EUSKARA ETA INFORMAZIOAREN TEKNOLOGIAK. EGUNGO EGOERATIK ETORKIZUNERA BEGIRA

Andoni Sagarna Izagirre,
Euskaltzaindiko ***sarrera-hitzaldia***

Ahaldun nagusi jauna, agintari agurgarriok, euskaltzainburu jauna eta euskaltzainok, etxekook, adiskideok, jaun-andreok: arratsalde on digula guztioi.

Hizkuntza eta teknologia, hara hor denok egunero erabiltzen ditugun bi gauza. Jende askorentzat, ordea, bata bestetik urrun dauden arloak dira. Behar bada mundu akademikoan letretako eta zientzietakotzat hurrenez hurren hartu ohi zirelako garai batean. Garai batean diot, zeren gaur egun urruntasun hori, neurri handi batean, desagertua baita.

Bi eremu horien atzoko, gaurko eta biharko hurbiltasunaz eta harremanez hitz egitera natorkizue hain zuzen ere.

Badakit zuetako askok *teknologia* hitza entzuten duzuenean, erreparo moduko bat sentitzen duzuela eta batek baino gehiagok pentsatuko zuen honezkero: zer egiten ote du ingeniari batek Euskaltzaindian?

Horrela bada, harridura astintzeko asmoz, lehenik eta behin, ingeniari bat hizkuntza-arazoez arduratzea, euskara lantzen saiatzea eta are Euskaltzain oso izatea ere, hain gauza bitxiak ez direla agertzen saiatuko naiz.

Urruti samarretik hasiko naiz, pixkanaka gure artera hurbiltzeko.

XX. mendearen hasieran, Alemaniako ingeniariak arduratuta zebiltzan euren lanbidean behar zituzten hitz teknikoez. Batez ere itzulpen teknikoan egiten ziren okerrak saihesteko asmoz, 1900ean gutxi gorabehera, VDI (Verein Deutscher Ingenieure) erakundeak, hau da, Alemaniako Ingeniarien Elkarteak, Hubert Jansen filologo eta lexikografo ospetsuari *Technolexikon* izenburua izango zuen aleman/ingeles/frantses hiztegi tekniko bat zuzentzeko ardura eman zion. Jansen, besteak beste, Natur Zientzietako eta teknikako maileguzko hitzak alemanez izan behar zuten ortografiari buruzko lan baten egilea izan zen. Egitasmo handia zen Technolexikon hori. Bi mila enpresa eta lankide boluntario agertu ziren, hiztegiari ekarpenak egiteko prest. Dozena erdi bat urtetan 3.600.000 fitxa egin zituzten, lexikografian ohikoa izan den bezala, alfabetoaren hurrenkeraren arabera hitzak aztertuz. VDIren kontseiluak 1907an egin zituen kalkuluen arabera, Technolexikon hiztegia burutzeko, ordu arteko

metodologia erabiltzekotan, beste 40 urte beharko ziren. Gehiegi ari zen korapilatzen gauza eta noraezean zebiltzan.

Bitartean, 1906an, Alfred Schlomann (1878-1952) ingeniariak makina-atalei buruzko hiztegitxo bat argitaratu zuen, sistematikoki ordenatua, kontzeptu-sistemaren arabera, alegia, eta irudiz hornitua, sei hizkuntzatan (alemana, frantsesa, ingelesa, errusiera, italiera eta espainiera). Lan honek harridura sortu zuen, hiztegi teknikoak lantzeko eta aurkezteko modu berri bat ekarri zuelako.

VDIk aholkua eskatu zion Koloniako Unibertsitatean filologia ingelesa irakasten zuen eta lexikografoa zen Arnold Schroer-i, Technolexikon-en garapenari buruzko erabaki bat hartu ahal izateko. Hark Schlomann ingeniariaren metodologia hobetsi zuen eta Technolexikon zelako hura bertan behera utzi eta Schlomann-en lana finantzatzen hasi zen VDI elkartea. Ondorengo urteetan, 1932.a bitartean, 17 tomo lodi egin zituzten. Lanean jarraitu zuten, harik eta naziek lana gerarazi eta 1935ean Schlomann-ek Alemaniatik erbesteratu beharra izan zuen arte.

Dena esan behar bada, Shlomann-en aurretik, Heinrich Paasch itsasgizon jatorriz alemaniar baina belgikar nazionalizatuak, *De la quille à la pomme de mât* izenburupean, argitaratu zuen itsasontzien terminologia, antzeko metodologiaz landua zen. Hala ere, Schlomann ingeniariak terminologia teknikoa lantzeko metodologiari bide berriak urratu zizkiola esan dezakegu.

Geroxeago beste ingeniari batzuek, hala nola Wüster austriarrak, Drezen letoniarrak, Lotte errusiarrak eta Selander suediarrak metodologia horri oinarri teorikoa ezarri zioten.

Aipatu ditudan ingeniari horiek, hiztegi teknikoak lantzen ibili ziren, euren lan-arlo propiotik hurbileko hizkuntza-arazotan beraz, baina beste batzuk izan dira Hizkuntzalaritzaren ur sakonagotan murgildutakoak ere.

Benjamin Lee Whorf 1918an Ingeniaritza Kimikoan graduatu zen Massachusetts-eko Teknologia-Institutu ospetsuan. Lanbide horretan lanean aritu zen aseguru-etxe batean, eta 1931n Hizkuntzalaritza ikasten hasi zen Yale-ko Unibertsitatean. Han Edward Sapir-en ikaslea izan zen lehenbizi eta lankidea gero. Bien izena darama, hain zuzen, Sapir-Whorf hipotesiak, alegia, pertsona batek hitz egiten duen hizkuntzako kategorien eta pertsona horrek duen mundua ulertzeko eraren artean erlazio sistematikoa dagoela dioen hipotesiak. Whorf-en ustez, pentsamendua hizkuntzak baldintzatzen du eta hizkuntza desberdinak hitz egiten dituzten pertsonek mundua ikusteko era desberdinak dituzte.

