprimakiak, datu kodetuak edo zenbakizko formatuan dauden informazioak...), datu horien edukia azalduko zaio erabiltzaileari bere hizkuntzan. PLANDoc sistemak telefono-enpresa batentzat honen telefono-sarerako hobekuntzak asmatzen ditu, baina erabiltzaileak ez du ikusten programaren emaitzaren kode ulertezina, berori azaltzen duen ingelesezko testua baizik. Forecast Generator sistemak (geroago aipatuko den METEO sistema ospetsuaren ondorengotzat hartua izan dena) ingeles edo frantsesezko testuak idazten ditu ordenadore batek kalkulatzen dituen eguraldi-iragarpen kodetuetatik abiatuz.

Itzulpen Automatikoa

Produktu ugari dago merkatuan salgai testu-itzulpenean laguntza emateko, baina euskara tratatzen duen sistemarik ez dago. Itzulpen perfektua egiten duen sistemarik ez dago inon, eta sistema batek berak ere ez ditu testu literarioak itzultzen. Guztiek itzulpen teknikoa dute erabileremu, testu teknikoetan hizkuntzen arteko hitzen eta esaldien korrespondentzian anbiguotasun gutxiago aurkitzen baita. Sistemaren batek tratatzen dituen testuak edozein motakoak badira, ziur emaitzaren kalitatea kaxkarra dela. Hala ere, gero azalduko dugun bezala, zenbait kasutan sistema horiek lagungarri izango dira.

Itzulpenaren automatizazioa ez da ia inoiz erabatekoa, eta automatizazio-mailaren arabera ondoko sailkapena egiten da: 1) Erabateko itzulpen automatikoa: errealitatea baino ametsa da gaur egun, non eta helburua ez den edukiaren ideia orokorra ateratzea. 2) Giza laguntzaz buruturiko ordenadore bidezko itzulpena: lanaren gidaria makina da, baina fase desberdinetan laguntzak eska ditzake; hitz baten adiera zuzena hautatzeko edo esaldi baten analisia nondik hasi behar den erakusteko adibidez. 3) Ordenadorez lagunduriko giza itzulpena: lanaren gidaria pertsona da, baina ordenadoreaz baliatzen da hiztegi berezituan kontsultak egiteko, testuaren formatua txukuntzeko eta zailtasunik gabeko testu-zatiak itzultzeko. Agian itzulpenaren zati handi bat ordenadoreak egingo du ia laguntzarik gabe, baina testua egokitzeko aurreprozesaketa edota emaitza zuzentzeko postedizioa ohikoak izaten dira. 4) Datu–banku terminologikoak: hiztegi berezituak erabiltzeko aukera hutsa eskaintzen duten laguntza-sistemak.

Testu itzulien erabilera nagusi bi bereizten dira: edukiaren ideia orokorra ateratzen dutenak, eta zabalkuntza handiko informazio zehatzak itzultzen dituztenak. Lehenengoaren adibide tipikoa "internautarena" dugu: hizkuntza arrotz batean itxura oneko web-orri bat aurkitu du, guztia zehatz-mehatz itzultzea denboraz edo diruz oso garestia litzateke, eta gainera, ia ziur oso-osorik ez litzaiokeela interesatuko gero. Guztiz zuzena ez den baina merkea den itzulpena ganbegiratuz jakin ahal izango du benetan interesatzen zaion partea zein den, eta gero zati horren itzulpen zehatza lortu. Oro har, denbora eta dirua irabaziko du honela. Beste aldetik, zabalkuntza handiko informazio zehatzen adibide gisa, etxetresna elektroniko baten erabilpenerako azalpenak ditugu. Testu horien zehaztasun-ulergarritasunak produktuaren arrakastarako giltza izango dira. Beraz, itzulpenak kalitate handikoa izan beharko duenez nahitaezko lana izango da giza-itzultzaile baten zuzenketa edota postedizioa. Postedizioa ekiditeko asmoz zenbait sistematan saiatu dira jatorrizko testuak mugatzen, erraz itzuli ahal izango denera mugatuz. Horrelakoetan analizatzaile bereziak definitzen dira jatorrizko testuetan lengoaia kontrolatutik ateratzen diren hitzak edo esaldiak salatzeko.

