Hizkuntza-Teknologiak: I+Grako garrantzia, emaitzak, eta beharrak

Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., ArtolaX.,
Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sarasola K., Soroa A.[1]

Ixa taldea

Sarrera gisa

Inprentaren sorkuntzak hizkuntzaren tratamendua eta zabalkuntza irauli bazituen, mende honetakoa dugun konputagailuak ez du iraultza txikiagoa ekarri. Hasteko, gero eta gehiago erabiltzen ditugu konputagailuak eta konputagailu-programak gure eguneroko jardunean, eta programa horietako askok eta askok testua nola edo hala "tratatu" egiten dute, prozesatu. Bestalde, konputagailuekiko komunikazioa hizkuntza arruntaren bitartez —eta ez lengoaia formal baten bidez— egin ahal izatea, gero eta normalago izango da. Gizarte eleanitzak hizkuntza batetik bestera egin behar izaten dituen joan-etorriak leuntzeko ere, aparteko lagun dugu konputagailua. Gainera, telekomunikazioetan gertatutako aurrerapen izugarriak eragin duen Internet fenomenoak, areagotu egin du hizkuntzaren tratamendu automatikoaren beharra; interesatzen zaigun informazioa ondo selekzionatzeko, esaterako, tratamendu linguistiko lagungarria ezinbestekoa baita.

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari *lengoaia naturalaren prozesamendua* (LNP) esaten diogu informatikariok, nahiz eta, batzuetan, hizkuntzalaritzaren ikuspuntutik erreparatuta, batez ere, *linguistika konputazionala* ere esan. *Hizkuntzaren industria* oso bat sortzen ari da, konputagailuaz baliatuz hizkuntza prozesatzea helburu duena. Artikulu labur honen helburua, beraz, *giza hizkuntzaren teknologia* esaten zaion hori zertan den azaltzea da, horretarako EuskoNews-en argitaratu genuen artikulua[2] eguneratu eta aberastu dugu datu berriekin.

Hasieran datuekin erakutsiko dugu zelako garrantzia ematen zaion gaiari Europako Batasunean eta Eusko Jaurlaritzan ikerketa eta garapena (I+G) suztatzeko deialdietan. Gero LNParen barruan azaltzen diren sistemak eta produktuak hobeto aurkeztearren, beste bi atal bereizi ditugu artikuluan: lehenengoan, "kaleko erabiltzailearentzat" salgai diren **aplikazioak** sartu ditugu, hizkuntza automatikoki tratatzeak zer helburu praktiko dituen azalduz; bigarrenean, aplikazio horiek sortuko badira zer-nolako azpiegitura behar den azaltzen saiatu gara, hizkuntza-softwarea sortzen dutenentzako **tresnak**, eta edozein aplikazio edo tresna garatzeko eratu behar diren **hizkuntza-baliabide** eta **-oinarriak** aztertuz.

Artikulu honetan aipatuko ditugunak Donostiako Informatika Fakultateko IXA taldearen[3] esperientziari dagozkio, gehienbat. Hizkuntzalariz eta informatikariz osaturiko ikertalde honetan, hamahiru urte inguru daramatzagu euskara idatziaren tratamendu automatikoan lanean, gure hizkuntza, arlo honetan, besteen pare egon dadin ahalegintzen, horretarako beharrezkoa den ikerkuntza sustatu eta azpiegitura prestatuz. Horregatik, artikuluan zehar

