ASMOZ
on-line prestakuntza
fundazioa

HIZNET
Hizkuntza Plangintza Ikastaroa

17. Euskara eta teknologia berriak

Irakaslea: © Andoni Sagarna

asmoz fundazioa

Soziologia 2 saila

AURKIBIDEA

1. GAINBEGIRATU BAT GAIARI

Euskararen geroa informazioaren eta jakintzaren gizarteak dituen eta izango dituen ezaugarriek baldintzatuko dute. Gizarte berri honek jakintza du garapenaren gakoa.

Orain arte ez bezala, iraultza ekonomikoaren erroak ez daude materiaren eta energiaren erabileran. 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 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.

Kontuan izan, ordea, gauza bat: hizkuntza abstraktuki ez dela gizartean erabiltzen dena, hizkuntza jakinak direla erabiltzen edo baztertzen direnak. Hizkuntza jakin bat, demagun euskara, informazioaren eta jakintzaren sorkuntzan eta transmisioan erabiltzen den heinean, ekonomiaren eragile da eta, neurri honetan, bizitza eta etorkizuna bermatuak ditu, baina beste batek kentzen badio lekua eginkizun horietan, jokoz kanpo geratuko da. Azkenik, arrazoiketa biribiltzeko, esan dezagun hizkuntza bat ez daitekeela izan gizarte berrian informazioaren eta jakintzaren euskarri, hizkuntza hori ez badago lotuta informazio-teknologien erabilerarekin eta ez badaude garatuta hizkuntza horretan bideratzen diren informazioak prozesatzeko baliabide teknologikoak.

Gaur egun, bizitza arruntean, ez dugu oraindik hainbeste nabaritzen hizkuntzaren tratamendu informatikorako tresnen beharrik, baina, goian esandako arrazoiengatik, zalantzarik gabe, ohartu ere gabe, ezinbestekoak bihurtuz joango dira.

Hizkuntza mintzatuaren eta idatziaren tratamendu automatikoaren garapena, hizkuntzaren industriak deritzenena, alegia, garrantzi handikoa da hizkuntzaren, teknologiaren eta ekonomiaren ikuspegitik. Horregatik, hain zuzen, gai honi dagokionez, koordinatu beharrekoak dira hizkuntza-plangintza, plangintza ekonomikoa eta ikerketa-, garapen- eta berrikuntza-planak.

Arestian aipatu ditugun hizkuntzaren industriek ahozko nahiz idatzizko hizkuntzaren tratamendu informatikoan oinarritzen diren produktuak eta zerbitzuak eskaintzen dituzte. Hauek lantegietan, etxeetan, zerbitzu publikoetan, etab. erabiltzen diren tresnetan egoten dira: ordenagailuetan, makinetan, telefonoetan, etab.

Tresna hauek mota desberdinetakoak izan daitezke: testuak idazteko eta editatzeko lagungarriak, dokumentuen kudeaketa elektronikoa, datu-baseak kontsultatzeko bilatzaileak, itzultzeko lagungarriak, hizketa ezagutzeko eta sintetizatzeko tresnak, hiztegiak, eskaner bidez digitalizatzen diren testuetako karaktereak ezagutzeko softwareak (OCR), etab.

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

2. INFORMAZIOAREN ETA JAKINTZAREN GIZARTEA

Informazioaren gizarteak muga politiko eta geografikoak deuseztatu egiten ditu gauza askotarako. Adibidez, ETB ikusteko, ez da jada beharrezkoa Euskal Herrian edo inguru hurbilean egotea, Amerikan ere ikus dezakegu sateliteari esker. Bestetik, informazioaren kudeaketa bihurtzen du lehiakortasunaren gako, informazioa lantzeaz arduratzen diren lanbideei garrantzi berezia ematen die, informazioa eta honen ordenagailu bidezko tratamendua izugarri ugaritzen ditu eta informazioaren transmisio-denbora guztiz murrizten du.

Globalizazioak ondasunen diseinuan, ekoizpenean eta kontsumoan du eragina. Informazioaren kasuan ere horixe bera gertatzen da.

Informazio- eta komunikazio-teknologiek hobekuntza nabarmenak eragiten dituzte industriaren produktibitatean eta zerbitzuen kalitatean. Administrazio publiko guztiek eta enpresa askok, esate baterako, "desmaterializatzera" jotzen dute.

Tresna informatizatuak hedatzen ari dira bizitzaren arlo guztietan. Gaur egun bulego guztietan daude garai batean espezialistek bakarrik ukitzen zituzten tresnak, eta pixkanaka etxeetako egongela eta sukaldeetara iristen ari dira. Horren lekuko dira bideojokoak edo agenda elektronikoak. Hizkuntza prozesatzeko tresnak, berriz, badatoz pixkanaka, adibidez, zuzentzaile ortografikoak.

Aldaketa hauek enpleguan eragin nabarmena izaten ari dira. Gizarte aurreratuetan duela 20 urtez geroztik langileen erdiak baino gehiagok ez dituzte objektu fisikoak maneiatzen. Horren ordez, informazioa sortzen, eraldatzen, transmititzen eta hedatzen dihardute. Eta portzentaia hau handitzen ari da.

Lantokietako harremanak ere aldatzen ari dira, informazioaren nagusitasun hau dela eta. Hierarkiak hausten ari da informazioa behar duenak zuzenean, bitartekaririk gabe, behar duen unean bertan eta inora mugitu gabe eduki beharrak.

Gero eta gehiago, langileak lantokira, ikasleak eskolara edo gaixoak erietxera mugitu gabe burutzen dira lana, ikasketak eta osasun zerbitzuak. Telelana, teleikasketak eta telemedikuntza beren aurrerapenak egiten ari dira.

Lanean, gero eta gehiago, idatzizko hizkuntzaren bidez gauzatzen dira eginkizunak, dela paperean, dela ordenagailuko pantailan. Horregatik, informazio-kopuru erraldoiak ekoizten dira. Informazio zientifikoa, adibidez, esponentzialki hazten ari da. Interneten dauden informazio-kopuruak sinetsi ezinezkoak dira: Google bilatzailearen hasierako orriak dioskunez (2001eko urrian) bilaketak 1.610.476.000 web orritan egiten ditu. Informazio hau dena erabiltzea ezinezkoa litzateke, bestalde, tresna informatikorik gabe.

3. INFORMAZIOAREN GIZARTEAREN ERAGINA HIZKUNTZAN

Informazioan oinarritzen den ekonomiak hizkuntza idatzia menderatzen duten pertsona gehiago eskatzen ditu, hizkuntzaren kalitatea areagotu beharra ekartzen du eta hizkuntzazko informazioa ordenagailu bidez tratatzera behartzen du.

Ordenagailua etengabe erabiltzeak, denbora errealean erabakiak hartuz enpresak kudeatzeak, informazioa lehenbailehen transmititu beharrak gero eta langile gehiago informazioa ekoitzi, transmititu, metatu edo analizatzen jarri beharra dakar.

Hizkuntza prozesatzen duten tresnak erabiltzeak hizkuntzaren kalitatea areagotzea beharrezkoa egiten du, zeren ordenagailuak nekez ezagut dezake mezu bat morfologia, sintaxi edo fonologia okerrez eratua baldin bada.

Kalitate oneko testuak erabiltzea beharrezkoa da, ordenagailuz testuak era fidagarrian analizatzeko, edo itzulpen automatikoa errazteko. Testuak idazten dituenak gero eta lasterrago burutu behar du bere lana, eta batere bitartekaririk gabe —zuzentzailerik edo idazkaririk gabe— iritsarazi hartzaileei bere idazkiak. Ordenagailua da, hain zuzen, kalitate ona lortzen gero eta gehiago laguntzen diona, zuzentzaileen, hiztegien edo itzulpenerako laguntzen bidez.

Gero eta gehiago, harremanak hizkuntza desberdinetako jendearekin izaten dira bizitza ekonomikoan edo kulturan. Partez ingelesa "lingua franca" gisa erabiliz bideratzen dira harreman hauek, zientzialarien artekoak adibidez, baina sarri gertatzen da hizkuntza hau behar bezala ezagutzen ez duten pertsonak harremanetan jartzea edota egoerak beste hizkuntza baten erabilera eskatzea. Kasu hauetarako, informazio-teknologiak soluzioak eskaintzen ari dira.