Montrealeko TAUM taldeak egindako METEO sistema da emaitzarik arrakastatsuena lortu duena. Parte meteorologikoak itzultzen ditu 1977tik hona, ingelesetik frantzesera, eta itzulpenaren %80 erabat zuzena da. Egunero oso antzekoak ziren itzulpen aspergarri hauek egiteko itzultzaileak bilatzea zaila zen, nahiz eta soldata ederrak eskaini. Urte hartatik hona lana egunero burutzen da METEOren laguntzaz. Hamaika saio egin da geroztik sistema honen diseinua beste gai batzuetara zabaltzeko, baina ezin izan da horren biribila den beste gai bat aurkitu. TAUM taldeak berak hegazkinetarako eskuliburuak itzultzeko saioak egin zituen, baina hasierako emaitza itxaropentsuek piztutako ametsak laster itzali ziren.

SYSTRAN Institutua 1970. urteaz geroztik itzulpen automatikorako tresnen saltzaile nagusia izan da. NASA, Europako Batasuna, General Motors eta Xerox dira bere bezerorik ezagunenak. Europako Ekonomi Elkarteak egokitzapen neketsua behar izan zuen —100.000 hitzeko hiztegia definitu behar izan bait zuen— frantses/ingeles itzulpena ahalbideratzeko. Baliagarritasun-mailaren berri emateko edo, aski izan daiteke aipatzea nola orain dela hamar urte 20 itzultzailek erabiltzen zuten sistema hau Luxemburg–en, hilabetean milaren bat orrialde ingeles/frantses, frantses/ingeles eta ingeles/italiera bikoteetarako itzultzen zutelarik. Kanadako General Motors–ek eskuliburuak itzultzen zituen ingelesetik frantsesera: 130.000 hitzeko hiztegia definitu ondoren, itzultzaileen lana lehen baino 3 edo 4 aldiz arinagoa zen, eguneko 1000 hitzetara helduz. SYSTRANen oinarri informatikoa, baina, guztiz atzeratuta dago, 1960ko hamarkadako teknologia erabiltzen baitu; hala ere, Systran itzulpen-sistema hoberenen eta erabilienen artean dago oraindik.

Dozenaka produktu dago ordenadore pertsonaletan itzulpenak egiteko hizkuntza bikote desberdinen artean. Adibidez, ingelesa-espainiera itzulpenak egiteko badira sistema komertzialak: Spanish Assistant, Dos amigos, Context, Translate, Globalink. Guztietan postedizioa beharrezkoa da, eta nolabaiteko elkarrekintza dago beti giza-itzulzailea eta programaren artean hitzen adiera zuzena hautatzerakoan eta. Bizkaiko GEINSA enpresa ere ingelesa-espainiera bikoterako sistema bat garatzen ari da, eta euskara ere lantzen dute maila apalago batean.

Ordenadoreen erabilera LNaren bidez

Aplikazio–mota honetako sistemek, ordenadore eta gizakiaren arteko komunikazioa errazten dute, erabiltzaileak bere hizkuntzaz lan egiteko aukera du eta. Horrelako sistemak inplementatzen zailak dira; galdera eta erantzunez osatutako elkarrizketa ulertu ahal izateko, partaideen planak eta helburuak aztertzeko tresnak behar baitira. Hiztun bakoitzak momentu bakoitzean zer dakien eta zer nahi duen asmatu behar da eta, gainera, ezagumendu horiek etengabe eguneratzen ibili behar da elkarrizketa aurrera joan ahala. Helburu orokorrekorik ez da luzaroan salgai egongo, baina badira dagoeneko aplikazio konkretuei lotuta dauden batzuk.

Datu-baseetarako galdeketa-sistema ugari dago, batez ere ingelesez. Datu–base konplexuei galderak egin ahal izateko lengoaia berezi bat ezagutu beharrak datu–baseen erabiltzaile potentzialen

kopurua murrizten duenez, galderak lengoaia naturalez egin ahal izatea oso interesgarria da bezero berriak harrapatzeko. Horrela produktura lehenengo aldiz hurbiltzen den erabiltzaile potentzial berriak oztopo gutxiago aurkituko du martxan ikusteko. Behin produktuaren funtzionamendua ezagututa motibazioa etorriko zaio probetxu handiagoa ateratzeko eta, hortaz, kontsulta-lengoaia berezia ikasteko. Izan ere, ordurako amua janda dauka eta programaren munduan sartuta dago.

Symantec–en "Question & Answer (Q&A)" sistemak arrakasta ederra izan du 1986 urteaz gero. Sistema hau analisi sintaktikoa bigarren mailarako lagatzen duten "gramatika semantikoetan" oinarritzen da. Galderak oso zailak ez badira, emaitza harrigarriak lortzen ditu.

