

02 IKERKETA LERROAK >>>

HIZKUNTZAREN TRATAMENDU AUTOMATIKORAKO TRESNAK

Elena García Berasategi
ELEKA

Helbide elektronikoa: elena@eleka.net

Teknologia berriek berebiziko garrantzia hartu dute azken urteetan gure bizitzan. Apenas konturatu gabe egunero egiten ditugun hainbat ekintzetan teknologien laguntza jasotzen dugu. Hizkuntzaren alorrean ere teknologia berriek ekarpen nabarmenak eskaintzen dizkigute, ondoren ikusiko den moduan.

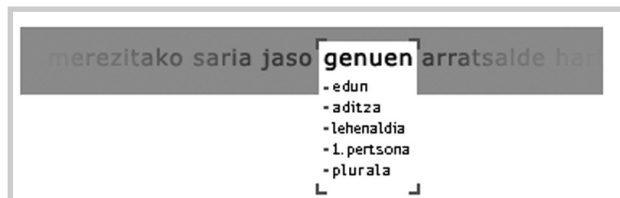
Zaila da ordenagailu bati hizkuntza bat ulertaraztea edo hizkuntza ezberdinak bereizten irakastea. Pixkanaka-pixkanaka, ordea, hizkuntzen zehaztasun gehiago ezagutzen dituzte ordenagailuek.

Hasteko, lexikoa edo hiztegia da ordenagailu batek hizkuntza bat ezagutzeko behar duen funtsezko zerbait. Lexiko horretan ahalik eta hitz gehien egotea komeni da eta, noski, dagokion kategoria eta azpikategoria. Hitzen kategoria eta azpikategoria oso baliagarriak izango zaizkigu, adibidez, pertsona izen bereziak eta leku izen bereziak bereizteko eta ondoren ordenagailuari esateko mota bakoitza nola deklinatu dezakeen.

Hori izango da hurrengo pausua. Hitzek zein aurrizki/atzizki har ditzakete? Bakoitzaren kategoria eta azpikategoria gordetzen denez, kategoria bakoitzak har ditzakeen bukaera, deklinabide atzizki eta aurrizkiak definituko dira. Honela, lema bakoitzaren forma posible guztiak zein izan daitezkeen jakingo du ordenagailuak. Adibidez, *etxe* beza-

lako izen arrunt batek lekuzko genitiboan *etxean*, *etxetan* edo *etxeetan* formak hartzen dituela zehaztuko da mugatu singularrean, mugagabea eta mugatu pluralean.

Behin lexikoa eta hartu ditzaketen atzizkiak definituta daudenean, lematizatzaile bat garatu ahal izango da, testu batean edozein forma hartuta, forma horren lema eta informazio morfologikoa emango diguna.



1. irudia: lematizazioaren adibidea

Hitz mailako forma guztiak definitu ondoren gramatika edo errege-lak definitu beharko dira, hitz horiek nola konbinatu daitezkeen jakiteko. Hitzak konbinatu ahala sintagmak osatuko dira eta sintagmen elkarketatik, berriz, esaldiak.

Hau guztia da, oso labur esanda, ordenagailu batek behar duen informazioa hizkuntza bat prozesatu ahal izateko. Jakintza horrekin, lematizazioa, analisisa eta sorkuntza egiteko gai izango da ordenagailua eta horiekin batera, ondoren azalduko diren tresnak: hizkuntza-ezagutzailea, zuzentzaile ortografikoak, Interneteko bilatzaile aurreratuak, entitateak erauzteko tresnak, terminologia erauzketa automatiko eleanitza, itzultzaile automatikoak eta orokorrean euskaraz idazten laguntzeko tresnak.

HIZKUNTZA-EZAGUTZAILEA

Adierazleak lortzeko aproposa den aplikazio honek idatziriko testuen hizkuntza detektatzen du. Horretarako, hizkuntza bakoitzeko hitz zerrendak eta karaktere-konbinaketa posibleen multzoak hartzen dira, bakoitzaren erabilpen-maiztasunekin batera. Horietan oinarrituz eta hizkuntza bakoitzak dituen ezaugarri bereizgarriez baliatuta, tresna honek edozein testu zein hizkuntzatan dagoen esango du.