Ordenagailuen ahalmena hazi den heinean, eta esponentzialki hazi ere, makinetan meta daitekeen eta denbora-unitateko prozesa daitekeen informazio-kopurua ere izugarri handitu da. Honi esker, gero eta hurbilago dago eguna hizkuntzaren aldibereko tratamendua egin ahal izango dena. Esan dezakegu hauxe dela, hain zuzen, informazio-teknologiek gaur egun duten erronkarik handienetako bat.

Tresna informatikoen erabilera erraztu beharrak ere hizkuntza naturalen (euskara, ingelesa, etab.) prozesamendua lantzea areagotzen du. Adibidez, teklatua tresna gogaikarria da informazioa sartzeko. Askoz hobe litzateke ordenagailuari aginduak ahoz ematea, noski, batez ere ezintasunen bat dutenentzat (itsuak, paraplegikoak, eskuak libre behar dituzten langileak, etab.) Hori aski aurreratua dagoen bidea da hizkuntza batzuentzat.

4. INFORMAZIOAREN GIZARTEA ESKATZEN ARI DENARI EZ ERANTZUTEAK DITUEN ONDORIOAK

Informazioaren gizartean murgiltzeak dituen ondorioak aztertzeko, hiru ardatz izan behar dira kontuan: teknologikoa, ekonomikoa eta soziala. Honen barnean kokatuko genituzke hizkuntzen erabilerari dagozkion arazoak.

Teknologia da informazio-gizartearen eragilea, eta era guztietako informazioaren digitalizazioa mugimendu honen giltzarria. Edozein eduki bi digitu erabiliz, 0 eta 1, kode daiteke, eta horrela kodeturiko edozer ordenagailuz tratatu. Informatizazioaren hasieran zenbakizko edukiak soilik kodetzen ziren kode bitar honen bidez. Zientzia eta teknologiako kalkuluak izan ziren kodeketa honen abantailak aprobetxatzen lehenak. Geroxeago kalkulu finantzarioa eta administratiboa. Orain, ordea, musika, bideoa, testuak, etab. ere digitalizatu egiten dira. Iraultza horren lekuko dira musikako CDak eta MP3 entzungailuak, argazki- eta bideo-kamera digitalak, CD-ROMean eskaintzen diren entziklopediak, bideojokoak, DVDan dauden filmak, posta elektronikoa, sakelako telefonoz transmititzen diren mezu laburrak eta laster transmitituko diren irudiak, eta satelite bidezko telebista. Gaur egun datuen tratamendu digitalak informazioaren prozesamendua eta telekomunikazioak integratu ditu.

Hizkuntzazko edozein eduki ere, dela idatzizkoa, dela ahozkoa, eta edozein hizkuntzatan gainera, trata daiteke teknologia hauek erabiliz. Ikuspegi teknikotik ez dago inolako ezintasunik. Alabaina, hizkuntza natural bakoitzak teknologia hauek erabiltzen dituzten produktu eta zerbitzuak berariaz garatzea eskatzen du. Hau, ordea, ez da gertatuko merkatuaren legeek bultzatuta soilik, gizartean abantailazko lekuan ez dauden hizkuntzen kasuan. Eleaniztasun-politika, egoera ahulean dauden hizkuntzen aldeko politika eta edonork (zaharrek, haurrek, ezinduek, hedadura handiko hizkuntzak ez dakizkitenek, etab.) teknologian oinarrituriko tresnak erabili ahal izatearen aldeko politika praktikatu beharko dira hizkuntzaren industrien arloan, desberdintasun sozialak handitzea eta gizarteko gatazkak larritzea ez bada nahi.

Euskaldunok antzeko egoera batean egonak gara lehen ere, inprimategiaren asmakuntzak ekarri zituen abantailetara mende bateko atzerapenarekin iritsi ginenean, hain zuzen. Ordukoa larria izan bazen, oraingoa askoz arriskutsuagoa da, gizarte-aldaketak orain duen lastertasunak eta sakontasunak askoz errazago bazter dezaketelako aldaketa bere alde jartzea lortzen ez duena.

Egunean jartzeko ahalegina garestia da, ordea, eta ezin izango dute egin aski baliabide ez dutenek. Arrisku galanta dago munduan milaka hizkuntza betiko desagertzeko, hautespen naturalaren legeak gupida gabe suntsitzen baitu aldaketara moldatzen ez dena.

Gaur egun, hizkuntzaren tratamendua, gehienbat, ingelesarentzat dago aurreratua, horixe baita munduko ekonomia eta teknologia menderatzen duten herrialdeetako hizkuntza. Azken finean, ekonomiak neurri handi batean mugatzen du hizkuntza batek bizirik irauteko duen aukera. Atzetik ekonomia sendoa ez duen hizkuntzak etorkizun iluna du, eta bizitza ekonomikoan komunikazio-tresna izateko prestatua ez dagoenak berdin. Bulego batean hizkuntza bat erabiltzeko erraztasun teknologiko guztiak baldin badaude eta beste bat erabiltzea gogaikarria baldin bada, lehenbizikoa erabiliko da eta ez bigarrena.

Euskararen prozesamendu automatikoa aurreratzen ez bada, bizitza ekonomikotik at eta bizirik irauteko aukerarik gabe geratuko da.

Neurri politiko egokiak hartu ezean, euskarak ez du garapen teknologikorik izango, merkatuko legeak ez dituelako alde, laneko hizkuntza, eta bizimodu modernokoa oro har, ez da izango eta etorkizunik gabe geldituko da. Zenbat eta beranduago hartu neurri hauek, orduan eta gainditzen zailagoa izango da atzerapena, ez denbora aurrerago joango delako bakarrik, baizik eta aurrera doazenen abiadura asko lastertuko duten faktoreak begi-bistan ditugulako, hala nola, ate joka dauden banda zabaleko komunikazioen eztanda.

5. ERRONKARI ERANTZUTEKO LANDU BEHARREKO ARLOAK

5.1. Gizakiak sortutako ahozko mezuen tratamendua

Gizakiak ahozko hizkuntzaz jaulkitako mezuen bidez makinei aginduak eta informazioa emateko sistemek irispide erosoa emango digute eguneroko bizimoduan hainbat informazio eta zerbitzutara. Erabiltzailearen hizkuntzan ahozko mezuak ulertzen dituen sistema da, zalantzarik gabe, natural, malgu eta eraginkorrena gizakiarentzat.

Teknologia eta aplikazio desberdinak behar dira makinak gizakiaren hizketa prozesa dezan.

Batzuetan ez da mezuaren edukia tratatu behar, baizik eta hiztuna edo honek egiten duen hizkuntza identifikatu behar da. Hiztunaren identifikazioa norbaiti leku batean edo zerbitzu batean (banku-zerbitzuetan, adibidez) sarbidea baimentzeko erabil daiteke, adibidez.

Gizakiak makinari hitz eginez aginduak eman ahal izateko, hizketa ezagutzeko sistema bat behar da, h.d. soinuzko informazioa testu bihurtzeko gai den programa bat. Horretarako behar den teknologia emaitza praktikoak eskaintzen hasia da. Badira, esate baterako, ahoz agindua emanez deia egiten duten telefonoak (*Deitu etxera,* adib.), datu sinpleak, hala nola kreditu-txartelen zenbakiak, ahoz sartzeko sistemak.

Hizketaren ezagutzan, mikrofono batek edo telefono batek jasotako soinuzko seinaleak prozesatu eta haiei dagozkien hitzak lortu behar ditu makinak. Hitz horiek izan daitezke makinarentzako aginduak, datu-base batean sartu beharreko datuak edo dokumentu batean sartu beharreko idatzizko hitzak, etab.

Maila desberdinetan garaturiko produktuak aurki daitezke arlo honetan. Badira hitz solteak ulertzeko gai diren sistemak, hiztunak hitzen artean pausaldiak egitea beharrezkoa dutenak eta badira hizketa jarraikia, pausaldi nabarmenik gabekoa, ere uler dezaketenak.

Arreta berezia jarri gabe mintzatzen garenean, idatzitakoa irakurtzen dugunean baino irregulartasun gehiago izaten ditu hizketak, eta nekezago uler dezake makinak. Hiztunak bere hizketaren erakusgarri batzuk ematea behar izaten dute sistema batzuek, lanean hasi aurretik, eta beste batzuek, aldiz, edonoren hizketarekin lan egiteko gai izaten dira.