---

1 Donostiako Informatika Fakultateko IXA taldekoak.

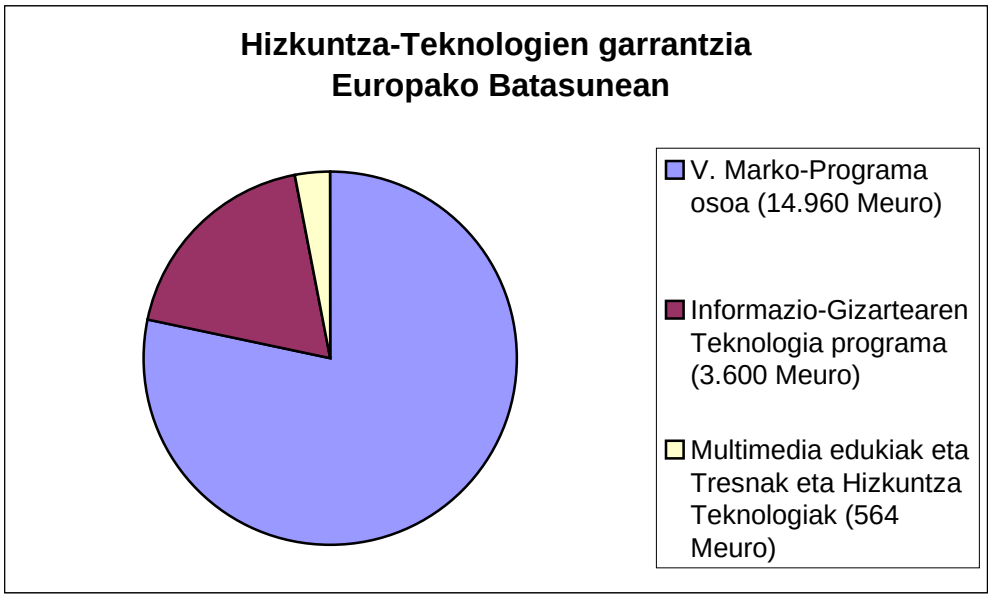
2 http://suse00.su.ehu.es/euskonews/0110zbk/frgaia.htm

3 http://ixa.si.ehu.es/

euskarari dagozkion oharrak egingo dira, bide luze honetan eginda dagoena eta egiteko dagoena zer den argitze aldera.

Hizkuntza-Teknologien garrantzia I+G deialdietan

Erakunde ofizialetatik bultzatu egiten da ikerketa lerro hau. Europako Batasunak, IV. eta V. Marko-Programei loturik, berrikuntza-programa europarrek gizartearen sektore ugariren interesa piztu edo suspertu egin du eta, horrekin batera, sortu berri baina biziki erakargarri diren merkatuetan etorkizunean aplikatzeko ideia fresko eta berrizaleen sorreraren eragile izan dira. Hizkuntzen Teknologiei aitortzen dion garrantzi handia erakusteko nahikoa da esatea DGXIII zuzendaritza orokorraren I+Grako aurrekontuaren %3,77 (564 milioi euro) bideratuko dela "Multimedia edukiak eta Tresnak eta Hizkuntza Teknologiak". Izan ere, V. Marko-Programaren aurrekontu osoa 14.960 milioi eurokoa[4] da, eta hortik 564 milioi euro bideratuko dira "Multimedia edukiak eta Tresnak eta Hizkuntza Teknologiak", berau izanda "Informazio-Gizartearen Teknologiak" programako (3.600 millioi euro) III. lerro estrategikoa (Key Action III).