Edozein erakundetan erabil daitekeen tresna da, bertan egiten den euskarazko produkzioa neurtu nahi denean adibidez. Honela, erakunde jakin bateko langileek posta elektronikoak euskaraz edota gazteleraz idazten dituzten jakin daiteke eta zein proportziotan.

XUXEN, ZUZENTZAILE ORTOGRAFIKOAK

Zuzentzaile ortografiko hau hainbat plataforma eta bulegotika paketetan integratzen da, hala nola, MS Office, OpenOffice, MAC Office, Windows

Jakintza horrekin, lematizazioa, analisisa eta sorkuntza egiteko gai izango da ordenagailua eta horiekin batera, ondoren azalduko diren tresnak: hizkuntza-ezagutzailea, zuzentzaile ortografikoak, Interneteko bilatzaile aurreratuak, entitateak erauzteko tresnak, terminologia erauzketa automatiko eleanitza, itzultzaile automatikoak eta orokorrean euskaraz idazten laguntzeko tresnak.

**XUXENeten ikuspegitik
hitzen zuzentasuna
egiaztatzeko hiru osagai
nagusi daude: lexiko
orokorra, lexiko
korporatiboa eta lexiko
pertsonala.**

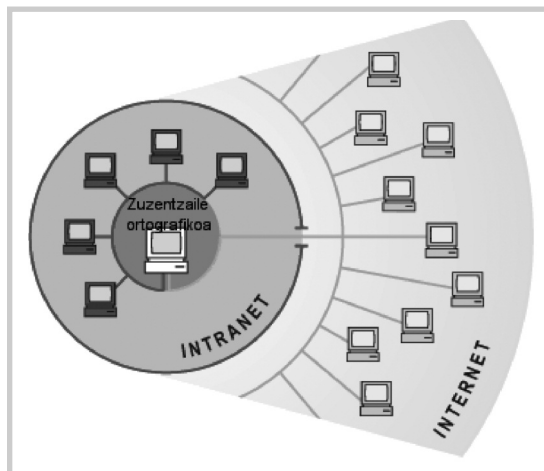
**XUXENetek, beste
zuzentzaile ortografiko
guztiek bezala, ez du
testuingurua kontuan
hartzen; beraz, ondoko
kasu hauetan ez du
laguntzarik emango.**

eta Linuxen. Egun bada Interneten testuak zuzentzeko aukera eskaintzen duen doako zerbitzua, ordenagailuan ezer instalatu gabe alegia. Atzerrian egonda adibidez, www.xuxen.com helbidean dagoen zuzenketa-sistema erabili ahal izango da testuak zuzentzeko. Bestela, euskaraz testuren bat idatzi nahi duenak Eusko Jaurlaritzako web gunetik deskargatu dezake doan http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_software/eu_9567/xuxen.html helbidean. Bertan programaz gain, instalatzeko laguntza eta eskuliburuak daude.

XUXENen helburua euskaraz idatzitako testuen akats ortografikoak detektatu eta zuzentzea da. Dokumentuak hitzez hitz tratatzen dira hizkuntzaren arabera, hitz bakoitza zuzena denentz erabakiz. Hitzak ezagutzen dituen bitartean lanean jarraitzen du etengabe, baina hitzen bat ezagutzen ez duenean abisua ematen du. Hau da ezagunena den XUXENen helburua baina bada XUXENet izeneko beste aplikazio bat, sarean lan egiten duena. XUXENeten ikuspegitik hitzen zuzentasuna egiaztatzeko hiru osagai nagusi daude: lexiko orokorra, lexiko korporatiboa eta lexiko pertsonala.