Bestalde, erabilera-arloak ere badu zerikusirik. Zailtasunak handitu egiten dira oso hiztegi zabalarekin lan egin behar denean edota antz handia duten hitzak erabiltzen direnean.
Hizkuntzalaritzaren ikuspegitik, berriz, batzuetan hitz-multzo jakin batekin eta hitz hauen arteko sekuentzia jakinekin lan egiten da. Hauxe da, hain zuzen ere, izan litekeen sistemarik sinpleena. Hori baino sistema landuagoek ezagutza gramatikala erabiltzen dute.

Euskararen kasuan fonologiak eta fonetikak zenbait zailtasun berezi dakarte. Adibidez, *atso* hitzaren *ts* fonema hizketan gauzatzen dutenean hiztun batzuek *tz* balitz bezala ahoskatzen dute edo idatziz bi hitzetan ematen dugun *ez dakit* esakunea ahoskatzen dugunean *sandhia* egiten dugu, *eztakit* esaten dugu.

Azkenik, badira beste faktore batzuk hizketa ezagutzeko sistemen lana zailtzen dutenak, hala nola hizketaldia burutzen den lekuan egon daitekeen zarata edo mikrofonoaren ezaugarriak eta kokapena.

Gaur egun merkatuan badaude salgai hizketa ezagutzeko programak. Adibidez, IBM etxearen *ViaVoice* Windows-entzat, Macintosh-entzat eta Linux-entzat. Programa honen bidez hainbat eragiketa ahotsa erabiliz burutu daiteke: sistema eragileari edo aplikazioei aginduak eman, Office-ko eta Lotus-eko aplikazioei testuak diktatu, diktaketa hauetan erabiliko diren izen propio, termino tekniko, akronimo eta helbideak hiztegian sartu, sistemak ezagut ditzan, ahotsaren bidez Interneten nabigatu, bilaketak egin, ordenagailuan metaturik dauden testuetako edukia entzun. Sistemak ikasi egiten ditu erabiltzailearen estiloa eta sarri erabiltzen dituen hitzak, zehaztasun handiagoz funtzionatzeko.

Bikainak dira, ikus daitekeenez, programa honek eskaintzen dituen ezaugarriak, baina esan beharra dago, halaber, ez duela edozein hizkuntzatan funtzionatzen, eta funtzionatzen duen hizkuntza guztietan ere ez dituela ezaugarri berberak.

Bertsiorik berrienean, 9an, japonieraz, ingelesez eta Brasilgo portugaleraz bakarrik lan egiten du eta bertsio zaharragoan beste dozena erdi bat hizkuntzatan. Euskara ez da inondik ere agertzen.

Alabaina, ez daiteke esan arlo hau euskaraz landu gabe dagoenik. Errentaren gaineko zergaren aitorpena egiteko epean, Gipuzkoako Foru Aldundiak telefonia-sistema automatiko bat erabiltzen du hiritarrei euskaraz nahiz erdaraz hitzordua emateko.

5.2. Testuen ezagutza automatikoa

Idatzizko hizkuntzaren ezagutza automatikoa, h.d. idatzizko hizkuntza datu elektroniko bihurtzeko sistema, garrantzi handikoa da, bai inprimaturiko dokumentu, aldizkari eta liburuetako informazioaren kasuan eta baita eskuz idatzitakoarenean ere, giza komunikazioan idatzizko mezuen trukeak duen pisuagatik.

Gaur egun dokumentuak ordenagailu bidez prestatzen dira formatu digitalean eta gero inprimatu egiten dira. Formatu digitalean dauden dokumentuak erabiliz gero, bertako edukia erraz prozesa daiteke automatikoki, baina batzuetan paperezko dokumentua eskuratzen dugu eta honen bertsio digitala ez dugu eskura. Horrelakotan, paperezko dokumentuak pasa daitezke formatu digitalera eskanerraren bidez. Makina horrek dokumentuaren irudi digitala eskaintzen digu, ordea, ez testua testu bezala, karaktere-segida bezala, alegia.

Eskuz idatzitako testuak digitaliza daitezke, inprimaturikoak bezalaxe, eskanerraren bidez, edota gainazal elektroniko baten gainean arkatz berezi batez idatziz.

Makinak idatzizko hizkuntza ezagut dezan, irudi bat interpretatu behar du eta hark duen balio sinbolikoa atera. Adibidez, zirkulutxo baten balioa "o" letra edo zero zenbakia izan daiteke.

Interpretazio hau OCR (Optical Character Recognition) deritzen programek egiten dute inprimaturiko testuen kasuan eta ICR (Intelligent Character Recognition) deritzenek eskuz idatzirikoen kasuan.

Inprimaturiko testuen interpretazioa errazagoa da eskuz idatzitakoena baino eta eskuzkoetan errazagoak dira karaktereak bananduta idatzitakoak lotuta idatzitakoak baino.

Kasurik errazenean ere badira faktore batzuk interpretazioa zailtzen dutenak: inprimaketaren kalitatea batetik eta erabiltzen diren letra-tipoak (Times, Helvetica, etab.) bestetik.

Zailtasun hauek gainditzeko era bat, karaktereak ezagutzeko algoritmoak erabiltzeaz gain, karaktere-multzoek osatzen dituzten hitzak ezagutzen ahalegintzea da. Honek esan nahi du, hizkuntza bakoitzarentzat moldatu egin behar direla programak.

Ikus dezagun zer gertatzen den kasu praktiko batean: sakelako ordenagailuek eskuz pantailaren gainean arkatz berezi batez idazten dena testu bihurtzeko gaitasuna izaten dute, baina hizkuntza jakin batentzat, normalean ingelesarentzat. Alabaina, bada programa bat, Paragon PenReader, ordenagailu txiki hauek beste hizkuntza batzuk interpretatzeko gaitzen dituena. Hona hemen gaur egun interpreta ditzakeen hizkuntzak:

Albaniera	Islandiera
Alemana	Italiera
Bulgariera	Kataluniera
Daniera	Kroaziera
Errumaniera	Letoniera
Errusiera	Lituaniera
Eslovakiera	Nederlandera
Esloveniera	Norvegiera
Espainiera	Poloniera
Estoniera	Portugalera
Frantsesa	Suediera
Greziera	Suomiera
Hungariera	Turkiera
Ingelesa	Txekiera

Ikus daitekeen legez, euskara ez dago hizkuntza horien artean.

5.3. Hizkuntzaren analisia eta ikerkuntza

Hiztunok hitzak kateatuz esanahia duten perpausak eratzen ditugu eta perpausak kateatuz berbaldiak edo testuak. Eraikuntza hauek nola osatzen diren aztertzen saiatzen da hizkuntzalaritza eta funtzionamendu horren eredua da gramatika. Honen barnean berriz sintaxiak elementuen konbinatoria erakusten digu eta semantikak, berriz, esanahiaren osaera.

Hizkuntza-teknologiako aplikazioetan erabiltzen den gramatikak hiztegia, erregela sintaktikoak eta erregela semantikoak izaten ditu. Erregela-multzo handiak erabiltzen direnez, arazo serioa izaten da sistemak aski lastertasun izatea interaktibitatea edo testu-multzo handiak prozesatzea beharrezkoa den aplikazioetan.

Dena den, aplikazio konplexuetara egun batean iristeko, lehenbizi maila apalagoko arazoak gainditu behar dira, eta, horregatik, hizkuntza-unitate sinpleen analisia

menderatzea funtsezkoa da. Analisi-motarik sinpleena analisi morfologikoa da. Horregatixe, hain zuzen ere, ordenagailu bidezko analisi morfologikoak egin ditu hizkuntzalaritza konputazionalaren beste edozein adarrek baino aurrerapen handiagoak azken 15 urteotan. Hau halaxe da euskararen kasuan ere. Duela hamabost bat urte EHUko Donostiako Informatika Fakultateko IXA taldeak, epe luzera begira helburuak jarri zituenean, itzulpen automatikoa zuen ametsa, baina garbi ikusi zuen analizatzaile morfologiko batetik hasi beharra zeukala urratsez urrats amets handi horretara hurbiltzen joateko. Halaxe egin dute urte hauetan zehar, eta ikerketa-lerro horretatik irten dira gaur egun ditugun euskararako tresna informatikoak (Morfeus analizatzaile morfologikoa, Xuxen ortografia-zuzentzailea, Euslem lematizatzaile automatikoa).