Whorf-ek Hizkuntzalaritzan ospe handia lortu zuen arren, inoiz ez zuen hartu Hizkuntzalaritza lanbidetzat eta esaten zuen, mundu akademikoaz kanpoko irabazpidea izateari esker, hobeto eta askatasun handiagoz ahalegindu zitekeela gai akademikoetan. Are gehiago, gauza jakina da aseguru-

-etxean egiten zituen lanek ekarri ziotela hipotesi famatu hori formulatzeko inspirazioa.

Industria-arloan gertatzen ziren suteak zerk eraginak ziren ikertzen ibiltzen zen Whorf, aseguru-etxearentzat. Sute bat gertatzen zenean, bertara joaten zen eta instalazio elektriko akastunen batek, baldintza txarretan metatutako erregairen batek edo beste eragile fisikoren batek sua piztu ote zuen ikertzen zuen. Ondoren txosten bat idazten zuen aurkitu zuena adieraziz. Alabaina, esperientziak irakatsi zion, gauza bat zela egoera fisikoa bera eta beste bat egoera horrek jendearentzat zuen esanahia. Esanahi hori, hain zuzen, istripua baldintzatzen zuen faktoreetako bat izaten zela, jendeak sutearen aurrean zuen portaeraren bitartez. Eta esanahiaren eragina argiagoa zen, egoera deskribatzeko erabiltzen zen hizkerarekin lotuta zegoenean. «Gasolina-bidoiak» zeudela esaten zen leku batean, arreta handiagoa jartzen zen «Gasolina-bidoi hutsak» zeudela esaten zen leku batean baino. Bigarrenean jendeak lasaiago pizten zituen pospoloak, esate baterako, nahiz eta bidoi hutsak, berez, arriskutsuagoak izan, lurrun lehergarria zeukatelako. Fisikoki egoera arriskutsua izan arren, «huts» hitzak arriskurik eza iradokitzen duela zioen Whorf-ek.

Hori bezala beste adibide asko erabiltzen zituen Whorf-ek, bere lanetan eta hitzaldietan, hizkuntzak pentsamendua eta portaera nola baldintzatzen dituen erakusteko.

Horretaz gain, Erdialdeko Amerikako hizkuntzak ikertu zituen.

Ikusi duzuen bezala, Whorf, ingeniaria izanik ere, Hizkuntzalaritzaren eta Hizkuntzaren Filosofiaren arloetan ibili zen.

Badira beste ingeniari batzuk, teknologia-arloko terminoak edo hizkuntza bera aztergaitzat hartuta lan egin ordez, hizkuntzak adierazpide den aldetik, ideiak zorroztasunez azaltzeko batzuetan eta jostatzeko beste batzuetan abilezia erakutsi dutenak.

Horietakoa izan zen Serafin Baroja Zornoza (1840-1912), Pio Baroja idazlearen aita. Meatzetako ingeniaria zen eta Huelvako Riotinton, Euskal Herrian, Valentzian, Madrilen eta Granadan lanbide horretan ibili zen. Gaztelaniaz eta euskaraz idatzi zuen. Liberala zen eta karlisten aurkako kantak idatzi zituen, Manterolaren *Cancionero Vasco* bilduman jasota daudenak.

Bere jaioterri Donostiari, ereserkiaren hitzak ez ezik, maitasunezko bertsoak ere eskaini zizkion.

Hark hizkuntzari etekina ateratzeko zuen gaitasuna ederki erakusten du behin Donostian «a» letrari buruz ordubetetik gorako hitzaldi umoretsua eman izanak.

Horrek ez du esan nahi, ordea, hizkuntzaren azterketa axola ez zitzaionik, gauza ezaguna baita Migel Unamunorekin batera izaten zela Bilboko tertulia batean eta biak euskal gramatikako gaiak eztabaidatzen zituztela.

Unamuno aipatu dudan honetan, hark bere adiskide ingeniari batekin bizi izan zuen pasadizoaz kontatzen omen zuena datorkit burura.

Behin batean, Unamuno eta Bilboko Ingeniaritza Eskolan Kimikako katedraduna zen adiskide katalan bat, industria-ingeniaria bera, Loiolako Santutegia bisitatzera joan omen ziren. Etxe hartako nagusiak lagundu omen zien bisitan eta, nobizio batzuk zeuden lekura iritsi zirenean, nagusi horrek, jesulagunek duten azkar-famaz harro, esan omen zien: ikus itzazue gazte horiek… hori da fintasuna, hori da arraza, hori da adimena!

Jakina, ba omen ziren han burutsu-itxura zuten ikasleak eta agian hain bizkorrak izango ez zirenak eta beste batzuk, alderantziz, hain argiak ez ziruditen arren, itxuraren kontrakoak izango zirenak, eta ba omen ziren beste batzuk motz-aurpegi nabarmena zutenak ere.

Haietako batengatik Unamunok galdetu omen zion jesuitari: «zuk uste duzu mutil horrek argia izateko itxura duela? Ezin dut esan hala ez dela, baina itxuragatik esaten dut.»

Unamunoren ateraldi horren aurrean, jesulaguna hitzak bilatu ezinda bezala hasi omen zen erdi totelka: *jakina … baliteke …* eta halako egoera gogaikarri bat sortu omen zen.

Orduan, ingeniari katalanak, naturaltasun osoz, ezer gertatu ez balitz bezala, esan omen zuen: berdin dio, ez arduratu, hori Txinan martiri izateko edukiko dute.