Alde ikaragarria dago ordenadore erraldoietarako eta mikroetarako egindako interfazeen artean; bai prezioz (milioiak eta hamarnaka mila pezeta inguru, hurrenez hurren) eta bai ahalmenez. Hizkuntzaren tratamendua askoz zabalago eta sakonagoa izateaz gain, sistema handietan erabiltzailearentzat laguntza eta erraztasuna handiagoa da. Erabiltzaile anitzi erantzun dakioke eta datu–base ahaltsuagoak atzitzeko aukera eskaintzen dute. Mikroetan kokatutako interfazeak guztiz desberdinak dira. Merkatuan 100 baino gehiago dira ingelesezko produktuak.

Ikerkuntzaren mundutik datozen ahalegin berrietan LNPko teknikak eta multimediakoak biltzeko saioak egiten dira, edo menuen bidez erakusten dira egin daitezkeen esaldien egitura eta kontzeptuak. Adibidez Texas Instruments enpresaren "Natural Link" paketean, erabiltzaileak ezin du galdetu edonola, berari aurkezten zaion menu moduko pantaila batean hitzak edo esaldi-zatiak hautatzen ditu nahi duen galdera osatzeko. Horrela esanda, menu hutsa dela dirudi, baina bere atzetik dagoen hizkuntz analizatzailea antzeko beste sistemen mailakoa da. Pakete honen ezaugarririk onena gardentasuna da: erabiltzaileak ondo daki zeintzuk esaldi ulertuko diren eta zeintzuk ez.

Datu–baseei buruz azaldu dugun hori guztia berdin esan daiteke gainontzeko aplikazio-programez ere. Gehienak adimen artifizialeko sistemetan integratuta daude. Baina bestelakoetan ere badira eta, adibidez, zenbait kontzeptu-bilatzailetan galderak lengoaia naturalez (ingelesez kasu guztietan) egin daitezke.

Ahozko hizkuntzaren tratamendua

Ahozko hitzak edo esaldiak ulertzea zaila da, hizkuntza idatzia ulertzeko arazoei ahozko hizkuntzaren problematika eransten zaiolako: hitzak ez dira guztiz bereizten hitz egiterakoan, esaldien hasiera eta bukaera erdikoa baino intentsitate txikiagoz ematen dira, eta, gainera, seinale fisikoen zaratak ohiko oztopoak izaten dira.

Sistema gehienek oso hitz gutxi ezagutzen dute, eta horien artean beti daude zenbakiak. Horrela erabiltzaileak zenbaki bat ahoskatuz aukera desberdinen artean hautatuko du behin eta berriz menu desberdinetan zer (edo zer eskatu) nahi duen ondo zehaztu arte. Merkatu handia zabaldu da horrelako sistemak telefono bidezko zerbitzuetan integratzeko: aurretiko hitzordua, produktu-eskaerak, e.a. Beste alde batetik, hizketaren ezagutzarik gabe, gero eta arruntago bihurtzen ari zaigu makinen ahots sintetizatuak entzutea gasolindegietan edo tabako-edariak saltzen dituzten makinetan.

Natural Vox enpresa arabarrak aurretiko hitzordua —medikuarenean eta errenta-aitorpena egiterakoan— automatikoki lortzeko sistema telefonikoak ezarri ditu azken urteetan, eta arrakasta handia izan du.

Ahozko hizkuntzaren tratamenduko teknikak antzeko beste aplikazioetan ere erabiltzen dira: eskuz idatzitako testuak ezagutzeko edota testu mekanografiatuen bertsio elektronikoa lortzen duten OCRetan (Optical Character Recognizer, karaktere-ezagutzaile optikoak).

Abiaburuak

Batez ere ingeleserako merkatuan aurki daitezkeen aplikazioak ikusi ondoren, artikuluaren bigarren zatian helburu horretara noizbait helduko bagara martxan jarri beharko genituzkeen abiaburuak deskribatuko ditugu, beti ere, IXA taldean markatutako estrategiari jarraituz eta bere mugekin, noski, eta azken hamar urte honetan sortu ditugun tresna eta oinarrietatik abiatuz. Abiaburuen artean aipatzekoa litzateke, jakina, arloko ikerkuntza. Hala ere, artikulu honetan aplikazio eta tresnen oinarri direnak azalduko ditugu batik bat, eta egun lantzen ari garen bestelako ikerketa-gaiak —teorikoagoak edo— aipatu baino ez ditugu egingo azkeneko atalean.

Tresnak

Atal honetan hizkuntzaren tratamendurako aplikazio-ekoizleentzat edo arloko ikertzaileentzat interesgarriak diren tresna batzuk ikusiko ditugu. Tresna horiek ez daude diseinaturik, oro har, erabiltzaile arruntarentzat.