5.4. Hizkuntza natural idatziaren sorkuntza

Ordenagailuan gordetako bestelako informaziotik abiatuz hizkuntza naturalezko testuak sortzeko gai diren programa informatikoak lantzen dira arlo honetan. Mailarik apalenean perpaus solteak sortzen dira eta jasoagoan testu osoak.

Eginkizun desberdinak dituzten sorgailuak egiten dira:

a) Testu- planifikatzaileak. Hauek informazio-multzo batetik irteeran eskainiko dena hautatzen dute eta testu-egitura bat sortzen dute
b) Perpaus- planifikatzaileak. Hauek maila lokalagoan jarduten dira: testuko perpaus bakoitzeko edukia antolatzen dute perpaus zuzenak eratzeko.
c) Azaleko egituraren mailan eratzen dituzte perpausak.

Arlo honetako produktu komertzialak egiten dituen enpresa bat EEBBetako *CoGenTex, Inc.* da. Etxe honek azkena aurkezturiko produktua *Project Reporter 2.0* izenekoa da. Hona hemen zertan datzan:

Gaur egun, leku desberdinetan lan egiten duten pertsonak proiektu berean ari badira lanean Internet erabiltzen dute komunikatzeko. Proiektuak kudeatzeko, berriz, merkatuan dauden horretarako programak erabiltzen dituzte, hala nola *Microsoft Project,* baina programa hauek ez dute izaten tresnarik, gordeta daukaten informazioa Webean jartzeko.

Arazo hau konpontzen du, hain zuzen, aipatzen ari garen programak, proiektuei dagozkien datuak web bidez argitaratzea ahalbidetzen baitie lantaldeei. Proiektuaren kontrolerako erabiltzen diren grafikoak ez ezik, proiektuaren egoerari buruzko txosten idatziak ere sortzen ditu, eta are erabiltzaile-talde desberdinei dagozkien txostenaren bertsio desberdinak ere.

Project Reporter-ek *Microsoft Project*-en bertsio hauek ditu bateragarriak: ingelesezkoa, alemanezkoa, frantsesezkoa, japonierazkoa, norvegierazkoa, Brasilgo portugalerazkoa, danierazkoa, italierazkoa, koreerazkoa, espainierazkoa, suedierazkoa eta hebraierazkoa. Euskarazko bertsiorik ez du Microsoft Project-ek.

Beste produktu bat *Multimeteo* izenekoa da. Software honen xedea eguraldiaren iragarpenak zenbait hizkuntzatan (ingelesez, frantsesez, alemanez, espainieraz eta nederlanderaz) automatikoki sortzea da, datu meteorologikoetatik abiatuz. Gaur egun ingelesez, frantsesez eta espainieraz bakarrik funtzionatzen du.

5.5. Hizketa naturalaren sorkuntza

Arlo honetan lanean hasi zirenean, gizakiak sortzen duen hizketa sintetikoki sortzea zen helburua, teoria elektroakustikotan oinarrituriko artikulazio-ereduak erabiliz. Gaur egun, arloa zabaldu egin da testutik hizketarako bihurketa ere hartuz.

Honelako sistema bat, euskarazko testuak ahots bihurtzen dituena, Bilboko Ingeniarien Goi Eskolan garaturiko *AhoTTS* sistema, http://bips.bi.ehu.es/tts/tts_eu.html helbidean proba daiteke.

5.6. Dokumentuen prozesamendua

Lan askotan dokumentuak erabiltzen dira, direla ikasliburuak, egunkariak, dokumentu juridikoak, gutun komertzialak edo beste hainbat motatakoak. Gizakiok ulertzeko antolatuta eta aurkeztuta dagoen informazio idatziz daude osatuak dokumentuak.

Dokumentuak sortzeko eta erabiltzeko prozesuan teknologiak produktibitate-hobekuntza nabarmenak eragin ditzake. Azken urteotan teknologia erabat sartu da bulegoetan, baina oraindik ere produktibitate-hobekuntza horiek maila apalean bakarrik lortu dira.

Dokumentu elektronikoen sorkuntzak bertan eduki berriak sartzeko aukera ekarri du, hala nola soinua eta bideoa, irudiez gain.

Hobekuntza hauen ondoan beste batzuk ere ari dira iristen teknologiaren eskutik: edukia, eta preseski hizkuntza naturalezko edukia, prozesatzeko, aztertzeko eta interpretatzeko sistemak.

Dokumentuen bizi-zikloan burutzen diren eragiketa nagusiak haiek sortzea, gordetzea, aurkeztea , banatzea, eskuratzea eta bilatzea dira. Gaur egun eragiketa horietan denetan esku hartzen dute informazio-teknologiek. Testu-prozesadoreek (MSWordek, adibidez) eta argitalpenak prestatzeko sistemek (QuarkXpress, Page Maker) sorkuntza fasean laguntzen dute, datu-base dokumentalek dokumentuak era kontrolatuan gordetzen. Banaketan teknologia desberdinak erabiltzen dira: faxa, inprimaketa, datuen truke elektronikoa (EDI edo electronic data exchange) eta posta elektronikoa. Batzuetan testu inprimatuak eskaner bidez, OCR programak erabiliz, dokumentu digitalak bihurtzen dira eta datu-baseetan gordetako dokumentuak bilaketa-motorren bidez bilatzen.

Eragiketa hauetan guztietan hobekuntza handiak lortzen dira hizkuntza naturala tratatzeko teknologiak erabiltzen direnean. Ikus ditzagun batzuk:

<u>Sorkuntzan</u>

Arlo honetako tresnarik sinpleena zuzentzaile ortografikoa da. Zorionez badugu euskaraz *Xuxen* izeneko era honetako tresna.

Ortografiaren zuzenketaren ondoren konplexutasun-mailan gramatika zuzentzailea dator. *Microsoft Office* paketeak, adibidez, badu *Multilingual Pack* izeneko bat eta honen barnean *Microsoft Office Proofing Tools* izeneko tresna-multzo bat:

- Zuzentzaile ortografikoa

- Gramatika-zuzentzailea
- Thesaurusa
- Hitz-mozketa
- Autozuzenketako zerrenda
- Autolaburtzailea
- Gramatikarako, hitz-mozketarako, ortografiarako eta thesauruserako hiztegiak

Gramatika-zuzentzaile hori gai da testua maila sintaktikoan aztertzeko eta baita maila sakonagoan ere, maila logikoan, subjektuen edo objektuen eta ekintzen arteko erlazioa hobeto ulertzeko. Adibidez, gai da esaldi bat aztertu eta forma pasibotik aktibora edo alderantziz bihurtuz berridazteko.

Hain aurreratuak ez diren beste gramatika-zuzentzaile batzuek analisi sintaktikoa egin gabe esaldiaren eta datu-base batean metaturik dituzten esaldiekin konparazioak eginez jokatzen dute.

Microsoft Office-ren hizkuntza desberdinetako bertsioek tresna-multzo desberdina dute. Euskarazko bertsioak Xuxen zuzentzaile ortografikoa erabiltzen du, baina ez du gramatika-zuzentzailerik

Bilaketan

Erabiltzaile batek, galdera bat eginez, galdera horri egoki erantzuten dion testu librez idatzitako dokumentu bat eskuratzeari esaten diogu bilaketa. Erabiltzaileak egiten dituen galderak izan daitezke hitz-zati bat, hitz bat, eragile boolearrez (ETA, EDO, EZ) loturiko hitz-konbinazio bat, esaldi bat edo esaldi-multzo bat. Bilaketa gehienak gaur egun hitz bat erabiliz burutzen dira web guneetan.

Euskaraz badugu bilatzaile bat euskarazko hitz baten edozein forma deklinatu galdetuz gero, hitz horren edozein forma deklinatu daukaten dokumentuak aurkitzen dituena. Ikus http://www.jalgi.com

5.7. Itzulpen-tresnak

Hizkuntza naturalen erabileran sarri aurkitzen dira eleaniztasunezko egoerak, eta gero eta sarriago gainera. Horren ondorioetako bat, gero eta itzulpen gehiago behar izatea da, eta beste bat itzulpen-lanak erraztuko dituzten tresnak garatzeko interesa handitzea.