Eusko Jaurlaritzako onartu berri duen Zientzia, Teknologia eta Berrikuntzarako Planak ere (ZTBP- PCTI 2001-2004), era berean, garrantzia handia eman dio ikerlerro berri honi. Horrela, Informazio-Gizartearen Teknologiei dagokien atalari portzentaia handiagoa ematen dio Europako Batasunak baino (Europan %24 eta hemen %31,4) eta atal horren barruan Infoingeniaritza Linguistikoa lerro estrategiko gisa azaltzen da beste hiru lerrorekin batera. Beraz, Eusko Jaurlaritzak 2001-2004 urteetarako aurreikusi dituen 614 milioi eurotik 10 milioi inguru Infoingeniaritza Linguistikoan inbertitu nahi izango dira. Estatutik, Europako Batasunetik eta iturburu pribatuetatik etorriko diren diru-ekarpenak ere kontuan hartuta lau urtetan batezbeste 6.000 milioi pezeta (360 Meuro) mugituko dira ikerketa lerro honetan.
Dagoenekoz enpresa, ikertalde eta erakunde ofizialen artean 20 baino gehiago dira Hizkuntza-Teknologiaren esparruko eragileak. Horien artean aipatu daitezke ondokoak: Adur Software Productions S. Coop., Ametzagaiña AIE, Aurten Bai Fundazioa, BAI & BY, ELHUYAR, Euskal Herriko Unibertsitateko Ingeniaritza Eskolako Ahots Taldea, Euskal Herriko Unibertsitateko Ixa Taldea, Euskal Herriko Unibertsitateko Zientzien Fakultatea,

[4] http://www.cordis.lu/fp5/src/budget.htm

Euskaltzaindia, Eusko Ikaskuntza, Eusko Jaurlaritzako Kultura Sailaren Hizkuntza Politikarako Sailordetza, Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Saila, Eusko Jaurlaritzako Industria, Merkataritza eta Turismo Saila, GEINSA, HABE, Ihardun Multimedia, Interlinea 2000, KAIXO, LKS S. Coop., Telefonica, UZEI eta Zabaltzen. Hauek guztiok bildu zituzten lehengo urtean Eusko Jaurlaritzako Industria, Kultura eta Hezkuntza sailek "Hizkuntzaren Industria" delako ekimen proposamena diseinatzeko.

Aplikazioak

Merkatuan aurki daitezkeen aplikazio gehienek hizkuntza "handiak" dituzte helburu, ingelesa, batik bat, baina baita, bigarren maila batean bada ere, frantsesa, alemanera eta espainiera bezalako hizkuntzak ere.

LNParen ia 50 urteko historian gorabehera handiak izan dira. Helburu liluragarriak lortzear zeudela uste zen une euforikoei, belarriak jaitsi eta helburu apal baina eskuragarriagoetara mugatzeko une pragmatikoak jarraitu zaizkie behin baino gehiagotan. Konputagailuek hizkuntza pertsonok ulertzen dugun moduan ulertuko duten eguna urrun da oraindik, baina horrek ez du esan nahi aplikazio interesgarri eta oso baliagarriak egin ezin direnik. Erabateko itzulpen automatikoa konputagailuen eskutik etorriko zela aurreikusi zuten 1954an Georgetown-eko Unibertsitatean. 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 konputagailuak hizkuntza arruntaz —lengoaia naturalean— programatu ahal izango genituela agindu zitzaigun. 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

Konputagailua etxeraino sartu bazaigu, "idazmakina azkar eta memoria onekoa" delako izan da, aurrena, eta baita, azken aldian batik bat, Interneten bidez hainbat informazio eskuratzeko tresna bikaina delako. Konputagailuak erraztasun handiak eskaintzen ditu, testuak sortu, kopiatu, osatu eta zuzentzeko. Eta, gainera, testu-egileari hizkuntzarekin zerikusi zuzena duten laguntza bereziak eskaintzen ahal dizkio. Hala nola:

- Ortografia-zuzentzaileak, gaur egun hizkuntza askotarako aurki daitezkeenak. Testuko hitz bakoitzaren ortografia egiaztatzen dute —testuingurua kontuan hartu gabe—, eta, okertzat jotakoan, ordezko posibleak proposatzen dituzte. Euskara bezalako hizkuntzen kasuan, hitzak kasu desberdinetan deklinatuta agertzen direnez, hitzaren analisi morfologikoa egin behar da. Euskarako egiaztatzaile/zuzentzailea, *Xuxen*, Microsoft Office-n integratua dago, eta doan eskura daiteke[5].
- Idazkera- eta gramatika-zuzentzaileak ere merkatuan dira hainbat hizkuntzatarako; eta hauek testuingurua kontuan hartzen dute, noski. Nahiz eta, gaur egun, hutsegite guztiak harrapatu ez, laguntza polita eskaintzen diote idazlariari.
- Hiztegi-laguntza integratuen arloan ere, era askotakoak aurki daitezke: sinonimo eta antonimoak ematen dizkigunetatik hasi, eta edozein hiztegi edo thesaurus testu-