Analizatzaile morfologikoa

Ingelesaren flexio-morfologia sinplearen eraginaz ordenadorez egindako analisi/sintesi morfologikoari kasu handiegia ez zitzaion egiten, eta askotan aplikazioak ibiltzen ziren hizkuntzaren forma guztiak zituen hiztegi batekin. Hau pentsaezina da euskara bezalako hizkuntza flexionatu eta eranskariaren kasuan, erro batetik sor daitezkeen hitz flexionatu posibleak asko eta asko baitira. Eta hori ez bakarrik euskararako, beste hizkuntza askotan (suomiera, turkiera, etab.) arazo bera baitzegoen. LNPko teknikak ingelesetik beste hizkuntzetara hedatzean eman zitzaion morfologiari daukan garrantzia.

Morfologia automatizatzeko orduan hizkuntzaren hiru aspektu deskribatu behar dira zehatz: lexikoa, hitzaren osaketa eta aldaketa fonologikoak. Lexikoa funtsezko elementua da eta lexiko oso eta orekatu bat eraikitzea lan izugarria da aurretik datu-base lexikal bat ez badago, lexikoko sarrera konbentzionalez gain morfema ez-independenteak ere behar dira eta gainera osagai guztiei dagokien informazio morfologikoa. Hitzaren osaketa (morfotaktika edo hitzaren gramatika deitu ohi da, hizkuntzak onartzen duen flexio-bideari jarraiki sarrera bakoitzaren ondoren etor daitezkeen elementuen deskribapen formala) deskribatu behar da sistema informatikoak jakin dezan nola lot daitezkeen erroak, aurrizkiak eta atzizkiak. Horrela lortuko dugu ez ezagutzea —eta ez sortzea—

etxeago baina bai *handiago*, eta ez *baitelako* baina bai *baita* eta *delako*. Azkenik, elementuak biltzean sortzen diren aldaketak (aldaketa fonologikoak) deskribatu behar dira; horrela azalduko da adibidez, aurretik azaldutako adibideari jarraituz, *bait* eta *da* biltzean ez dela *baitda* gertatzen, *baita* baizik.

Hiru atal horien bidez egiten dira gaur egun hizkuntzen deskribapenak morfologia automatikoari begira. Kasu gehienetan flexio-morfologia hartzen da kontuan baina ez eratorpena eta elkarketa, azken horiek ez baitira erregularrak.

Deskribapen horiek egiteko modua erabiliko den programaren menpe egongo da (programak hizkuntzatik bereizi egiten dira gaur egun). Horrekin lor daiteke analizatzaile/sortzaile morfologiko bat, programa hauek askotan gai izaten baitira hitza emanda analisia lortzeko eta erro batetik abiatuta deklinabide osoa lortzeko. Horixe izan da IXA taldean euskararako egin dugun lehen tresna.

Analizatzaile morfologikoa oinarri bat da hainbat aplikaziotarako. Honako hauek dira garrantzitsuenak:

- Zuzentzaile ortografikoa. Hitza analizatzerik baldin badago hitza zuzena izango da eta bestela abisu bat emango da. Hori eginda dago euskararako.
- Tutore-sistema automatikoak hizkuntza ikasten ari den jendearentzat. Erroreak detektatzeko, ariketak programatzeko etab.erako oso elementu interesgarria da. Honetan ari gara lanean gaur egun.
- OCR dokumentuen irakurketan (eskanerrak erabiltzean) sor daitezkeen erroreak detektatzeko.
- Hizketaren sintesia edo testu-sorkuntza lortzeko sorkuntza morfologikoa funtsezko osagarria da.
- Hizkuntz aplikazio sofistikatuagoetarako —sintaxian oinarritutakoak, itzulpen automatikoa, etab.— lehen urrats gisa.

Lematizatzaile/etiketatzailea

Morfologiatik sintaxira doan bidea oso luzea gertatzen da LNParen munduan, morfologia-sistema osoak eraikitzea posiblea den bitartean, gaur egun ez baita oraindik posible sistema sintaktiko automatiko oso bat garatzea. Are gutxiago ingelesa bezala ikertu ez den hizkuntza baterako. Hori dela-eta tarteko bideak hartu dira eta analizatzaile sintaktiko orokorrak baino sinpleagoak diren tresnak bultzatu dira azken urteetan. Arrakastatsuenak etiketatzaileak izan dira eta, haiekin batera, lematizatzaileak. Etiketatzaileek testuko hitz bakoitzak dituen analisi desberdinen artean zuzena dena aukeratu behar dute; lematizatzaileek, aldiz, lema posibleen artean dagokiona. Adib. *zuen* hitza analizatzean posible da *ukan* aditza lehen aldian izatea edo *zuek* izenordaina genitiboan. Testuinguruaren arabera erabaki behar du lematizatzaileak zein den hitzari dagokion etiketa zuzena (aditza edo izenordaina) edota zein den bere lema (*ukan* edo *zuek*). Beraz, lan hau konplexua da, ez baita posible hitz isolatuak aztertzea, eta sintaxiaren lehen urrats gisa hartzen da tresna hauen lana.