Abisu baten aurrean erabiltzaileak honako aukera hauek ditu:

- Hitza onartu agerpen konkretu horretan.
- Hitza zuzendu teklatutik, erabiltzaileak hitz-forma egokia dakienean.
- Proposamenak eskatu, erabiltzaileak hitza nola idazten den ez dakienean edo zalantzan dagoenean.
- Hitza onartzeaz gain hitzaren erroa hiztegi pertsonalean edo korporatiboan sartu; horrela hitz hori ez ezik, bere erroatik sortutako forma guztiak (flexioak) ere onartuko dira dokumentu horretan eta etorkizunekoetan. Hiztegi korporatiboan datuak sartzeko baimen berezia beharko da.



2. irudia: XUXENet zuzentzaile ortografikoa sarean

XUXENeten prozesua konplexu samarra da, batez ere euskararen kasuan, euskara eranskaria denez ezin izan baita erabili inguruko erdaretan bezala hitz-forma guztiekin osatutako zerrenda.

Zer da oraingoz egiten ez duena? XUXENetek, beste zuzentzaile ortografiko guztiek bezala, ez du testuingurua kontuan hartzen; beraz, ondoko kasu hauetan ez du laguntzarik emango:

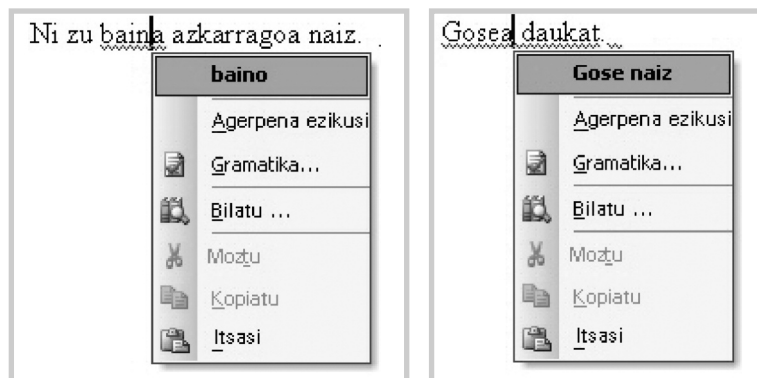
- Arazo sintaktikoetan: “Ni dut” bezalakoak ontzat emango ditu, hitzez hitz hartuta horiek denak zuzenak baitira.
- Akats baten ondorioz sortutako hitz egokietan: “klaseak hazi dira” idatziz gero, “hazi” beharrean “hasi” jarri nahiko zen baina hori ez litzateke harrapatuko. Berdin gertatzen da gaztelaniaz “Ahí que hacer” idazten bada.

Gerta daiteke, baita ere, abisua ematea ondo dauden hitzetan, hauen erroak XUXENeten lexikoetan ez daudenean, baina lexikoan barneratu ondoren hau ez da errepikatuko.

Goian aipatu diren komuntzadura erroak topatu eta konpontzen laguntzeko sintaxiaren ezagutza sakona behar da. Euskal Herriko Unibertsitateko IXA Taldea (<http://ixa.si.ehu.es/Ixa>) eta Eleka Ingeniaritza Linguistikoa (www.eleka.net) elkarlanean ari dira etorkizun hurbil batean arazo honi konponbidea eskaintzeko. Orain arte lortutako emaitzen artean estilo zuzentzaile bat garatu da, komak behar dituzten lokailuek izango dituztela bermatuko duena adibidez edota errore tipiko batzuk harrapatzen dituen.

**Euskal Herriko
Unibertsitateko IXA
Taldea**
(<http://ixa.si.ehu.es/Ixa>)
**eta Eleka Ingeniaritza
Linguistikoa**
(www.eleka.net)
**elkarlanean ari dira
etorkizun hurbil batean
arazo honi konponbidea
eskaintzeko.**

**XUXENek hemendik
gutxira bere bizkaierazko
baliokidea izango du:
XUXENb.**



3. irudia: Estilo-zuzentzailearen proposamenak

XUXENek hemendik gutxira bere bizkaierazko baliokidea izango du: **XUXENb**.