5 http://www.euskadi.net/euskara_hizt/indice_e.htm
http://www.sc.ehu.es/xuxen-e.htm

prozesadoretik irten gabe kontsultatzeko aukera eskaintzen digutenetaraino. Teknologia prest dago, eta aurki izango ditugu horrelakoak gure artean, euskaraz idazten duenarentzat lagungarri.

- Itzulpen-lanetarako programak ere prozesadore zabalduenetan integratzen dira, eta glosategi, hiztegi eta itzulpenen berrerabilpenerako laguntzak —itzulpen-memoriak, adib.— eskaintzen dituzte, antzeko testuak itzuli behar direnean, testuen bertsio berriak egiterakoan, etab., itzultzaileari lana erraztuz. Ezagunenetako bat *Trados* izenekoa da[6].

Testu-masa handiak tratatu edo kudeatzerakoan, berriz, aplikazio hauek aurkituko ditugu:

- Kontzeptu-bilatzaileak, datu-base dokumentaletan bilaketak egiten dituztenak. Sistema hauek orain, hitz gakoen konbinazio boolear hutsetik harantzago, LNPko teknika gero eta sofistikatuagoak erabiltzen dituzte, hala nola, lematizazioa, perpausen bukaeren detekzioa, akronimoen zabaltzea eta kalkulu estatistikoak. Ametzagaiña taldeak kaleratutako *Kapsula* softwarea[7] euskarazko dokumentu-baseen gestiora zuzendua dago. Ixa taldearen lematizazioan oinarritutako bilatzailea integratuta dago Euskaldunon Egunkariaren hemeretokan[8] eta ZientziaNet[9] web-zerbitzuetan.
- Kategorizazio-sistemak oso baliagarriak dira makina bat dokumentu (telefonoetako matxura-parteak, albisteak, adib.) 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-konpainiak daukan sistemak matxura-parteak automatikoki bideratzen ditu, konponketaz arduratu beharko den bulegoraino.
- Informazio-erauzketako sistemek, hizkuntza arruntean idatziriko testuetatik abiatu eta datu-base egituratu bat osatzen dute (ekintza edo gertaeraren nor-noiz-nongoak zehaztuz), gero informazioa errazago aurkitu ahal izan dadin.
- Testu-sorkuntza automatikoa informazio-erauzketaren alderantzizkoa da. Kasu honetan, konputagailu barruan dauden datu egituratuetatik abiatuz, datu horien edukia azalduko zaio erabiltzaileari bere hizkuntzan. *Forecast Generator* sistemak ingeles edo frantsesezko testuak idazten ditu konputagailu batek kalkulatzen dituen eguraldi-iragarpen kodetuetatik abiatuz. Eguraldi-iragarpenon testua euskaraz —eta frantsesez, ingelesez, alemaneraz, nederlanderaz, gaztelaniaz, katalanez, eta galegoz— ematen duen sistema bat garatzen ari da gaur egun, *MultiMeteo* proiektuaren barruan[10].