Interes horren fruitu dira sortu diren tresna desberdinak: itzulpen-sistema automatiko eta erdi-automatikoak, itzultzaileei laguntzeko ordenagailu bidezko lagungarriak eta zenbait hizkuntzatan aldi berean testuak ekoizteko sistema automatiko edo erdi-automatikoak.

Itzulpen automatikoaren arloa oso zaila da. Ikusi besterik ez dago arlo honen garapena duela berrogei urte baino gehiago hasi zela eta oraindik ere emaitzak ez direla oso nabarmenak. Izan ere, hizkuntza batean dagoen testu bat beste hizkuntza batean jarri eta irakurlearengan inpresio berbera eragiteak hizkuntza-arazo zailak gainditzea eta beste gauza asko eskatzen ditu, bereziki munduaren ezagutza.

Dagoeneko merkatuan produktu asko aurkitzen dira, adibidez:

Ekoizlea	**Produktua**	**Ezaugarriak**
Lernout & Haspie-Apptek http://www.apptek.com EEBB	TranSphere Webtrans	Itzulpen automatikoa, honako hizkuntza bikote hauen artean: Ingelesa/Arabiera Ingelesa /Txinera Ingelesa /Koreera Ingelesa /Persiera Ingelesa /Turkiera Ingelesa /Japoniera Webean oinarrituriko sistema, bilaketa motorra eta itzulpen automatikoa konbinatzen dituena. Galderak egiteko hizkuntzak: ingelesa, arabiera, koreera, farsiera, txinera eta turkiera.
Atril Software, S.L. http://www.atril.com/ Katalunia	Déjà Vu	Ordenagailuz lagunduriko itzulpen-sistema. Itzulpen-oroimena (TM: Translation memory) eta adibidean oinarrituriko itzulpen automatikoa (EBMT: Example-Based Machine Translation) erabiltzen ditu. Edozein hizkuntzatarako balio du
Fujitsu http://software.fujitsu.com/jp/lingo/index-e.html Japonia	Atlas 7.0 Translingo	Itzulpen automatikoko sistema. Ingelesa-japoniera/japoniera-ingelesa Weberako bilatzailea eta itzultzaile automatikoa. Ingelesezko galderak japonierara itzultzen ditu eta japonierako emaitzak ingelesera itzultzen. Japoniera jakin gabe japonierazko informazioa bilatu eta ulertzeko tresna.
Heisoft http://www.heisoft.com Alemania	Reverso Pro	Itzultzaile automatikoa Hizkuntza-bikoteak: alemana-frantsesa alemana-ingelesa alemana-errusiera
Systran http://www.systransoft.com/ Frantzia	Systran	Itzultzaile automatikoa. Hizkuntza bikoteak: ingelesa-frantsesa ingelesa-alemana ingelesa-italiera ingelesa-espainiera ingelesa-portugalera
Dagostini http://www.dagostini.it/Hypertransinfo/Hypertransinfo Italia	Hypertrans	Ordenagailuz (Macintosh) lagunduriko itzulpena. Binaka konbina daitezkeen hizkuntzak: ingelesa alemana, frantsesa

		italiera espainiera
IBM http://www-3.ibm.com/software/ad/translat EEBB	TranslationManager Personal Translator Petra	Ordenagailuz lagunduriko itzulpena, itzulpen-oroimenarekin. Erabil daitezkeen hizkuntzak: afrikaans kataluniera txinera daniera nederlandera ingelesa suomiera frantsesa alemana greziera islandiera italiera japoniera norvegiera portugalera errusiera espainiera suediera Itzulpen automatikoa. Hizkuntzak: alemana-ingelesa ingelesa-alemana Itzulpen automatikoa. Hizkuntzak: ingelesa-italiera
Saillabs http://www.gmsmuc.de/		Testuen eta dokumentuen itzulpena online. Testu bat bidali eta berehala itzulita jaso. Hizkuntzak: ondokoen edozein bikote. Ingelesa, frantsesa, alemana, italiera eta espainiera
Language Partners International http://www.languagepartners.com/ EEBB	SDLX 4.0 Transcend	Itzulpen-oroimena Itzulpen automatikoa. Hizkuntzak: espainiera→ ingelesa ingelesa→ espainiera frantsesa → ingelesa ingelesa→ frantsesa alemana→ ingelesa ingelesa→ alemana ingelesa→italiera ingelesa→portugalera
Multiling Corporation http://www.multiling.com/fortis.htm EEBB	Transit 3.0	Ordenagailuz lagunduriko itzulpena itzulpen-oroimenarekin

LogoVista Corporation http://www.logovista.co.jp/english/ Japonia	LogoVista Multilingual 1.0	49 konbinazio honako hizkuntza hauen artean: ingelesa, japoniera, frantsesa, alemana, italiera, portugalera, espainiera eta koreera
Lingvistica http://www.ling98.com/ Herbehereak	Pars 5.0	Ingelesa→errusiera Errusiera→ingelesa Ingelesa→ukrainera Ukrainera→ingelesa
Language Engineering Corporation http://www.lec.com EEBB	Translate	Itzulpen automatikoa. Hizkuntzak: ondokoen bikoteak. ingelesa frantsesa italiera espainiera japoniera koreera poloniera portugalera errusiera ukrainera txinera

Goiko taulan aipatzen diren produktu asko "itzulpen automatiko"rako tresna gisa aurkezten dira eta zein hizkuntzatatik zein hizkuntzatara egiten duten itzulpena esaten da, adibidez: *TranSphere*, *Atlas 7.0* edo *Systran.*

Beste kasu batzuetan "ordenagailuz lagunduriko itzulpen-sistemak" direla eta "itzulpen oroimenean" oinarrituak daudela. Ikus dezagun zein diren desberdintasunak, makina bidezko itzulpenaren historia nolakoa izan den eta etorkizunari begira nolakoak diren joerak, zeren eta garrantzizko ondorioak dituzte gai hauek teknologiak hizkuntzaren bizitzan izan dezakeen eraginaren aldetik.

Makina bidezko itzulpenaren esparrua 1950eko hamarkadan jaio zen. Denbora luzea izan arren, hurrengo berrogei urtetan zehar ez zen askorik aldatu. Ikertzaileen artean eztabaidak zeuden hasiera hartan eta gutxi gorabehera eztabaidagai haiek hortxe jarraitzen dute.

Hizkuntza batetik bestera testuak bihurtzeko, tarteko hizkuntza bat, *interlingua* bat, erabiltzea proposatzen zuten batzuek eta transferentzia zuzena egitea beste batzuek. Hau zen eztabaidetako bat. Beste alde batetik, batzuek zioten makina bidezko itzulpena hizkuntzalaritza-arazoa dela nagusiki eta beste batzuek, aldiz, zentzu arruntarekin eta munduaren ezagutzarekin dagoela lotua batez ere. Azken urteotan ikerketa-lerro berriak ireki dira itzulpen berriak lehen egindako itzulpenetatik ateratako informazioan oinarritzea bilatzen dutenak. Joera honek hartzen duen formetako bat adibidean oinarrituriko makina bidezko itzulpena da. Sistema hauek informazioa prozesatzeko ohiko sistemak dira, jakintza linguistikorik gabeak. Beste sistema batzuetan itzulpen-corpus handi baten ezaugarri estatistikoetatik automatikoki ateratzen da itzulpenak egiteko ia jakintza guztia.

Transferentziaren bidea batez ere Japonian jarraitzen da. Bide honen zaleek diotenez, planteamendu hau egokiagoa da emaitza onak laster lortzeko.

Tarteko hizkuntza bat erabiltzearen aldekoek, berriz, hobesten duten planteamenduaren bi abantaila aipatzen dituzte: lehenbizikoa, hizkuntza-multzo bateko hizkuntza-bikote guztien arteko itzulpenak lortzeko aski dela tarteko hizkuntzatik hizkuntza bakoitzerakoa eta alderantzizkoa lortzea.

Gai hauetan benetan jarria ez dagoen batek baino gehiagok, oraindik ere, uste du hurbil egon daitekeela edozein testuren kalitate handiko itzulpen guztiz automatikoa lortzeko helburua, baina gauzak ez dira horrela, sistema askok itzulpen-lanari ekin aurretik edizio-lan bat egitea eta itzulpenaren ostean beste edizio bat egitea eskatzen dute.