Unamunok pentsatu omen zuen jesuitak gaizki hartuko zuela hura, baina hara non grazia egin eta barreka hasi zen.

Ingeniari katalan hura Pompeu Fabra i Poch zen, kataluniera estandarraren aita**.** Ingeniaria zen eta baita filologoa ere. Berak esaten omen zuen uste zuela bere ingeniaritzako formazioak ez ziola batere kalterik egin Filologian gauzak planteatzeko garaian.

1906an, Katalunieraren Nazioarteko Kongresua egin zenean, artean Fabra Bilbon Kimikako katedradun zen, baina kongresuan izan ziren hiru mila partehartzaileen arteko gogotsuenetako bat izan zen. Ordurako urteak zeramatzan katalana aztertzen. L'Avenç izeneko taldearen barruan katalanaren ortografia berritzeko kanpaina bultzatu zuen 1890-1891n eta 1904an *Tractat de ortografia catalana* argitaratu zuen Jaume Massó eta Joaquim Casas-ekin batera. 1912an *Gramatica de la llengua catalana* plazaratu zuen, aurreko urtean Bilbo utzi eta Bartzelonara aldatu ostean, hango Diputazioak sortu zuen katalunieraren katedradun eta sortu berria zen Institut d'Estudis Catalans-eko Filologia saileko kide gisa. Beranduago Instituteko buru izatera iritsi zen.

Institutek Ortografia-arauak 1913an ezarri zituen, 1917an Ortografia Hiztegia argitaratu zuen eta katalanaren gramatika ofiziala 1918an.

Hala ere, bere lanik ospetsuena eta mardulena, *Diccionari general de la llengua catalana,* Instituten hiztegi ofiziala bihurtu zena, 1932an atera zuen.

Euskaldunok lan horietan hasteko Euskaltzaindia sortu genuen garai berean, haiek oinarrizko lan guztia egina zuten, gidari eta ardatz Pompeu Fabra zutela.

Euskaltzaindiaren lau fundatzaileen artean, berriz, hor genuen Luis Eleizalde, ingeniaria ez bazen ere, Zientzietan lizentziaduna eta Matematikako katedradun izan zena.

Inazio Maria Etxaide (1884-1962) ingeniari donostiarra, Gipuzkoako telefono-sareko zuzendari eta telefonia automatikoa ezartzen aitzindari izan zena, euskaltzaina izan zen 1942 ezkeroztik eta euskaltzainburua 1952-1962 bitartean. Hainbat lan egin zituen euskal hitz eratorri eta elkartuez, aditzaz eta euskararen jatorriaz.

Louis Dassance (1888-1976) Laborantzako ingeniari ospetsua, *Gure Herria* aldizkariaren sortzaileetakoa eta bertako administratzaile izan zen, «Ikas» elkarte pedagogikoko ohorezko lehendakari eta euskaltzain oso 1949tik aurrera.

Pedro de Yrizar, (1910-2004) Industria- eta Geografia-ko Ingeniaritzan doktore zenak, euskal aditzari, euskararen atlasari, Bonaparte printzeari, euskararen eta Kaukaso aldeko hizkuntzen kidetasunei eta euskalkiei buruzko hainbat lan egin zituen eta euskaltzain urgazle eta gero ohorezko izan zen.

Gaur bertan, nire aurretik izendatutako hiru euskaltzain oso dira ingeniariak: Jean-Louis Davant, Laborantzako ingeniaritzan doktorea, batetik eta Ibon Sarasola eta Mikel Zalbide industria-ingeniariak bestetik.

Gaurko euskaltzain urgazleen artean hor ditugu Txillardegi, Martzel Ensunza, Joan Mari Irigoien, Ramuntxo Camblong, Manuel Ruiz Urrestarazu, Xabier Artola eta Jose Ramon Etxebarria, denak arlo bateko edo besteko ingeniariak. Eta badira beste batzuk ingeniariak izan gabe zientzietako lizentziadunak eta doktoreak direnak

Ezin dut esan, beraz, bide-erakuslerik izan ez dudanez. Horrek, jakina, haien eredua jarraitzeko ahaleginak egingo ditudala agintzera behartzen nau eta halaxe agintzen dut.

Hizkuntza-arazoek hainbat ingeniariri sortu dioten kezka eta eragin dioten lan egiteko bultzada ikusi ondoren, nire kezka eta interesgune batzuen berri eman nahi nizueke.

Aurreko belaunaldietako nire lanbidekideek izan ez duten aukera bat dut gaur egun: hizkuntza eta teknologia uztar ditzaket; bereziki hizkuntza eta informazioaren teknologiak. Ez daukat bata lantzeko bestea baztertu beharrik,

aitzitik, esan genezake bi arloak hertsiki lotuta aurkitzen ditugula hainbat esparru teoriko eta praktikotan.

Ikus dezagun begirada laster batean zein diren hizkuntzaren eta informatikaren arteko topaguneak:

Informazio-teknologiak eta hizkuntza uztartzen dituzten hiru lan-arlo zabal dira Hizkuntzalaritza Konputazionala, Hizkuntza Naturalaren Prozesamendua eta Hizkuntzaren-Ingeniaritza. Horretan bat datoz, baina definizio zabal horren barruan gauzak zehazten hasi orduko, zalantzak hasten dira. Hori beste esparru batzuetan ere sarri gertatzen da.

Ez da inoiz erraza izaten hurbileko eremuetako jakintzagaien arteko mugak zehaztea, eta, are gutxiago, horietan lan egiten dutenak tankera horretako auzietan ados jartzea. Hori jakinik abiatzen gara bereizketak egiteko ahalegin honetan eta, gainera, egia da arlo horietan egiten diren lanek bilbe tapituegia osatzen dutela, sinplekeriatan jausi gabe, behin betiko sailkapenak egiten hasteko. Nolanahi ere, gauzak ulertzen lagunduko duelakoan, termino bakoitzaren sintagma-burura joko dugu, haren esanahiaren oinarriaren bila:

Hizkuntzalaritza konputazionala: Hizkuntzalaritza arloko jakintzagaia da, sistema informatikoen bidez hizkuntza aztertzeaz arduratzen dena.