Itzulpen automatikoa

Produktu ugari dago merkatuan salgai, testu-itzulpenean laguntza emateko. Itzulpen perfektua egiten duen sistemarik ez dago inon, eta sistema bat bera ere ez da gai testu literarioak behar bezala itzultzeko. Gehienek itzulpen teknikoa dute erabileremu, testu teknikoetan anbiguotasun gutxiago egoten baita hizkuntzen arteko hitzen eta esaldien korrespondentzian.
Itzulpenaren automatizazioa ez da ia inoiz erabatekoa, eta automatizazio-mailaren arabera ondoko sailkapena egin ohi da: 1) erabateko itzulpen automatikoa: errealitatea baino, ametsa da gaur egun, non eta helburua ez den edukiaren ideia orokor bat ateratzea; 2) giza laguntzaz egindako konputagailu bidezko itzulpena: lanaren gidaria makina da, baina fase desberdinetan

---

6 http://www.trados.com/

7 http://www.kapsula.com/

8 http://www.egunkaria.com/hemeroteka/

9 http://www.zientzia.net/

10 http://www.inm.es/wwi/MultiMeteo/Multimeteo.html

laguntzak eska ditzake, hitz baten adiera zuzena hautatzeko edo esaldi baten analisiari nondik ekin behar zaion galdetzeko, adibidez; 3) konputagailuz lagunduriko giza itzulpena: gidaria pertsona da, baina konputagailuaz baliatzen da hiztegi berezituan kontsultak egiteko, testuaren formatua txukuntzeko, eta zailtasunik gabeko testu-zatiak itzultzeko. Kasu honetan, batzuetan itzulpenaren zati handi bat konputagailuak egiten du ia laguntzarik gabe, baina beharrezkoak izaten dira aurreprozesaketa —testua egokitzeko— eta postedizioa —emaitza zuzentzeko.

Sistemak aipatzen hasita, Montrealeko TAUM taldeak egindako *Meteo* sistema da emaitzarik arrakastatsuena lortu duena. Parte meteorologikoak itzultzen ditu, 1977tik hona, ingelesetik frantsesera, eta itzulpenaren %80 erabat zuzena da. Bestalde, SYSTRAN Institutua izan da, 1970. urteaz geroztik, itzulpen automatikorako tresnen saltzaile nagusia, eta NASA, Europako Elkartea, General Motors eta Xerox ditu bere bezeroen artean. Interneteko *Altavista* bilatzailean ere eskaintzen da itzulpen-zerbitzu automatiko bat, *Systran*-en oinarritua[11]. Siemens-ek garatu *METAL* da beste sistema sonatu bat, testu teknikoen itzulpenera zuzendua.

Konputagailu pertsonaletan, berriz, dozenaka produktu dago itzulpenak egiteko: *Spanish Assistant*, *Power Translator*, etab. Guztietan beharrezkoa da postedizioa, eta nolabaiteko elkarrekintza dago beti giza itzultzailea eta programaren artean, hitzen adiera zuzena hautatzerakoan eta.

Katalunian, *El Periódico* egunkaria gaztelaniaz eta katalanez kaleratzen da egunero, itzulpen-sistema bati —eta postedizioaz arduratzen den 20 pertsonako lantaldeari— esker. Bestalde, gaztelaniatik katalanera itzultzen duen sistema bat ere proba daiteke doan Internet bidez[12].

Konputagailuen erabilera LNaren bidez

Aplikazio-mota honetako sistemek, konputagailu eta gizakiaren arteko komunikazioa hizkuntza arruntean bideratzea dute helburu. Horrelako sistemak inplementatzen zailak dira: galdera eta erantzunez osatutako elkarrizketa ulertu ahal izateko, mintzakideen planak eta helburuak aztertzeko tresnak behar dira. Hiztun bakoitzak momentu bakoitzean zer dakien eta zer nahi duen asmatzeko gai izan behar du sistemak, eta, gainera, ezagumendu horiek etengabe eguneratzen ibili behar du elkarrizketa aurrera joan ahala. Helburu orokorrekorik ez da luzaroan salgai egongo, baina badira dagoeneko aplikazio konkretuei lotuta dauden batzuk.