Arrakastarik handienak testu-mota jakinekin lan egiteko prestaturik dauden sistemek lortu dituzte, hala nola, Kanadan eguraldiari buruzko txostenak lantzeko erabiltzen den METEO sistemak.

Orain arte arlo honetan arrakasta handiagoa lortu ez izateko arrazoi nagusia soluzioa hizkuntzalaritzatik etorriko zela uste izatea izan da. Gutxietsi egin izan dira giza itzultzaileek erabiltzen duten zentzu arrunta eta munduaren ezagutza, baina frogatua dago garrantzi handikoak direla. Horregatik, gaur egun gizakien eta makinen arteko elkarlana hartzen da benetan bideragarritzat.

Itzulpena esaten dugunean zer adierazi nahi dugun ere argitu egin beharko genuke. Itzulpen-mota desberdinak baitaude:

- **Birsorkuntza**. Poesiak edo publizitateak, esate baterako, mezuak hizkuntza batetik bestera pasatze hutsa baino askoz gehiago eskatzen dute, inpresioa transmititzea baita genero horietan garrantzitsua eta ez hitzen ohiko esanahia itzultzea.

- **Lokalizazioa**. Programa, eskuliburu eta sistema informatikoen itzulpena da hau.

- **Informazioaren zabalkunderako itzulpena**. Informazio teknikoa hizkuntza desberdinetako erabiltzaileei banatu behar zaienean garrantzitsuena informazio guztia aldakuntzarik, gehikuntzarik eta murriztapenik gabe, zehatz eta argi transmititzea da, dotorezia gora behera.

- **Gutxi gorabeherako itzulpenak**. Adibidez, ingelesez ez dakien norbaitek erabil dezake Systran-en online bertsioa (http://www.systransoft.com/) testuak itzultzeko, baina ezingo du espero itzultzaile profesional batek egin dezakeen itzulpenaren kalitatea. Hala ere, sistemak hauek badute beren praktikotasuna.

Goiko lehen bi arloetan ia ezinezkotzat hartzen da itzulpen guztiz automatikoa, batez ere lehenbizikoan. Hirugarrenean urrunago joan daitekeela ikusten da eta emaitza politak ere lortu dira, Siemens-Nixdorf etxearen METAL sistema, esate baterako. Azkenik, gutxi gorabeherako itzulpenak egiten dituen sistema asko daude dagoeneko.

Sistema hauek guztiek hazkunde handia dute gaur egun, mundu globalizatu batean hizkuntza desberdinetan idatzitako testu-masa handiak erabili behar izaten direlako. Izan ere, mundu guztiak nahiago du bere kulturako hizkuntzan testuak irakurtzea, baita beste hizkuntzak ongi samar menderatzen dituenean ere, bestela denbora eta zehaztasuna galtzen dituelako. Gaur egun ingelesa *lingua franca*-tzat

hartu arren, ikuspegi ekonomiko hutsetik ere ikusten da soluzio hobea dela teknologiaren laguntzaz nork bere hizkuntza erabili ahal izatea.

Hau ulertzean baino areago gertatzen da norberarena ez den hizkuntza batean informazioa jarri nahi denean. Txosten tekniko edo komertzial bat atzerrira bidali nahi duenarentzat oso erosoa litzateke bere hizkuntzan lan egin eta itzultzaile profesional baten beharrik gabe, makinarekin elkarrizketan, atzerriko hizkuntzara txostena itzuli ahal izatea. Hau nahikoa bideragarria da gaur egun.

Itzultzaile profesionalek behar dituzten tresnak beste era batekoak dira. Hauek, gaur egun dituzten sistemarik egokienak itzulpen-oroimenaren bidezkoak dira, goiko taulan aipaturiko Trados edo DejaVu-ren tankerakoak.

5.8. Hizkuntza desberdinetan bilaketak egiteko tresnak

Testuzko edukia duten datu-baseetan, hizkuntza bat baino gehiago erabiliz, kontsultak egiteko arazoa, gorago, 4.6 puntuan aipatu dugun bilaketaren arazoaren konplikazio bat dela esan genezake.

Gai honi eman izan zaizkion erantzunak aztertzen hasi aurretik ikus dezagun lehenbizi dokumentazio-sistema batean dokumentuak gorde eta gero aurkitzeko erabili izan den metodoa zein izan den.

Enpresa edo erakunde bateko dokumentazio-departamentura egunero dokumentu asko iristen badira (txostenak, arauak, kontratuak, etab.) eta ez badira jasotzen gero erraz aurkitzeko moduan, laster sortuko da sekulako anabasa.

Horrelakorik gerta ez zedin, iristen zen dokumentu bakoitzari fitxa bat egiten zitzaion eta bertan, dokumentuaren edukia ongi adierazten zuten hitz gako batzuk jartzen ziren.

Demagun dokumentu batek nekazaritzan erabiltzen diren berotegietarako plastikoen konposizioaz hitz egiten duela. Dokumentu horren fitxan jarri beharreko hitz gakoak "nekazaritza", "berotegiak", "plastikoa" izan litezke. Orain, pentsa dezagun gai horri buruzko dokumentuak artxiboan bilatzen dituen batek kontsulta egiten duenean "laborantza", "negutegiak" eta "polimeroak" hitz gakoak erabiltzen dituela. Jakina, ez luke interesatuko litzaiokeen dokumentu hori eskuratzerik izango. Arazo hau saihesteko "thesaurus"ak deritzen hiztegi kontrolatuak erabiltzen dira. Hiztegi hauetan sinonimoak daudenean bat hobesten da eta besteak baztertzen dira, hitzen arteko erlazio hierarkikoak definitzen dira, zein den termino zabalagoa eta zein adiera murritzagokoa, etab. Horrela, esan lezake:

> Laborantza ez erabil, erabil nekazaritza
> …
> plastiko, termino zabalagoa polimero
> …
> negutegi ez erabil, erabil berotegi

Dokumentua gordetzen duenak eta bilatzen duenak hizkera arautu hau erabiltzen badute, bigarrenak ez du arazorik izango dokumentu hori eskuratzeko garaian.

Jokabide hau erabiliz hizkuntza desberdinetan kontsultak egiteko sistema bat eraiki nahi izanez gero, thesauruseko termino bakoitzaren itzulpena beste hizkuntzetan zein izango den erabaki behar da. Horrekin, ordea, ez dira arazo guztiak

konpontzen, zeren hizkuntza bat baino gehiago erabiltzen direnean, terminoen hedadura semantikoak ez dira bat etortzen. Egia da, esate baterako euskarazko "haragi" eta "okela" hitzen ordaina dela espainierazko "carne", baina alderantzizkoarekin kontuz ibili behar da, zeren "carne" ez da beti "okela".

Arazo hau konpontzeko era bat, espainierazko bi hitz balira bezala, "carne (alimento)" eta "carne (en general)" bereiztea izan liteke.

Lan hauek automatizatu ere egiten dira. Horretarako, indexatzailea deritzon programa informatiko batek testua hustu eta hitz gakoak izan daitezkeenak identifikatzen ditu. Gero, testu osoan azaltzen diren hitzen eta hitz gakoen arteko loturak eskuz landu daitezke edota lehendik indexaturik dauden dokumentuetatik atera, ikasketa-prozesu automatiko batez baliatuz. Kontsulta egiteko garaian ere jokabidea antzekoa izango da, h.d. kontsultan erabiltzen diren hitzei zein hitz gako dagozkien ondorioztatu beharko du sistemak. Testu-hitzen eta hitz gakoen arteko loturak hizkuntza desberdinetako hitzen artekoak izanez gero, plantea daiteke bilaketak hizkuntza batean baino gehiagotan egitea, dokumentuaren hizkuntza kontuan izan gabe.

Gaur egun, ordenagailuak azkarragoak direnez, dokumentuen bilaketak egiteko sistemek ez dute zertan hitz gakoak erabili behar, testu libreko bilaketak egiten dira normalean. Indexatzaileak testu-hitz guztiak husten ditu eta haien kokapenaren datuak indize batean gordetzen ditu. Bilaketa egiteko garaian bilatzaileak ez du bilatu nahi den terminoa testuak burutik burura arakatuz (sekuentzialki) bilatzen, baizik eta indizean egiten du bilaketa, honen emaitza oso laster aurkitzen du eta bilatzen dugun terminoaren agerpenak zein dokumentuetan dauden esaten digu berehala.