Atal nagusia desanbiguazioa bada ere, beste zereginak ere badaude halako tresna bat garatzean, esate baterako, hitz anitzeko unitate lexikalen identifikazioa (lokuzioak, hitz-elkarketak, pertsona-izen osoak, etab.). Desanbiguatzeko teknikak bi ildo nagusitatik doaz: metodo enpirikoak edo estatistikoak batetik, eta ezagumendu linguistikoan, erregeletan, oinarritutako metodoak bestetik. Gero eta joera gehiago dago bi metodo-motak konbinatzeko. Lortzen diren emaitzak ez dira zeharo fidagarriak baina %95-98 tarteko fidagarritasuna lortzen da. Sistema hauek garatzen ari dira hizkuntza askotarako eta gu euskararena ere egiten ari gara.

Tresna hauek izan duten arrakasta beren aplikazioetan datza, oso aplikazio interesgarri eta aktualak baitituzte lematizatzaile/etiketatzaileek:

- indexazioa: testuak indexatu nahi direnean ez zaigu forma interesatzen, lema eta kategoria baizik. Indexazioa da oinarria gaur egun hain modan dauden datu-base dokumentaletan eta Interneteko bilatzaileetan. Adibidez, testu batean *kalekoak*, *kalera* eta *kalejiratik* agertzen badira, lehen biek azaldu behar dute *kale*az galdetzen dugunean, baina hirugarrenak *kalejira*z egiten dugunean.
- terminologia/lexikografia: automatikoki lemak ondo identifikatzen badira eta dagozkien etiketak egokitzen bazaizkie lan lexikografikoa erruz errazten da, eta testu batetik terminologia automatikoki erauzteak ez dirudi oso lan zaila.

Analizatzaile sintaktikoa

Analizatzaile sintaktikoen zeregina testuetako osagai sintaktikoak ezagutzea da: hitz isolatuez osatu sekuentzietan elkarrekin lotuta dauden egitura sintaktikoak (perpausak, izen-sintagmak, aditz-sintagmak, izen-lagunak, eta abar) ezagutuko dira. Analisiaren oinarria lexikoa eta gramatika izango dira, hizkuntzako hitzen ezaugarri sintaktikoak eta egitura sintaktikoen osaketa posibleak definituko dituztenak.

Prozesu honek anbiguetate handia sortzen du, esaldi bakar baterako analisi posible anitz lor baitaitezke. Kontuan hartu, gainera, hitzen analisi morfologikoa aldez aurretik egin behar dela, eta hitz bakoitzeko analisi morfologiko anitz sortzen direla (euskararen kasuan eta gure datuen arabera, hitzeko 2,7 aukera desberdin batez beste). Anbiguetatea eragozpen handia da erabateko analisi automatikoa lortzeko. Erraza da laborategiko esaldi multzo bat prozesatuko duen analizadorea eraikitzea, baina oso zaila testu libreekin lan egingo duena egitea. Ingeleserako ALVEY sistemaren gramatikak edozein esaldi tratatzen duela diote, baina gero ez da oso erabilgarria beste analisi edo sistemetan aplikatzeko: darabilgun hizkuntzaren anbiguotasuna dela-eta testu arruntetako esaldietan batez beste 100 bat analisi desberdin lortzen baitute.

Formalismo asko dago gramatikak definitzeko, baina gehienak Chomsky-k definitutako Testuingururik Gabeko Gramatiken gainean egindako hedapenak dira. Baterakuntza-gramatiketan erregela bakoitzeko osagaien gainean ekuazio-multzo bat definitzen da osagaien gaineko komunztadura egiaztatzeko eta egitura konposatuak osatzeko. Beste planteamendu berri bat agertu da zenbait sistema berritan: xedea ez da esaldi oso-osoa analizatzea, hori bilatuz gero gehienetan porrota izango baita

emaitza, eta nahikoa da esaldiaren azaleko analisia lortzea, hau da, bereiztea zein diren osagai posibleak eta beren arteko loturak. Emaitza hauek teoria linguistikoen ikuspuntutik ez dirudite oso dotore, baina emaitza horiekin beste hainbat tratamentu informatikori atea zabaltzen zaio. Beste alde batetik, azken analizatzaile horiek konputazionalki askoz azkarragoak dira. Murrizpen Gramatikak dira adibide ezagunena.