Datu-baseen galdeketa-sistema ugari dago, batez ere ingelesez. Datu-base konplexuetan kontsultak egin ahal izateko lengoaia berezi bat ezagutu beharrak datu–baseen erabiltzaile potentzialen kopurua murrizten duenez, galderak hizkuntza arruntean egin ahal izatea oso interesgarria da. Symantec-en *Question & Answer (Q&A)* sistemak arrakasta ederra izan du, 1986az gero. Merkatuan 100etik gora dira horrelako produktuak, denak ere ingelesezkoak. Zenbait kontzeptu-bilatzailetan ere egin daitezke galderak ingeles arruntean[13].

Ahozko hizketaren tratamendua

Merkatu handia zabaldu da ahozko hizketa prozesatzen —eta ulertzen— duten sistementzat. Sistema hauek, batez ere telefono bidezko zerbitzuetan integratzen dira oraingoz: aurretiko hitzordua, produktu-eskaerak, ikuskizunetarako erreserba-eskea, telefonogune automatikoak, e.a. Baina badaude bestelakoak ere: diktaketa automatikoa, adibidez.

Egun, aurretiko hitzordua ematen duten sistema gehienek zenbakiak eta astegunen izenak besterik ez dituzte ulertzen, baina, hala ere, ekonomikoki interesgarriak diren aplikazioak egin dira horrela. Natural Vox enpresa arabarrak[14] aurretiko hitzordua —medikuarenean, eta errenta-aitorpena egiterakoan— automatikoki lortzeko sistema telefonikoak ezarri ditu azken

---

11 http://babel.altavista.com/translate.dyn

12 http://www.softcatala.org/traductor/

13 http://www.askjeeves.com/

14 http://www.natvox.es/

urteetan, eta arrakasta handiz, gainera. Sakelako telefonoen munduan eta Internetekoan hainbat produktu ari da kaleratzen, non informazio idatzia "ahoz" ematen baitzaio erabiltzaileari. Euskaltelek eta Telefónica-k garatuak dituzte, mezuak euskaraz irakurtzen dituzten oinarrizko sistemak.

Aplikazioak garatuko badira, zer-nolako azpiegitura behar da?

Artikuluaren bigarren parte honetan, eta oso labur, halabeharrez, abiaburuak deskribatuko ditugu, aipatu ditugun aplikazioak eta produktuak sortzera helduko bagara antolatu beharko genituzkeenak, beti ere gure taldean markatutako estrategiari jarraituz. Abiaburuon artean funtsezkoa dugu, jakina, arloko ikerkuntza. Hala ere, artikulu honetan aplikazio horiek garatzeko tresnak, eta aplikazio eta tresnok egin ahal izateko oinarriak azalduko ditugu batik bat, ikerkuntzarekin zer ikusia dutenak beste baterako utziz.

Tresnak

Atal honetan hizkuntzaren tratamendurako aplikazio-ekoizleentzat edo arloko ikertzaileentzat interesgarriak diren tresna batzuk ikusiko ditugu. Tresna horiek ez dira sortu, beraz, "kaleko erabiltzailearengan" pentsatuz.

Ahozkotik idatzira

Lehen oinarria, hizketa prozesatu nahi bada, ahozkoa testu idatzi bihurtuko digun tresna da. Ahozko hizketa ezagutzea ez da erraza: hitzak ez dira ongi bereizten bata bestetik, intonazioa dago, eta, gainera, seinale fisikoen zarata ere oztopo da. Euskal Herrian badira gai honetan diharduten bizpahiru ikertalde —Bilboko Ingeniaritza Eskolako Aholab izenekoa bat[15], Leioako Zientzia Fakultatean beste bat[16].