Euskaraz, oso hizkuntza eranskaria izaki, indizean ez dira testu-hitzak gordetzen, hauei dagozkien lemak baizik. Honek esan nahi du indexatzaileak testu-hitz bakoitzaren analisi morfologikoa egin behar duela. Zorionez, badugu euskaraz lan hau egiten duen tresnarik.

Euskarazko hitzak erabiliz ingelesezko dokumentuetan bilaketak egin nahi bagenitu, sistemak honela funtzionatu beharko luke:

1. Ingelesezko dokumentua hustu testu-hitz guztiak ateratzeko eta indizea sortzeko.
2. Ingelesezko termino bakoitzari dagokion euskal lemaren berri ematen duen hiztegi bat izan.
3. Galdera egiteko erabiltzen den euskal hitza lematizatu
4. Ateratzen den euskal lemari ingelesezko zein hitz dagokion begiratu hiztegian
5. Indizean begiratu zein dokumentutan agertzen den ingelesezko hitz hori.

Horrela, euskaldun batek "arrainak" galdetzen badu, lematizatzaileak "arrain" lema aterako du, hiztegiak esango du ingelesezko baliokidea "fish" dela eta bilatzaileak hitz hori zein dokumentutan dagoen.

5.9 Hizketaren prozesamendua hizkuntza ezberdinetan

Hizkuntza desberdinen prozesamendua ez da testutan bakarrik behar; ahozko hizkuntzan ere bai. Honen adibideak dira hizkuntza desberdinetan lan egin behar duten telefono bidezko informazio-zerbitzu automatikoak.

Hizketaren prozesamendua berez konplikatua baldin bada, hizkuntza bat baino gehiago erabili behar dituen sistema baten konplexutasuna are handiagoa da. Horregatik, arlo honetan egin diren aurrerapenak ez dira oraindik oso handiak, baina interes handia dago eta ikerketa asko egiten da.

Adar asko dituen ikerketa-arloa da:

Hizkuntza mintzatuaren identifikazioa. Hiztun batek erabiltzen duen hizkuntza zein den automatikoki ezagutuz gero, bidera daiteke bere hizkuntzan erantzungo dion zerbitzura. Hau erabilgarria da hizkuntza desberdinak erabiltzen dituen jendeari zerbitzuak eskaintzen dizkioten bulego publikoetan.

Hizketaren ulerkuntza hizkuntza batean baino gehiagotan. Dagoeneko badira merkatuan hizkuntza batean baino gehiagotan erabil daitezkeen diktaketa-sistemak. Hauetako bat Philips etxearen FreeSpeech 2000 softwarea da.

Mikrofonoa eta bozgorailua integratuak dituen sagu bat etortzen da softwarearekin batera. Hau ordenagailuari konektatu, programa instalatu eta testu batzuk irakurri behar zaizkio, ordenagailuak duen soinu-txartelera eta gelan dagoen zarata-mailara egokitu dadin. Gero sistemari irakatsi egin behar zaio. Ordu laurden batez edo horrela testu batzuk irakurtzen zaizkio, erabiltzailearen tonua, ozentasuna eta ahoskera ezagut ditzan.

Irakasteko erabiltzen den testua irakurri ahala, hitzok berdez agertzen dira pantailan. Hitz bat ezin ulerturik geratzen bada, hartzen duena gorriz agertzen da eta berriz irakurri behar zaio ongi ezagutzen duen arte.

Prozesu hau amaitu ostean, testu-prozesadorea irekitzen da eta has daiteke diktatzen. Ordenagailu on batean hitzak esan eta handik pare bat segundora agertzen dira pantailan. Diktaketa egiteko ez ezik ordenagailua ahotsaz gobernatzeko ere balio du programa honek. "Ireki", "leihoa handitu" eta horrelako aginduak eman dakizkioke ordenagailuari.

Zenbat eta gehiago erabili programa, orduan eta hobeto interpretatzen du esaten zaiona. Hizkuntza batean baino gehiagotan funtziona dezake, baina ez dakigu, egia esan euskaraz erabiltzeko moldatzeak zer zailtasun izan lezakeen.

Ahozko hizkuntzaren itzulpena. Arlo hau ikerketa-mailan dago gaur egun. Proiekturik aurreratuenetako bat JANUS izenekoa da. Garatzen ari diren hizketa itzultzeko sistemak giza interpretari batek egiten duen lanaren antzekoa egiten du. Gai jakin batzuk bakarrik itzultzeko dago prestatua: hitzorduak emateko, hoteletan erreserbak egiteko edo bidaiak programatzeko, adibidez.

Giza interpretari batek ez bezala datu-baseak denbora errealean kontsulta ditzake emendiozko informazioa emateko, hala nola trenen ordutegiak edo hirietako planoak.

Teklaturik erabili gabe funtzionatzeko prestatua dago. Hizketa aski ez denerako, eskuzko idazketa edota keinuak ezagutzeko ere prestatua dago.

Solas arruntean itzulpena egin ahal izateko, gai izan behar du hizketan arruntak diren akatsak, etenaldiak, inguruko zaratak, etab. jasateko. Itzulpen literala baino areago interpretazio lagungarria du helburu.

Hitzorduak erabakitzeko behar den hizketarako 3.000tik 5.000 hitzera bitarteko hiztegia behar du. Bidaiak programatzeko behar den hizketaren arloan ingelesa eta alemana onartzen ditu sarrerako hizkuntza gisa eta irteerakoak, berriz, ingelesa, alemana eta japoniera izan daitezke. Beste arlo batzuetarako koreerarako eta espainierarako moduluak ere garatu dira.

Beste proiektu interesgarri bat Verbmobil izenekoa da. Hizkuntza desberdinak egiten dituzten solaskideei ahozko elkarrizketan laguntzeko sistema bat da, dela aurrez aurre, dela telekomunikazioak erabiliz. Verbmobil sistemak zerbitzari batean lan egiten du eta saketako telefonoaren bidez komunikatzen da erabiltzailea sistemarekin. Onartzen dituen itzulpenak honako hauek dira:

alemana ←→ ingeles amerikarra
alemana ←→ japoniera
alemana ←→ txinera

Proiektu honetan Siemens eta Philips enpresak eta unibertsitate batzuk izan dira partaideak.

5.10. Hizkuntza mintzatuaren identifikazio automatikoa

Zenbait egoeratan oso inportantea da hiztun batek egiten duen hizkuntza zein den identifikatzea, esate baterako hainbat hizkuntza desberdin tratatzen dituen itzulpen-sistema batek lehenik eta behin hiztunak egiten duen hizkuntza zein den igarri behar du.

5.11. Hizkuntza idatziaren identifikazioa

Aurreko puntuan agertzen diren sistemen helburu bera dute, baina testuekin lan egiten dute. *Xerox* etxearen Grenobleko ikerketa-laborategiak garatua du 47 hizkuntza desberdin bereiz ditzakeena, 5 hitz baino gehiagoko testu bat emanez gero. Hizkuntza horien artean euskara ere badago. Ikus:
http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html

5.12. Hizkuntza naturalaren prozesamendurako hizkuntza-baliabideak

Goian ikusi ditugun sistemak garatzeko, hobetzeko eta ebaluatzeko, makinek irakur ditzaketen hizkuntza-datuak eta -deskribapenak dira hizkuntza-baliabideok.

Hizkuntza-baliabide hauen adibideak dira testu- eta hizketaldi-bilduma edo corpusak, datu-base lexikalak, gramatikak eta terminologiak.

Sistema hauen interesa handitu egin da, gainera, teknika estatistikoetan oinarrituriko sistemen garapena indarrean dagoenez geroztik, sistema hauek corpus handiak behar dituztelako.

Zuzentzaile ortografikoak, testu-prozesadoreetan lerro-bukaerako hitz-mozketa egokiak egiteko sistemak, gramatika-zuzentzaileak eta antzeko aplikazioak garatzeko beharrezkoak dira testu idatzien corpusak.