Oinarriak

Datu-base lexikala eta morfologiaren deskribapena

Datu-base lexikala da hizkuntzaren lexikoaren biltegi erraldoia. Hiztegi elektroniko moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahi horrek dituen eskakizunak kontuan harturik antolatua. Horrek eskatzen du, noski, lexikoaren antolakuntza gero zertarako erabiliko den kontuan hartuz egitea, eta lexiko-deskribapenaren sistematizazio bat: sarreren kategoria-sistema bateratu eta homogeneoa erabiltzea, kategoria bakoitzeko elementuak behar den bezala deskribatzeko beharrezko diren ezaugarriak zehaztea, etab.

Euskararen kasuan, IXA taldean Xuxen ortografia-zuzentzailearen prestatze-lanari ekin genionean sortu zitzaigun halako lexiko-biltegiaren premia. Gorago esan bezala, baina, zuzentzaile hori oinarrizkoagoa zen analizatzaile morfologikoaren azpiproduktutzat hartzen genuen guk, eta datu-base lexikala ere ez genuen antolatu nahi izan zuzentzaile horretarako hiztegi edo hitz zerrenda soil gisa, etorkizunean euskararen tratamendu automatikoaren arloko beste edozein tresna edo aplikaziotarako oinarri lexikal sendo gisa baizik. Eta horrela sortu zen EDBL, Euskararen Datu-Base Lexikala, harez gero gure lanetarako oinarri lexikala izan dena, etengabe eguneratuz joan dena, eta gaur edo bihar komunitate zabalago bati bere ateak irekiko dizkiona, oinarriak prestatze-bide honetaz beste batzuk ere balia daitezen.

Datu-basea diseinatzerakoan garrantzi handia eman zitzaion, bada, etorkizunean izan ditzakeen hedapenak onartzeko behar bezain malgua izateari eta, bereziki, bertan jasoko zen informazio linguistikoa ahalik eta erarik neutralenean deskribatzeari, hau da, formalismo edo teoria linguistikoetatik ahalik eta erarik independenteenean.

EDBLk gaur egun 70.000 sarrera inguru biltzen ditu, hiru atal nagusitan sailkatuak: hiztegi sarrerak (izenak, adjektiboak, aditzak, e.a.), adizkiak (aditz-forma jokatuak) eta morfema ez-independenteak (atzizki, aurriki, e.a.). Sarrera-kategoria bakoitzeko aldez aurretik definiturik dauden ezaugarri edo atributuak erregistratzen dira, eta kasu guztietan, lehen aipatu bezala, sarrerari dagokion morfologia ere deskribatzen da (informazio morfotaktikoa), horretarako bi mailatako formalismoaz baliatuz[5].

5 Koskenniemi K.. *Two-Level Morphology: A general Computational Model for Word-Form Recognition and Production*. Ph.D. tesia, Helsinkiko Unibertsitatea. Pubs. no. 11. 1983.

EDBL egun datu-baseen kudeaketarako sistema baten pean dago eta halako sistemek ohikoak dituzten erraztasunak eskaintzen dizkio hizkuntzalariari —hizkuntzalariak baitira bere erabiltzaile nagusiak—: interfaze atsegina lanerako, informazioa egunean mantendu eta berorren kontsistentzia bermatzeko erraztasunak, behar den aplikazioetarako informazioa behar bezala iragazteko aukerak, eta abar. Euskararen baturatze-bidean izandako azken gertakariak —Euskaltzaindiaren erabakiak, batik bat— egunean mantentzeko ere ezinbesteko tresna bihurtu da datu-basea, eta etorkizunean nahi genukeen zabalkundea izan dezanean EDBLk bete dezakeen lan inportanteetariko bat izan daiteke azken erabakien berri emango duen tresna izatea.

Hiztegi elektronikoak

Hizkuntzaren datu-base lexikal orokorra oinarri dela, horren inguruan biltzen ahal dira beste zenbait tresna lexikal ere: definizio-hiztegiak, hiztegi terminologiko berezituak, hiztegi elebidunak, eta beste. Horrelakoen garrantzia ere ukatu ezina da, batez ere hizkuntzaren semantika tratagai denean edota itzulpenaren arloko aplikazioak egiterakoan.