Hizkuntza Naturalaren **Prozesamendua:** Informatika arloko jakintzagaia da, hizkuntza naturala automatikoki sortzeaz eta ulertzeaz arduratzen dena.

Hizkuntzaren **ingeniaritza**: ingeniaritza den heinean, erabilera praktikoa duten sistemak eraikitzeaz arduratzen da eta kasu honetan hizkuntza da sistema horiek duten langaia.

Jakina, ez dago erabilera praktikoa izango duten sistemak garatzerik, hizkuntza naturala prozesatzeko tresna egokirik gabe, hauek ez dago lortzerik hizkuntza aztertu gabe eta makinek ulertuko dituzten hizkuntza-ereduak landu gabe, baina, bide batez, hizkuntzalariek tresna eraginkorrak behar dituzte euren ikerlanak burutzeko, etab., etab. Guztia guztiari lotuta dago.

Gaur egun beste jakintzagai askotan gertatzen den bezala, hauetan ere ezinbestekoa da hainbat arlotan trebatutako profesionalak lankidetzan jardutea: hizkuntzalariak, informatikariak, fisikariak, elektronikan adituak, psikologoak, matematikariak, etab.

Agian, jarduerak zehazki zedarritzen saiatzea baino egokiagoa izan liteke hizkuntzalaritza teoriko hutsetik hizkuntzaren industriek ekoizten dituzten erabiltzaile arruntentzako produktuetara bitartean etenik gabeko langintza-kate bat badela esatea.

Teorikoenetatik praktikoenetara doan bide horretan, muturrik teorikoenean, Hizkuntzalaritza Konputazional Teorikoa dugu.

Jakintzagai hori Hizkuntzalaritzaren adar bat da, hizkuntzaren funtzionamenduari buruzko hipotesiak frogatzeko informatikak sortutako tresnez baliatzen dena, baina baita Psikologiaren adarra den Zientzia Kognitiboaz ere, gizakiok hizkuntza nola erabiltzen dugun aztertzea eta portaera horren eredu informatikoak sortzea beharrezkoa duelako.

Azken urteotan, ikerketa horiek hartu duten konplexutasunaren ildotik Psikolinguistika Konputazionala ere sortu da. Jakintzagai horrek hitzak, hitz konposatuak, perpausak, etab. ikusiz eta entzunez ulertzeko edota ahoz eta idatziz sortzeko gizakiok erabiltzen ditugun prozesuen eredu informatikoak lantzen ditu.

Teoria hutsaren eta tresna praktikoen artean Hizkuntza Naturalaren Prozesamendua dago.

Hizkuntza naturalezko informazioa sistema informatikoen bidez ulertzea eta ekoiztea da jakintza-arlo horren egitekoa.

Esan gabe doa oso arazo korapilatsuak aurkitzen dituztela jarduera horietan ari direnek. Izan ere, hizkuntza aztergai eta langai bihurria da makinentzat

1950eko hamarkadan hasi ziren, Amerikako Estatu Batuetan, ordenagailuak itzulpen-lanetan erabiltzen, bereziki errusierazko zientzia-aldizkariak ingelesera itzultzeko asmoz, pentsatuz makina horiek oso bizkorrak direla kalkuluak egiten, eta antzeko trebezia erakutsiko zutela hizkuntza prozesatzeko ere. Uste zuten algoritmo egokiak eta haiek exekutatuko zituzten programa informatikoak garatzea zela kontua, Fisikako fenomenoak aztertzeko egiten denaren antzera, baina ustea erdia ustel gertatu zitzaien. Lauzpabost urteko epean itzulpen automatikoa menperatuta edukiko zutela iragarri zuten adituek, baina agudo hasi ziren zailtasunak agertzen, porrota ez esateagatik. Besteak beste, hiztegi-mailako anbiguotasunei aurpegi eman ezinda aurkitu ziren berehala. Anbiguotasun-arazoak konpondu nahian gerora hainbat ikerketa-lerro garatu dira, hain zuzen: kategoria gramatikalak etiketatzea, hitzen adierak desanbiguatzea, desanbiguazio sintaktikoa, etab.

1966an aitortu zuten hamar urteko ikerketak ez zituela aurreikusitako emaitzak lortu, eta aurreko hamarkadan ikerketa-arlo horretarako diru ugari izan zen bezala, iturria itxi zen eta lehengo ahoberokeriak ezkortasun bihurtu ziren. Hizkuntza aztertzea eta fenomeno fisikoak aztertzea ez baitira hain antzekoak.

Hizkuntzalaritzak ikertzen dituen fenomenoak askoz konplikatuagoak dira. Ikusi besterik ez dago zenbat teoria desberdin agertzen diren hizkuntzalaritzan eta nola ez den askotan bat gailentzen: Saussure-ren estrukturalismoa, Amerikako estrukturalismoa (Bloomfield, etab.ena), Chomsky-ren gramatika sortzailea eta gramatika transformatzailea, testuingururik gabeko gramatikak, Mel'čuk-en eta kideen esanahi/testu teoria (Meaning <->Text Theory edo MTT), besteak beste.

Bidenabar, esango dizuet zientzia zehatzen eta teknologiaren esparruetan hezitakooi kosta egiten zaigula batzuetan lur mugikor horietan ibiltzea, ez gaudelako horrelako ikuspegi-aniztasunarekin lan egiten ohituta.