XUXENb-k izango dituen baliabideak eta erabiltzeko erak, XUXEN estandarrak dituen berak izango dira. Gainera, posible egingo dio erabil-

Euskaraz Lan Egiteko Tresnak: ELET. Web aplikazio honetan Elekan dauden oinarritzko teknologiak bildu dira. ELETek eguneroko lanean sortzen diren beharrei on-line erantzun bat eman nahi die, euskararen erabilera sustatzeko asmoz. Honek langileen autonomia areagotzen du eta autoikasketa bultzatzen lagunduko luke.

tzaileari testu jakin baten aurrean dagoenean, bizkaierazko edo euskara batuzko zuzentzailea erabiltzea testuaren izaeraren arabera. Hots, XUXENb-k batuaz ere zuzentzeko aukera izango du integratuta.

Bizkaiera euskalkiak hainbat berezitasun ditu. XUXENb-ren xede nagusia hauek ezagutzea eta zuzenketan kudeatzea da:

- Lexikoa: Hainbat hitz propio (*berba, gatzatu, ugaraxo, laga, ezpabere...*) eta aldaera (*barri/berri, emon/eman, pekatu/bekatu, igon/igo...*) ditu bizkaierak.
- Morfema dependenteak:
 - deklinabide-atzizkiak (*diruAGAZ*)
 - erlazio atzizkiak (*dakiALA*)
- Adizkiak: laguntzaileak eta trinkoak (*dabe, zatozie, dakarre*)
- Morfotaktika: lema bakoitzak har ditzakeen atzizkiak.
- Morfofonologia erregelak: *eskola + a > eskolEA; da + Ela > dAla*

Oraingo XUXEN estandarra bizkaieraz idatzita dagoen testu bati aplikatzen zaionean, izatez ondo dauden hitzak okertzat jotzen ditu, beheko adibidean ikusten den bezala.

Urteetako · ibilbide · horretan · bizkaiera · eredu ·
bat · finkatzen · joan · da, · denporearen · buruan ·
formalizazino · handiagoa · emon · jakona ·
Irakasle, · idazle · eta · ikertzaile · mordoaren ·
izenak · aitatu · leitekez, · bide · horretan · lehen ·
urratsen · emoile · izan · ziran · Mikel · Zarate, ·
Karmelo · Etxenagusia, · Ander · Manterola, ·
Lorentzo · Zugazaga · eta · Jose · Mari ·
Irazolarenak · kasurako. ¶
¶

4. irudia: bizkaierazko hitz zuzenak txartzat hartuta

XUXENb-k hitz horiek ontzat eman eta ez ditu okertzat joko.

Euskaraz Lan Egiteko Tresnak: ELET. Web aplikazio honetan Elekan dauden oinarritzko teknologiak bildu dira. ELETek eguneroko lanean sortzen diren beharrei on-line erantzun bat eman nahi die, euskararen erabilera sustatzeko asmoz. Honek langileen autonomia areagotzen du eta autoikasketa bultzatzen lagunduko luke.

ELET aplikazio malgua da. Moduluka antolatzen da eta hainbat baliabide linguistiko eskaintzen ditu. Horien artean:

- Hiztegiak: Elhuyar Fundazioko *Euskara-Gaztelania/castellano-vasco* *Elhuyar Hiztegia*, *Euskara-Frantsesa/français-basque* *Elhuyar Hiztegia*,

Euskara-Ingelesa/English-Basque Elhuyar Hiztegia, Hiztegi Modernoa eta Sinonimoen Kutxa hiztegietan bilaketa bakar batean bilatu ahal izateko modua, hain zuzen.

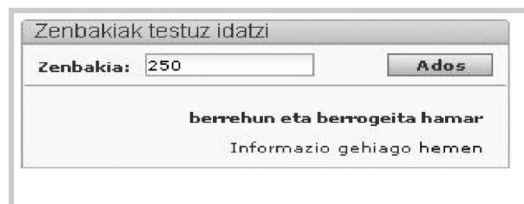
- Gramatika: Gramatikako modulu honek hainbat edukietan bilaketak egitea ahalbidetzen du eta iturrietara joateko estekak eskaintzen ditu: HABEko deklinabide taulak, adizkien paradigma, idazkera erregelak, Euskaltzaindiko *Hiztegi Batua*, etab.