Gaur egun, eta hemendik aurrera zer esanik ez, ia ateratzen diren hiztegi guztiak ateratzen dira euskarri elektronikoren batean (CD-ROMean, batez ere), eta horietaz baliatzea ere helburu da hizkuntza baten tratamendu automatikorako oinarri lexikala prestatzeko orduan.

Gurean hor ditugu UZEIren EuskalTerm datu-banku terminologikoa, I. Sarasolaren Euskal Hiztegia, eta Elhuyarrek, Harluxet Fundazioak eta Adorez taldeak, besteak beste, kaleratutako hiztegi-lanak euskarri elektroniko desberdinetan; etorkizunean, eta datu-base lexikal zentral batekin behar bezala loturik, adibidez, hainbat lanetarako oinarri lexikal osagarri bihur litezke lan horiek, nahiz eta hasieran helburu horrekin sortu ez.

IXA taldean, oraintxe, Euskal Hiztegiarekin eta Aulestiaren ingelesa-euskara makinaratutako bertsioarekin ari gara lanean, EDBLri lotuz honek oraindik ez duen osagai semantikoa (definizioak) eta itzulpenezkoa (ingelesa) gehitzeko asmoan.

Gramatika konputazionalak: sintaxiaren deskribapena

Sintaxia ere funtsezkoa dugu hizkuntzaren tratamenduaren arloko edozein lani ekiteko, hizkuntza ezagutzea nahiz sortzea delarik helburua. Hizkuntzaren gramatika formalizatu eta konputazionalki tratatzeko moduan adierazi behar da, morfologiaz harantzago joan nahi duen edozein aplikazio edo tresnatan erabiliko bada.

Euskararen kasuan morfologia eta sintaxiaren arteko lotura estua hartu behar da kontuan lehenik. Horrek eraman gaitu tratamendu morfosintaktikoa analizatzaile morfologikoan —morfosintaktikoan, hobeto esanda— integratzera. Hitzaren barruko gramatika modu bat definitu da horrela, eta, horri esker, analisi morfologikotik beretik prestatzen zaio bidea gero etorriko den analisi sintaktikoari. Erregela multzo baten bidez deskribaturik daude, bada, hitz barruko morfemen arteko erlazioak, erlazio horietatik hitz osoaren analisiarentzat garrantzizkoa den informaziorik eratortzen denean.

Baina horretaz aparte, euskarazko perpausen joskera, sintaxia, ere deskribatu beharra dago. Lehen esan den bezala, analisi zein sorkuntza sintaktikoak ezinbesteko tresnak baititugu aplikazio gehienetan. Horrela bada, beste hizkuntza batzuetarako baliagarri suertatu diren formalismoak eta analisi-teknikak erabiltzen ari gara gu ere. Horien artean azpimarratu nahi genuke Murrizpen Gramatika[6], analisi morfologikotik ateratzen den anbiguotasun maila jaisteko eta esaldien analisi sintaktiko azalekoa egiteko erabiltzen ari garena. Edozein testu —mugarik gabe— analizatzea helburu duen formalismo horretan hitzen "hurbileko gramatika" deskribatzen duten 1.000 erregela inguru idatziak ditugu oraingoz. Horretaz gain, PATR-II izeneko baterakuntza-formalismoaz[7] euskarazko izen-sintagma eta perpaus bakunen egitura deskribatzen duen gramatika konputazionala ere garatu dugu.

Taxonomia semantikoak

Hizkuntza ulertzea xede denean, baina, ez da aski morfologia eta sintaxiarekin, semantikaz ere jakin behar izaten baitu programak. Anbiguotasun linguistikoa ebazteko beste modurik ez dago, asko eta askotan, semantikaz baliatzea baino.

Hizkuntza baten tratamendurako azpiegituran osagai semantikoak ere behar du bere lekua, beraz. Semantika lexikala da, beharbada, osagai horren prestakuntzan landu beharreko estreinako alderdia. Semantika lexikalak hitzen semantika biltzen du, lexikoko elementuen artean dauden erlazio lexiko-semantikoak: sinonimia, antonimia, hiperonimia/hiponimia (klase/azpiklase erlazioak), erlazio meronimikoak (zatia/osoa, osagaia/osoa, e.a.), eta beste. Geroago etorriko da esaldien adierazpide semantikora eramango gaituen analizatzailea.

Hiztegi arruntetan hitzen semantika lexikalari buruzko hainbat eta hainbat informazio dago. Informazio hori bertatik erauzteko —erdi-automatikoki, askotan— makina bat saio egin dira, baita gurean ere. Lan horien helburua, gehienetan, lexikoko unitateen artean erlazio lexiko-semantiko horiek esplizituki errepresentatzea izan ohi da, azkenik sare semantiko moduko bat lortzeko. Ingelesezko sare semantikoen artean ezagunena-edo WordNet izenekoa genuke[8], eta euskararako halako sare bat eratzea genuke guk ere geure epe ertaineko jomugen artean.