1970eko hamarkadan Adimen Artifizialak esperantza handiak ekarri zituen eta aurrerapen batzuk bai, baina iraultzarik ez. 1980ko hamarkadan, makina ahaltsuago eta aldi berean merkeagoen etorrerak metodo estatistikoetan oinarritutako itzulpen automatikoari zabaldu zizkion ateak. Bide horretatik ere ez da mirariric etorri, baina bai itzultzaileen lana arintzen duten tresnak. Arlo jakin batzuetako eta hizkuntza jakin batzuen arteko itzulpen teknikoak egiteko oso lagungarriak dira. Murriztapen horietatik pixka bat urrundu orduko, ordea, irristaka hasten dira. Alegia, tresna elektroniko baten eskuliburua itzultzea egingarria den bezala, eleberri bat ondo itzultzea guztiz ezinezkoa da gaur-gaurkoz.

Metodo probabilistikoek duten abantailetako bat ikaskuntza automatikoaz baliatzeko aukera da.

Ikaskuntza automatikoa adimen artifizialaren adar bat da, ordenagailuei «ikastea» ahalbidetzen dieten algoritmoak eta teknikak garatzeaz arduratzen dena, oso datu-kopuru handiak erabiliz. Geroxeago ikusiko dugun bezala, horrek esan nahi du testu-corpus handiak erabiliz.

Esaldi batzuk eta are orrialde batzuk itzultzeko gai diren «jostailuzko» sistemak asmatzea ez da hain zaila. Askok izan dute bide horretatik abiatuta zerbait lortuko zuten itxaropenarekin lan egiteko tentazioa, eta baita hemendik oso urrutira joan gabe ere, baina, lehentxeago edo geroxeago, ez aurrera eta ez atzera gelditzen dira.

2007ko maiatzaren 1ean Berria egunkarian ederki islatuta agertu zen gai hau. Eusko Jaurlaritzak euskarazko itzultzaile automatikoa garatzeko prozesua abian jarri zuela eta, horren garapenean jardungo zuen enpresa hautatzeko lehiaketa publikoan parte hartuko zuten enpresetako adituei iritzia eskatzen zitzaien.

Zalantzan jartzen zuten haiek euskarazko itzultzaile automatiko orokor batek eskain dezakeen zehaztasuna. Hogei urte arlo horretan diharduten aditu euskaldunek zioten gaur egun ez dela posible edozein testuren itzulpen zehatzak egiteko sistema bat lortzea, baina posible litzatekeela arlo jakinetarako, arlo horietako testu corpus handiak egongo balira.

Aditu horietako batek zioen Jaurlaritzak euskarazko itzultzaile automatikoa garatzeko abiatu duen prozesua, baliatu beharreko aukera polita dela, ez jostailu bat sortzeko, ez dagoen azpiegitura bat sortzeko baizik. Nire ustez erdiz erdi asmatu zuen esan zuenean azpiegitura eta sarea direla beharrezkoak une honetan, euskararen eta euskararentzako teknologiaren ikerketan diharduten arteko lankidetza bilatu behar dela eta corpusak biltzeak eta lantzeak lehentasuna dutela. Ni ere bat nator iritzi horrekin eta uste dut Euskaltzaindiak baduela zereginik helburu hori lortzeko. Geroxeago itzuliko naiz puntu honetara.

Hizkuntzalaritza aldetik oinarri sendoak behar ditu hizkuntza naturalaren prozesamenduko sistema batek, itzulpen automatikoa baino askoz arazo sinpleagoak ere arrakastaz gainditzeko.

Arlo horretako mende-erdiko eskarmentuak zenbait irakaspen ekarri digu, hala nola:

– Hizkuntza Naturalaren Prozesamendua ez da gramatika- eta semantika--kontua bakarrik; gizakiok dugun munduaren ezagutza gabe ezinezkoa gertatzen da.

– Hizkuntzaren maila guztietako arazoak gainditu behar dira: fonologikoak, morfologikoak, lexikalak, sintaktikoak, semantikoak, pragmatikoak, testu-gramatikakoak, etab.

– Oinarrizko teoria sendorik gabe ezin da hizkuntza naturalaren prozesamendurik serioski garatu eta hau gabe ezin dira tresna praktiko eraginkorrak lortu. Arlo horiek eraikin baten egituraren antzera antolatu behar dira: teoria zimenduetan, prozesamenduko tresnak gainean eta erabiltzaile arruntentzako produktuak gailurrean. Hemen ere ez da bidezkoa etxea teilatutik hastea.

Zailtasunak zailtasun, ordea, hainbat tresna baliagarri presta daitezke arazo horiek gainditzen joan ahala.

Merkatuan arlo horretako hainbat produktu aurkitzen dira, esate baterako:

– itzulpen automatikorako lagungarriak
– zuzentzaile ortografiko eta gramatikalak
– informazio-bilatzaile aurreratuak
– testuen laburpenak egiteko tresnak
– testuak automatikoki sortzeko tresnak, adibidez eguraldi-iragarpenak egitekoak
– hizketa sintetizatzen eta ezagutzen duten sistemak
– ahozko hizketa testu idatzia bihurtzeko tresnak eta alderantzizkoak
– eskaneatutako testua ezagutzeko tresnak (OCR)
– eskuzko idazketa ezagutzen duten sistemak

Hainbat tresna berezitu daude hizkuntzalarientzat ere:

– Analizatzaile morfologikoak
– Analizatzaile sintaktikoak
– Kategoria gramatikalen etiketatzaileak
– Izen propioen ezagutzaileak
– Konkordantziak egiteko programak
– Azterketa estatistikoak egiteko programak, etab.