- Deklinabidea: Modulu honetan erabiltzaileak lema bat idatzi, deklinabide kasu bat hautatu edota lema horren forma deklinatu guztiak ikusi ahal izango ditu.

- Aditzen jokatzeari: Modulu honetan aditza jokatzeko laguntza eskainiko da. Kontsulta bi erataraino egin ahal izango da. Batetik, erabiltzaileak pantaila batean aditzaren modu/denbora eta pertsona aukeratu ditu. Horietatik, ELETEk aditz jokatu egokia proposatuko du. Bestetik, adizki bat aztertzeko aukera eskaintzen du ELETEk. Aukera honetan erabiltzaileak adizki bat testu-kutxa batean idatzi eta honen analisia lortuko du, zalantza kasuarentzat lagungarria izango dena.

- Datak eta orduak nola idatzi: Modulu honetan zenbakizko data edo ordu bat idatzi eta sistemak zenbakitik letrara bihurtuko du, adibide batzuk gehituko direlarik. Emaizaren leihuan gai honi dagokion Euskaltzaindiako araura joateko esteka ere eskainiko da.

- Zenbakiak testuz idatzi: Zenbakiak letraz idaztea tokatzen denean, zalantza ugari sortu daiteke. Modulu honen bitartez, erabiltzaileak zenbaki bat testu-kutxa batean sartu eta botoi bat sakatuz, letraz idatzitako ordaina jasoko du. Horretaz gain, Euskaltzaindiako zenbakiak idazteko araua kontsultatzeko esteka bat aurkituko du.



5. irudia. Zenbakiak testuz idatzi modua

- ZalaNet: Hizkuntza-zalantzak argitzeko sistema bat da. Datu-base moduan moldatuta dago eta galdera erantzun moduan antolatzen da informazioa. Informazioa eskuratzea galdera ireki baten bidez egin ahal izango da. *ZalaNet* modulu bizi moduan planteatzen da. Izan ere, datu-basean gordetzen den informazioa etengabe hazten joatea aurreikusten da. *ZalaNetek*, jatorrizko datu-basean aurkitzen ez diren dudak eta zalantzak argitzeko, aditu bati galdera egiteko aukera eskainiko du eta, adituak egokia baderitzo, galdera eta erantzun horiek datu-basean txertatuko dira.

MIATU euskaraz idatzitako web guneak indexatzeko bereziki prestatutako indexatze eta bilaketa tresna da. Horrek ez du esan nahi MIATU beste hizkuntzetan idatzitako dokumentuak indexatzeko gai ez denik. Izan ere MIATUk gaztelera, ingelesa eta frantseserako ere ezaugarri aurreratuak eskaintzen ditu eta beste edozein hizkuntza ere modu arruntan kudea dezake.

MIATUk lematizazioa erabiltzen du indexatu behar diren hitz formen lemak lortzeko eta bilaketak lehen gainean egiteko aukera ematen du.

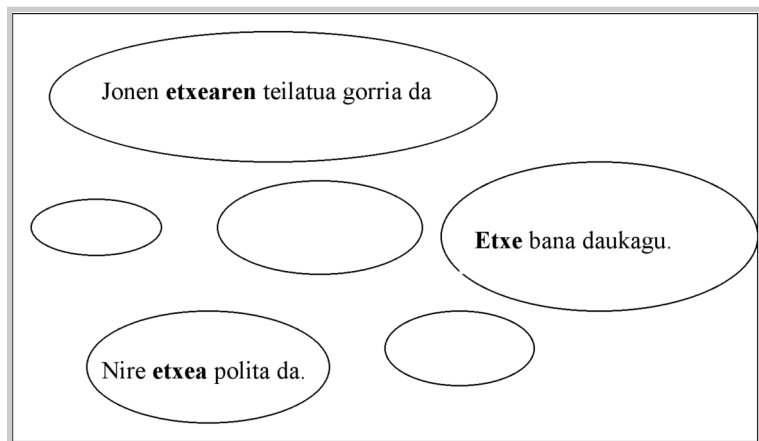
Noski, hala nahi bada posible da MIATUri forma jakin baten agerpenak bakarrik bilatzeko eskatzea eta MIATUk hala egingo du.