Hitzen forma ezagutzetik harantzago joango litzatekeen testu-zuzentzaile batek ezinbesteko luke halako tresna bat, eta gauza bera euskaraz idatzitako testuetan informazio-bilaketa egingo lukeen hizkuntzari zuzendutako tresna batek. Adiera mailako desanbiguazioan ere beharrezkoa litzateke halako sare semantikoa.

6 Karlsson F., Voutilainen A., Heikkila J., Anttila A.. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyer. 1995.

7 Shieber M.. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes, no. 4. 1987

8 Miller G.. Five papers on WordNet. *Special Issue of International Journal of Lexicography* 3(4). 1990.

Testu-corpusak

Ikerkuntza-arlo honen azpiegituran nahitaezkoa den beste elementu bat testu-corpusak ditugu. Testu-corpusak testu-masa handiak dira, informazio linguistikoaren iturri nagusietako bat eta gorago aipatutako aplikazio, tresna eta oinarrietarako probaleku ezinbestekoak. Hizkuntz corpusak lexikografian duen garrantzia ezaguna den bezala, LNPrako lexikoi bat —datu-base lexikal orokorra bera— prestatzerakoan ere premiazkoa dugu. Gramatika bat ezin da hutsetik asmatu: testuak ditugu hizkuntzaren erabileraren lekuko. Egindako tresnak eta aplikazioak ezin dira probatu laboratorioko hitz, perpaus eta esaldiekin soilik: testu errealak behar dira, egindako programa horiek gero, benetako testuei aurre egin diezaietenean, porrot egingo ez badute.

UZEI eta Euskaltzaindiaren EEBS corpusa[9] genuke, guk dakigula behintzat, egun Euskal Herrian lengoaia naturalaren prozesamenduko lanetarako dagoen corpus sistematizatu bakarra. Baina corpus hori, 3.000.000 hitzekoa izanik ere, ez da nahikoa. Testu-corpus horien biltze-lan eta antolaketa sistematikoari ekin egin behar zaio lehenbailehen, modu planifikatu batean. Lan horretan toki askotako jendeak hartu behar luke parte —Euskaltzaindia, UZEI, komunikabideak, argitaletxeak, eta abar— uste baitugu halako lan bat behar-beharrezkoa dela, honetan ari garenontzat ez ezik, baita beste ikertzaile askorentzat ere. Gaur egun ez da duela urte gutxi batzuk bezala, testuak euskarri elektronikoan egunero sortzen baitira, pilaka. Kontua da horiek biltzea, txukuntzea, eta ikertzaileon eskura jartzea.

Ikerkuntza

Bukatzeko, lengoaia naturalaren prozesamenduaren arloko ikerkuntza hartuko dugu hizpide, izan ere ikerkuntza baita hizkuntzaren tratamendu automatikoa helburu duen programa ororen oinarri ezinbestekoa. Baina, gorago esan bezala, oraingoan aipatu baino ez ditugu egingo gure taldean gaur egun esku artean ditugun proiektu eta ikergaiak.

Morfologia konputazionalaren alorrean, murrizpen gramatiken bidezko desanbiguazioa eta estatistikan oinarritutakoa ditugu aztergai. Sintaxiarenean, berriz, sintaxian oinarritutako zuzenketa, gaur edo bihar gramatika-zuzentzaile bat egitera eramango gaituena. Aipatu bi arloak ukitzen dituela, Konputagailuz Lagunduriko Hizkuntzen Irakaskuntzan ere aurreratua da tesi-lan bat.

Euskal aditzen azpikategorizazioa da beste ikergai teoriko bat, sintaxia ez ezik semantikaren arloa ere ukitzen duena, eta etorkizunean aplikazio praktiko handikoa izatea espero duguna. Semantika eta lexikoaren alorretan, berriz, semantikan oinarritutako zuzenketak dituen arazoak aztertzea eta gorago aipatu dugun taxonomia kontzeptuala eraikitzea lirateke helburuak. Horretaz gain, hitz-adieren desanbiguazioa, hiztegi arruntetatiko informazio-erauztea, ezagutza lexikalaren errepresentazio-moduen azterketa, eta giza-erabilerarako hiztegi adimendunak eta itzulpen lexikalerako laguntza-sistemak.

9 Egungo Euskararen Bilketa Sistematikoa.