Une honetan Hizkuntza Naturalaren Prozesamenduan egiten ari diren aurrerapenek mehatxu eta aukera berriak dakarzkigute euskaldunoi. Alderdi bioi buruzko gogoeta egin nahi nuke orain zuekin batera.

Has gaitezen mehatxuak zein diren ikusten. Oso tresna baliagarriak daude dagoeneko, hedadura handiko hizkuntzentzat behintzat edo, hobeto esan, ia hedadura handiko hizkuntzentzat bakarrik.

Horixe da, hain zuzen, nire kezkarik handienetako bat. Informazio eta Komunikazio Teknologiek azken hamarkadetan ekarri dizkiguten baliabide gehien-gehienak ederki baliatu ditugu euskaldunok: testu-prozesadoreak, kalkulu-orriak, datu baseak kudeatzeko sistemak, irudiaren eta soinuaren tratamendu digitala, etab. Eduki digitalekin lan egiten hasi aurretik oso gogaikarriak, garestiak edo ezinezkoak ziren lanak arras erosoak, merkeak eta bideragarriak bihurtu zaizkigu. Digitalizazioa benetako bedeinkazioa izan dugu orain arte.

Kontuz!, ordea, hemendik aurrera; hizkuntzaren prozesamenduko tresna asko hizkuntza jakinetarako bakarrik baitira. Zenbait adibidek argituko digu esan nahi dudana.

Gaur egun badira programa batzuk eskuz idatzitakoa testu editagarria bihurtzen dutenak, esate baterako. Arkatz berezi batez erabiltzaileak idazten duena digitalizatzen dute eta teklatuaz idatzitako testua balitz bezala gordetzen dute ordenagailuan. Programa informatiko batek arkatzaren muturrak jarraitzen duen ibilbidea interpretatzen du.

Sistema horiek interesa dute, esate baterako sendagileentzat: gaixoekin dituzten topaketetan ordenagailu eramangarri baten pantailan eskuz idazten dituzte jasotzen dituzten datuak eta programak fitxa osatzen du.

Horrela esanda, badirudi ez duela horrek medikuak erabiltzen duen hizkuntzarekin zerikusirik, baina zoritxarrez ez da horrela. Sistema informatikoak ez du arkatzaren ibilbidea besterik gabe aztertzen; gainera hizkuntzaren prozesamendua ere erabiltzen du, idatzitakoa interpretatzeko. Edozein tresnak gutxienez hiztegi bat behar du eta euskararen kasuan, oso izaera eranskaria duelako gure hizkuntzak, analizatzaile morfologikoa ere bai. Sistema hizkuntza bakoitzerako prestatu behar da. Horren ondorioa da merkatuan hizkuntza jakinetarako bakarrik aurkitzen direla sistema horiek.

Medikuntza-arloan beste adibide batzuk aipa nitzake, esate baterako txostenen transkripzioarekin gertatzen ari dena. Garai batean, sendagileak magnetofono batean grabatzen zuen txostenaren edukia eta idazkari batek transkribatzen zuen. Gaur egun, ahozko azalpenaren grabazio digitala jaso ostean, hizketa ezagutzeko programa batek grabazioa testu editagarria bihurtzen du zuzenean eta idazkariak edo sendagileak berak aski du programak egin dituen interpretazio-akatsak zuzentzea, behar izanez gero. Sistema berria erabilita, produktibitatea %30 handitzen da eta batez besteko txosten bakoitzeko 1,5 € aurrezten dira eta denbora %50-%75 jaisten da. Hori guztia ingelesez, alemanez, suedieraz, danieraz, nederlanderaz, suomieraz, frantsesez, norvegieraz, portugesez, espainieraz, katalanez, italieraz, hungarieraz edo grezieraz egiteko, dagoeneko badira tresnak merkatuan. Euskaraz gaur-gaurkoz ez dago aukera hori.

Interneten edo dokumentu-biltegietan egiten diren bilaketak ere aldatzera doaz.

Jakina denez, orain amaraunean bilaketak egiten direnean, esate baterako, hitzak eta esaldiak karaktere-katea gisa tratatzen dira eta ez haiek duten esanahiaz baliatuz. Web 3.0 edo web semantikoa deritzonean, berriz, erabiltzaileek esanahian oinarritutako interakzioak izango dituzte ordenagailuekin. Orain, edozein bilaketa egiten dugunean, sekulako emaitza-mordoa lortzen dugu, baina zerrendan agertzen diren lehenbizikoetako batzuk baizik ez ditugu benetan ikusten eta ez dakigu interesatzen zitzaigunen bat ez ote zegoen askoz beherago ezkutatuta.

Orain web orriak egiteko nagusiki erabiltzen den HTML lengoaiaren bidez, orriaren itxura besterik ez da zehazten, baina web semantikoan datuen izaera deskribatzen duen teknologia bat erabiliko da. Makina gai izango da zenbaki bat posta-kodea, fakturaren zenbatekoa edo telefonoa den bereizteko, adibidez.

Web semantikoak, partez behintzat, irisgarriago bihurtuko du informazio hori guztia eta gainera egiten den kontsultaren testuinguruan esanguratsua dena eskainiko du.

Enpresek eta erakundeek beren dokumentu-biltegietan duten jakintza ustiatzeko askoz baliabide aberatsagoak izango dituzte etorkizun hurbilean, dokumentu horietako eduki semantikoa bereizteko aukera izaten dutenean. Hizkuntzaren eduki semantikoa automatikoki tratatzeko, ordea, hizkuntza bakoitzerako eta lan-arlo bakoitzerako prestatzen ari dira tresna informatikoak.

Terminologiaren estandarizazioa garapen horretan funtsezkoa da. Datu-base batean gaur egun «beheko masailezur» terminoa baldin badago erregistro batean, ez dugu aurkituko kontsultan «behe-baraila» erabiltzen badugu. Web semantikoan ez da hori gertatuko, sistemak ulertuko du egiten zaion galdera eta erabiltzaileak ez du galduko informazio hori, baina jakina sinonimia horren berri eman egin behar zaio sistemari.