Analizatzaile morfologikoa

Hizkuntza guztietan beharrezkoa, eta euskara bezalako hizkuntza flexionatu eta eranskarien kasuan ezinbestekoa, analizatzaile (eta sintetizatzaile) morfologikoaren zeregina hitz-forma osatzen duten morfemak ezagutzea (eta konposatzea) da, eta morfema bakoitzari dagokion informazio morfologiko-lexikala ematea. Erreminta hau oinarri da hainbat aplikaziotan, hala nola, zuzentzaile ortografikoa, karaktere-ezagutze optikoa (OCR), eta aplikazio sofistikatuago guztietan —itzulpen automatikoa, adib.—. Euskarako analizatzaile/sintetizatzaile morfologiko orokorra egina dago, eta *Xuxen*-en funtsa da.

Lematizatzaile/etiketatzailea

Lematizatzaile/etiketatzailea analizatzaile morfologikotik eratortzen da, eta hitz-forma baten lema eta kategoria ematen ditu, anbiguotasuna saihestu edo gutxitzearren testuingurua aintzat hartuz. Zeregin nagusia desanbiguazioa bada ere, beste egitekorik ere badu halako tresna batek, esate baterako, hitz anitzeko unitate lexikalen identifikazioa (lokuzioak, hitz-elkarketak, pertsona-izenak, etab.). Oso aplikazio interesgarriak dituzte lematizatzaileek: indexazioa —Interneteko bilatzaileetan, adib.—, terminologia eta lexikografia, etab.

15 http://bips.bi.ehu.es/ahoweb/

16 http://sirius.we.lc.ehu.es/ (Reconocimiento automático del habla)

Euskarako lematizatzaile orokorrari *EusLem* izena eman diogu, eta ezarrita dago jadanik Euskaldunon Egunkariaren eta Jalgi zerbitzariko bilatzaileetan[17].

Analizatzaile sintaktikoa

Analizatzaile sintaktikoen zeregina, testuetako osagai sintaktikoak ezagutzea da: perpausak, izen-sintagmak, izen-lagunak, etab. Analisiaren oinarria lexikoa eta gramatika izango dira, hitzen ezaugarriak eta egitura sintaktikoen osaketa posibleak definituko dituztenak. Hau ere ezinbesteko tresna dugu hizkuntza-aplikazio askotan, itzulpen automatikoan, esate baterako. Euskararen kasuan, azaleko analizatzaile sintaktiko orokorra egina dugu —*EusMG*—, eta zuhaitz sintaktiko osoa emango digunaren ikerbideak nahiko aurreratuta daude.

Hizkuntza-baliabideak eta -oinarriak

Azkenik, aplikazio eta tresnon zimentarria diren hizkuntza-baliabide eta -oinarriak hartuko ditugu hizpide, artikulua bukatzeko.

Datu-base lexikala eta morfologiaren deskribapena

Datu-base lexikala da hizkuntza-lexikoaren biltegi orokorra. Hiztegi elektroniko moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahiak dituen eskakizunak kontuan harturik antolatua. Horrek lexiko-deskribapenaren sistematizazio bat eskatzen du: sarreren kategoria-sistema bateratu eta homogeneoa, kategoria bakoitzeko elementuak behar den bezala deskribatzeko beharrezko diren ezaugarriak zehaztea, etab. *EDBL*, euskararen datu-base lexikalak[18], 75.000 sarrera inguru biltzen ditu egun —hiztegi-sarrerak, adizkiak eta morfema ez-independenteak—, eta IXA taldea arduratzen da egunean mantentzeaz.

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.

Gramatika konputazionalak: sintaxiaren deskribapena

Sintaxia ere funtsezkoa dugu hizkuntzaren tratamenduaren arloko edozein lani ekiteko, helburua hizkuntza ezagutzea nahiz sortzea dela ere. Hizkuntzaren gramatika formalizatu, eta konputazionalki tratatzeko moduan adierazi behar da, morfologiaz harantzago joan nahi duen edozein aplikazio edo tresnatan ustiatuko bada. Euskararen kasuan, gainera, morfologia eta sintaxiaren arteko lotura estua hartu behar da kontuan. Horrek eraman gaitu tratamendu morfosintaktikoa analizatzaile morfologikoan integratzera: *Morfeus* izeneko analizatzaile morfosintaktiko orokorra da emaitza.