Lexikografoek ere, hitzen erabilerak aztertzeko, testu idatzien corpusak erabiltzen dituzte. Euskararen kasuan bi corpus erabiltzen dira, bata Orotariko Euskal Hiztegikoa, XX. mendea arteko testu gehienak biltzen dituena, eta bestea Egungo

Euskararen Bilketa-lan Sistematikoa (EEBS) deritzon proiektuan erabiltzen dena eta XX. mendearen hasieratik honako testuen lagin adierazgarri bat biltzen duena. Azken hau kontsultagai dago http://www.uzei.org helbidean.

Euskararen tratamendu automatikorako oinarri lexikala, berriz, EDBL (Euskararen Datu-Base Lexikala), da. Bera da *XUXEN* zuzentzaile ortografikoaren, *MORFEUS* analizatzaile morfologikoaren eta *EUSLEM* lematizatzailearen oinarri lexikala. Etorkizunean, analisi sintaktiko-semantikorako ere erabiliko da.

70.000 bat sarreraz osaturik dago, hiru atal nagusitan banatuak: hiztegi-sarrerak (hiztegi konbentzional batean aurkitzen direnak bezalakoxeak), aditz-formak eta morfema ez-independenteak, bakoitza bere informazio morfologikoarekin.

6. HIZKUNTZA-PLANGINTZA ETA ZIENTZIA, TEKNOLOGIA ETA BERRIKUNTZA-PLANAK

Aurreko puntuetan ikusi dugun legez, hizkuntzaren industriek eragin handia izaten hasi dira hizkuntzaren bizitzan, informazioaren eta jakintzaren gizarte honetan. Bestalde, garbi ikusi ahal izan dugu hedadura handiko hizkuntzek ateratzen diotela batez ere probetxua aurrerapen horri eta besteek arrisku handia dutela gehiago ahultzeko.

Bistan da merkatuaren indarrek ez dutela hizkuntza ahulenen alde jokatzen. Inor gutxik izan lezake interesa hedadura txikia duten hizkuntzetarako tresnen garapenean inbertitzeko. Ezinbestekoa izango da hizkuntzaren bizitzaz arduratzen diren erakundeek hau kontuan izanik hizkuntzaren industrien garapena bultzatzea.

Arazoak hizkuntza-plangintzaren esparrua askogatik gainditzen duenez, zientzia, teknologia eta berrikuntza-planetan ere kontuan izan beharreko arazoa da zalantzarik gabe.

Hainbat ekintza-mota dira arlo hauetan aurreikusi beharrekoak:

- Hizkuntzaren erabilera errazten duten produktuen eta zerbitzuen sorkuntza eta zabalkundea sustatzea.
- Hizkuntza naturalaren prozesamenduaren ikerketa eta irakaskuntza bultzatzea.
- Ahozko eta idatzizko hizkuntzaren corpus handiak sortzeko, komunikabideei eta industria editorialari lankidetza eskatzea.
- Hizkuntza-industrien eta honek sortzen dituen produktuen garrantziaz gizartea informatzea eta kezkaraztea.
- Zerbitzu publikoetan hizkuntzaren industrien produktuak erabiltzea.

Eusko Jaurlaritzak, erronka horri erantzun nahian, Ikerkuntza Zientifikoaren 2000-2003 Plan Nazionalaren "Informazioaren gizartea" Area sektorialerako "Hizkuntzaren Industria" delako Ekimen-proposamena egin zuen: http://www.euskara.euskadi.net/r59-734/eu//datos/azkentxostena.pdf

Europako Elkarteak ere badu MLIS *(MultiLingual Information Society)* izeneko programa bat 1996ko azaroan abiarazi zuena, 1997-99 bitartean Europako Elkartean hizkuntza-aniztasuna Informazioaren Gizartean sustatzeko. Programa honen barnean landu diren proiektuak hiru helburu hauetara bideratuak izan dira:

- Europako hizkuntza baliabidetarako (hiztegi elektroniko eleaniztunak, datu-base terminologikoak, eta testu-corpusak) zerbitzu-multzo baten sorkuntza sustatzea
- Hizkuntza-teknologien, baliabideen eta estandarren erabilera sustatzea.
- Hizkuntza-tresna aurreratuen erabilera sustatzea, bai Europako Elkartearen mailan eta baita Estatu Kideen sektore publikoetan ere.

Bestalde, 2001eko IST (Information Society Technologies) programaren barnean III. Ekintza Gakoan (Multimedia eduki eta tresnak) bada Ekintza-lerro bat, III.3 HLT (Human Language Technologies) deritzona.

HLT programaren xede nagusia ekonomikoa da, e-business edo sare telematikoen bidezko negozioa bultzatzea, testuinguru global batean, mundu guztiari aukera berdinak emanez. Hori lortzeko, teknologia eleaniztunak garatu beharra ikusten

dute eta ikerkuntza generikoa edo aplikatua sustatzen. Ekintza-lerro honek bi adar ditu:

IST2001 - III.3.1 Web eleaniztuna

Interneten eleaniztasuna sustatzeko, hizkuntza idatziaren eta mintzatuaren itzulpen automatikoa barne.

IST2001 - III.3.2 Elkarrekintza natural eta eleaniztuna

Hizkuntza desberdinetako hiztunei lagunduko dieten teknologiak garatzeko. Esate baterako:

- **Etxean**: sare telematikoetara konektaturiko informazio- eta entretenimendu-tresnekiko elkarrekintzan, etxeko zerbitzu konplexuei aginduak emateko eta haiek kontrolatzeko, are hizketaren ulerkuntza eta sintesiaren bidez.

- **Lanean**: pertsona arteko eta taldeko komunikazioa erraztuko duten teknologien bidez, hala nola bilkura birtualetan eta elkarrizketa eleaniztunetan.

- **Telekomunikazio mobiletan**: sakelako telefono eta ordenagailuetan, adibidez.

7. BIBLIOGRAFIA

- ABBOU, A.: LEFAUCHEUR, I. eta T. (1987): *Les industries de la langue : applications industrielles du traitement de la langue par les machines*. Éditions DAICADIF. Paris.
- CARRÉ, R.; DÉGREMONT, J.F.; GROSS, M.; PIERREL, J.M.; SABAH, G. (1991) : *Langage humain et machine.* Presses du CNRS. Paris.
- CHAUMIER, J. (1988) : *Le traitement linguistique de l'information,* 3. ed. ESF. Paris.
- DAOUST, F.(1996) : *SATO (Système d'analyse de texte par ordinateur), Version 4.0, Manuel de référence, Service d'analyse de texte par ordinateur (ATO).* Université du Québec à Montréal, 256 orr.
- DEWEZE, A.; DENIEL, Y. (1993): *Informatique documentaire,* 4. Masson. Paris.
- DÍEZ CARRERA, C. (1994): *Las industrias de la lengua : panorámica para los gestores de la información.* Biblioteca Nacional: FESABID. Madrid.
- GOUVERNEMENT DU QUEBEC, MINISTERE DES COMMUNICATIONS, DIRECTION GENERALE DES TECHNOLOGIES DE L'INFORMATION (1994): *Guide d'analyse et pratiques recommandées en gestion de l'information textuelle au gouvernement du Québec.* Québec.
- GRABE, W.; KAPLAN, R. B. (1986): *Science, technology, language, and information: Implications for language and language-in-education planning*. International Journal of the Sociology of Language 59, 47-71.
- DRAGO, T.; RUIZ DE GOPEGUI, L.A. (1991): *Industrias de la lengua*. VI Encuentro Iberoamericano de Comunicación. Buenos Aires.
- VIDAL BENEYTO, J. (1991): *Las industrias de la lengua*. Fundación Germán Sánchez Ruipérez, D.L. Madrid.
- MANIEZ, J. (1987) : *Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires.* Éditions d'Organisation. Paris.
- MATHIAS, J. & KENNEDY, T. L. (Eds.) (1980): *Computers, language reform, and lexicography in China;* A report by the C(hinese) E(nglish) T(ranslation) A(ssistance) Delegation. Pullman. Washington State UP
- SABAH, Gé.(1988-1989). *L'intelligence artificielle et le langage.* Hermès. Paris.
- SIBERTIN-BLANC, Ma. (1994): *Nouvelles technologies et communication de l'information. De l'analyse des besoins à l'ingénierie documentaire.* ADBS Éditions. Paris.