Teknologia horiek erabili ahal izateko, ez da izugarrizko aldaketarik egin beharko, ez datu-iturrietan, ez tresnerian, baina bai programetan eta datuak etiketatzean, metadatuak ezartzean.

Langile batzuek trebakuntza berezia beharko dute eta ikasi beharko dute informazio-atalen arteko harremanez arduratzen. Ontologia-garatzaileak beharko dira, hau da lan-arlo jakinetako kontzeptuei dagozkien terminoak eta haien arteko erlazioak landuko dituztenak.

Jakina, terminoak hizkuntza jakinetakoak izaten dira. Nor arduratuko da euskarazko ontologiak lantzeaz makina erremintaren munduan, aseguruenean, finantza-zerbitzuenean, herri-administrazioarenean eta beste arloetan? Web semantikorako ez badugu euskara prestatzen, gizarte digitalaren Harri Aroan geratuko gara.

Antzeko zerbait gertatzen da beste hainbat arlo eta aplikaziotan ere. Esate baterako, biltegietan erabiltzen den «ahotsaren bidezko pikina» deritzon sisteman. Langileak biltegi bateko hainbat apaletan dauden produktuak bildu behar ditu, bezero baten eskaria osatzeko. Entzungailuak eta mikrofonoa ditu langileak. Biltegiko ordenagailuak irrati bidez galderak egiten eta aginduak ematen dizkio langileari eta honek esanez sartzen dizkio datuak ordenagailuari, mikrofonoaren bidez. Makinaren eta langilearen elkarrizketa gertatzen da. Hor programa informatiko bat ordenagailuko informazio idatzia entzungarri bihurtzen ari da eta langilearen ahotsezkoa idatzizko. Hori ere hizkuntza jakinetarako bakarrik aurkitzen da merkatuan.

Zer gertatuko da teknologia horiek hizkuntza batzuetan bakarrik erabil badaitezke? Jokoz kanpo geratzen diren hizkuntzak ez dira praktikoak gertatuko eta hainbat esparrutatik kanpora geratuko dira.

Aipatu ditudan tresna horiek ez dira ez jostailuak eta ezta bitxikeriak edo zientzi fikziozko filmetako kontuak ere. Informatikaren eta hizkuntzaren ezkontzak eragin handia du gaur egungo gizartean eta askoz handiagoa izango du laster.

Pixkanaka, hizkuntza era batera edo bestera prozesatzen duen softwarea sartzen ari da ordenagailu pertsonaletan, telebistan, sakelako telefonoetan, bideojokoetan, etab.

Gizarte berri honek jakintza du garapenaren gakoa. Orain arte ez bezala, iraultza ekonomikoaren erroak ez daude materiaren eta energiaren erabileran soilik. Informazio-teknologien etorrerak garapenaren eragilea beste arlo batera eraman du. Materia erabili beharrean, ikurrak, kodeak, mezuak erabiltzen dituzte teknologia hauek eta substantzia ukiezin hau da gizartea mugitzen duen motorraren erregai berria.

Aldaketa hauek ez dute ekoizpenean eta ekonomian bakarrik eragina, bizitzaren arlo guztietan baizik, eta hizkuntzan bereziki, hizkuntza delako, hain zuzen, gehien-gehienetan, informazioaren euskarria. Informazioaren tratamendua eta jakintzaren transmisioa gero eta gehiago daude lotuta hizkuntzaren tratamendu informatikoarekin. Horrela bada, ekonomiak, informazioak, jakintzak eta hizkuntzak lotura estuak dituzte.

Mehatxuak nondik etor dakizkigukeen aipatu ondoren, abaguneez ere hitz egin nahi nuke, baina horren aurretik, bereziki aipatu nahi nuke arlo horretan corpusek duten garrantzia.

Ekoitzi izan diren testu-multzo handiak ikertuz aztertzen du hizkuntza Corpus-hizkuntzalaritzak. Azterketa horien bitartez, hizkuntza gobernatzen duten lege abstraktuak aurkitzen saiatzen da. Antzina arakatze-lana eskuz egin ohi zen, baina gaur egun lan hori modu automatikoan edo erdi-automatikoan egiten da gehienbat eta gero zuzendu egiten da.

Ordenagailuen ahalmena handitu eta haiek merkatu ahala, corpusak hizkuntzaren ikerketan erabiltzea guztiz hedatu da.

Corpusei esker, hizkuntzalariak datu enpirikotan oinarritu dezake bere lana eta ez aurreiritzitan edo hizkuntzaz duen ezagutza nahitanahiez mugatuan.

Corpusak erabiltzeak abantaila handiak ditu hizkuntzalariarentzat. Esate baterako, hitz jakin baten adierak eta erabilerak aztertu nahi baditu, erraz bilatuko ditu hitz horren agerpen guztiak eta ondoz ondoko lerro banatan jarri, ezkerretik eta eskuinetik dituzten testuinguruekin. Eskuineko testuinguru guztiak alfabetoaren araberako hurrenkeran jar ditzake, hitz horren kokapenak aztertzeko.

Gaur egun, Hizkuntzalaritza ez ezik, Hizkuntzaren Ingeniaritza ere corpusetan oinarritzen da.

Esan dezakegu, hortaz, corpusak eraikitzea beharrezkoa dela, bai hizkuntza aztertzeko eta baita hizkuntzaren ingeniaritzako produktuak garatzeko ere.

Hainbat corpus-mota daude, baina estrategikoki funtsezkoena erreferentzia-corpusa da.