Taxonomia semantikoak

Hizkuntza ulertzea xede denean, baina, ez da aski morfologia eta sintaxiarekin, semantikaz ere jakin behar izaten baitu programak. Anbiguotasun linguistikoa ebatzi ezina da, askotan, semantikaz baliatu ezean. Hizkuntza baten tratamendurako azpiegituran, osagai semantikoak ere behar du bere lekua, beraz. Eta semantika lexikala da, beharbada, osagai horren

---

17 http://www.egunkaria.com/egun1999/sarrera.html
http://www.jalgi.com/

18 http://sipl54.si.ehu.es/

prestakuntzan landu beharreko estreinako alderdia. Semantika lexikalak lexikoko elementuen artean dauden erlazio lexiko-semantikoak biltzen ditu: sinonimia, antonimia, hiperonimia/hiponimia (klase/azpiklase erlazioak), eta beste. Erlazio lexiko-semantiko horiek sare semantiko moduko batean adierazten dira esplizituki. Ingelesezko sare semantikoen artean ezagunena-edo *WordNet* izenekoa dugu, eta haren euskararako egokitzapenari *Euskal WordNet* deitzen diogu.

Testu-corpusak

Eta azkenik, ikerrarlo honen azpiegituran nahitaezkoa den beste elementu bat aipatuko dugu: testu-corpusak. Testu-corpusak testu-masa handiak dira, informazio linguistikoaren iturri nagusia, eta gorago aipatu aplikazio, tresna eta oinarrietarako probaleku ezinbestekoak. Hizkuntza-corpusek lexikografian duten garrantzia ezaguna da. Era berean, LNPrako lexikoi bat edo gramatika konputazional bat ezin dira hutsetik asmatu, eta, horretarako, corpusak ezinbestekoak dira. Bestalde, garatutako tresnak eta aplikazioak ezin dira probatu laborategiko hitz, perpaus eta esaldiekin soilik: testu errealak behar dira.

Testu-corpusen biltze-lan eta antolaketa sistematikoari ekin egin behar zaio lehenbailehen, modu planifikatu batean. Lan horretan, arlo askotako jendeak ez ezik, instituzioek ere parte hartu behar lukete, halako testu-bilduma handi bat behar-beharrezkoa baitugu, honetan ari garenok zein beste hainbat ikertzailek ere. Testuak euskarri elektronikoan egunero sortzen dira pilaka —argitalpen hauxe dugu adibide—: kontua da horiek sistematikoki biltzea, txukuntzea, eta ikertzaileon eskura jartzea.

Bukatzeko

Artikuluaren helburua Hizkuntza-Teknologien arloaren ikuspegi orokor bat ematea izan da. Gauza asko aipatu dira, baina, ezinbestean, oso labur. Hala ere, espero dugu irakurlearentzat lagungarri izango direla orri-oineko oharretan ipini ditugun web helbideak, interesik izanez gero, haritik tira eta informazio aberatsago eta sakonagoa eskura dezan.

2000ko azaroa.

IXA taldea 1988an sortu zen euskararen tratamendu automatikoan ikertzeko. Gaur egunean taldean hamazazpi informatikari eta hamar hizkuntzalari biltzen gara, denetara hamabi doktore garela. Emaitzen artean lau aplikazio orokor kaleratu ditugu: XUXEN zuzentzaile ortografikoa, Elhuyar hiztegia Word2000 editorean integratua, GAIN web-bilatzailea, eta eguraldi-iragarpenen euskarako bertsioaren sorkutntza automatikoa Multimeteo sisteman. Harremanak ditugu enpresa hauekin: Microsoft, Lexiquest France, Eatoni, UZEI, Elhuyar, Plazagunea, Euskaldun Egunkariarekin eta Hizkia.

ixa.si.ehu.es