Hizkuntza bati buruzko ahalik eta informaziorik osatuena emateko prestatuta dagoen corpusari erreferentzia-corpusa esaten zaio. Hizkuntzaren ahalik eta aldaera gehienen berri emateko, behar den adinako tamaina izan behar du. Tamaina garrantzitsua da, baina, gainera, eredu egoki baten arabera hautatutako testuak eduki behar ditu erabileremu, genero, erregistro eta gainerako bereizgarriak kontuan hartu eta orekatua izateko. Hartara, erabil daiteke gramatikak, hiztegiak, tesaurusak eta beste hainbat tresna lantzeko.

Erreferentzia-corpusen xedeak hainbat izan daitezke eta izan beharko lukete, bai hizkuntzalarientzat eta baita hizkuntza-ingeniarientzat ere. Tamainaz oso handiak izaten dira, 50 milioi testu-hitzetik gorakoak, behintzat, eta 100 milioikoak badira, hobe.

Euskaltzaindiaren historian egin den gauzarik onenetako bat corpusak eratzeko erabakia izan da, zalantzarik gabe. Orotariko Euskal Hiztegiaren corpusa eta EEBSrena garrantzi handiko bi altxor dira. Ederki ikusten ari gara hori Hiztegi Batuaren lanetan murgilduta gabiltzanok.

Hala ere, horiek ez dira erreferentzia-corpusak, tamainagatik, hartzen duten epeagatik, duten lantze-mailagatik. Antzeko zerbait esan daiteke ereduzko prosaren biltegiaz. Elhuyar-en Zientzia eta Teknologiaren corpusa, berriz, eredugarri izan daiteke erabili diren metodo eta tresnen aldetik eta baita, neurri batean behintzat, lortutako lantze-mailagatik, baina esparru mugatukoa da. Corpus berezituen sailean kokatzen da.

Euskarak erreferentzia-corpusa beharrezkoa du. Hori, jakina, ez da testuak euskarri digitalean biltzea soilik.

Corpus bati benetako etekina ateratzeko, landu egin behar da, oharreztatuz. Hizkuntzaren maila guztietan jar daitezke ohar horiek: maila lexikalean, sintaktikoan, semantikoan, pragmatikoan, diskurtsuarenean. Horrekin, testuetan dagoen informazio inplizitua informazio esplizitua bihurtzen da eta horrela lasterrago eta aiseago aurkitzen eta aztertzen du informazioa bai gizakiak eta baita makinak ere.

Horrelako corpus bat prestatzea ez da txantxetako lana. Behar dugu, baina ez dugu berehala eskueran izango. Bitartean ez dugu, ordea, zertan geldi egon beharrik eta ez gaude geldi. Euskaltzaindiak badu dagoeneko bide hori urratzen joateko egitasmoaren lehen zirriborroa. Bertan corpus monitore bat eraikitzen joatea planteatzen da, alde batetik euskara idatzizko hedabidee-

tan ia unean-unean nola erabiltzen ari den jakiten lagunduko duena eta bestetik gorputza hartzen joango dena erreferentzia corpuseranzko bidean.

Abaguneei dagokienez, nire ustez, ikuspegi estrategiko zabal baten barruan bilatu behar dira. Estrategia hori marrazten hasteko, hasierako baldintzak izan behar ditugu gogoan eta nire ustez hauek dira:

1. Ez gara hutsetik abiatzen eta gutxiagorik ere. Baditugu euskal munduan beharrezkoak ditugun hainbat arlotan eskarmentua duten pertsonak eta erakundeak. Hizkuntza Naturalaren Prozesamenduan, esate baterako, ikerketa-proiektu asko garatu dira, hainbat doktoretza-tesi ere bai, kongresuetan lanak aurkeztu dira, tresnek praktikan erakutsi dute sendoak eta eraginkorrak direla.

2. Aurreko baldintza aldekoa dugun bezala, eta oso aldekoa gainera, orain aipatuko dudana ahulezia bat da: ez dugu lortu oraindik ustiapen industrial eta komertzial errentagarririk Hizkuntzaren Ingeniaritzan. Gehiago esango dut: zaila ikusten dut hori lortzea planteamendu batzuk ez baditugu aldatzen, baina zorionez uste dut gure eskuetan dagoela aldaketa hori eragitea. Geroxeago zehaztuko dut hau gehiago.

3. Euskara erabilera murritzeko hizkuntza izatea oztopo larria da. Honek, jakina, asko baldintzatzen du aurreko puntuan aipatu dudan errentagarritasunaren arazoa.

4. Mundu-mailan Hizkuntzaren Ingeniaritzako produktuen eta zerbitzuen merkatua interes handia sortzen ari da eta aplikazio-eremu berri asko ari dira agertzen.

5. Euskal Herrian euskaraz aparte, hedadura handiko bi hizkuntza oso ondo menperatzen ditugu: espainiera eta frantsesa.

Baldintza hauek kontuan izanda, esaldi batean laburbiltzen ahaleginduko naiz estrategia:

Dagoeneko metatua dugun jakintzan eta praktikan oinarrituz, Hizkuntzaren Ingeniaritzako produktuen eta zerbitzuen industria bultzatzea, hedadura handiko hizkuntzen merkatuetan errentagarritasuna bilatuz, bide batez euskarak datorren munduan bizirik irauteko beharrezkoak izango dituen tresnak landuz.

Mundu globalizatuan buru-belarri sartuta gaudelarik, jakintzan oinarritutako organizazioak behar ditugula esaten digute adituek. Hara hor bete-betean baldintza hori betetzen duen jarduera bat. Baliteke ni ameslari bat izatea, baina uste dut areto honetan bertan dauden dozenaerdi bat lagunek estrategia horretan sinetsiko balu, bideragarria litzatekeela.

Esker mila zuen arretagatik.