Badira ildo honetan sortu diren beste prototipo batzuk. EHUko IXA Taldearen teknologietan oinarrituta, hizkuntza-ariketak automatikoki sortzeko tresna garatu da, idazten laguntzeko aplikazioa (ELETen aipatu diren modulu batzuk integratzen dituena) eta idazlanak automatikoki ebaluatzeko sistema, besteak beste. Ariketak automatikoki sortzen dituen tresnak hutsuneak bete, okerrak zuzendu eta ikasleak erantzuteko galderak sortzen ditu. Idazlanak ebaluatzeko aplikazioak berriz, hitzen zuzentasun ortografikoa, hitz kopuru totala, kategoria jakin bateko hitz kopurua totalarekiko eta beste irizpide batzuetan oinarritzen da. Guzti hauek noski, pertsona ordeztu baino, erabiltzaileei laguntza eskaintzea dute helburu.

Teknologia berrien garapenean Internetek izugarritzko eragina izan du. Ondoren, Internet oinarri edo baliabide bezala hartzen duten aplikazio batzuk azalduko dira.

INTERNETEKO BILATZAILE AURRERATUAK: MIATU.

MIATU euskaraz idatzitako web guneak indexatzeko bereziki prestaturiko indexatze eta bilaketa tresna da. Horrek ez du esan nahi MIATU beste hizkuntzetan idatzitako dokumentuak indexatzeko gai ez denik. Izan ere MIATUk gaztelera, ingelesa eta frantseserako ere ezaugarri aurreratuak eskaintzen ditu eta beste edozein hizkuntza ere modu arruntan kudea dezake.



6. irudia. Dokumentu multzoa

Esan bezala MIATU euskararen berezitasunei aurre egin eta etekina ateratzeko prestatuta dago. MIATUk lematizazioa erabiltzen du indexatu behar diren hitz formen lemak lortzeko eta bilaketak lehen gainean egiteko aukera ematen du. Honek zer esan nahi du? Irudiko dokumentu multzoan hiru dokumentu agertzen dira besteak beste. Hiru dokumentu

hauek “*etxe*”ren formaren bat dute baina “*etxe*” hitza zehazki bere edukian duen bakarra “*Etxe bana daukagu*” dokumentua da. Bilatzaile arruntetan “*etxe*” hitzaren gaineko bilaketa egin ezker, “*Etxe bana daukagu*” dokumentua bakarrik eskuratuko genuke. “*Etxearen*” bilatu ezker “*Jonen etxearen teilatua gorria da*” bakarrik lortuko genuke. “*Etxeetatik*” edo “*etxeen*” bilatu ezker berriz, ez litzateke emaitza bakar bat bera ere lortuko. MIATUk arazo hau gainditzen du: bilaketa egiteko “*etxe*”ren edozein forma erabiltzen dela ere, emaitzetan forma guztiak agertuko dira modu lehenetsian. Hau da, adibidean “*etxe*” bilatuta hiru dokumentuak lortuko dira. “*etxeetakoen*” bilatzen bada ere berdin, nahiz eta hitz hori dokumentuetako batean ere ez den agertzen.

Nabarmena da funtzio hau oso erabilgarria dela euskararako. Noski, hala nahi bada posible da MIATUri forma jakin baten agerpenak bakarrik bilatzeko eskatzea eta MIATUk hala egingo du. Funtzio aurreratu hau beste hizkuntzetan ere erabil daiteke baina kasu gehienetan abantaila ez da euskararen kasuan bezain nabarmena izango.

Edozein web gunetan integratzeko aukera eskaintzen du MIATUk eta integrazio hori egiteko modua erabakitzeko malgutasuna ematen dio web garatzaileari. Bilatzaile honen doako bertsio bat Hizkuntza Politikarako Sailburuordetzaren web gunean topa daiteke deskargagarri: http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_softwarea/eu_9567/web_lib.html

Euskarazko edukien gaineko bilatzailea Interneten: Elebila.

Elebilak (www.elebila.eu) Interneteko bilatzaile ezagunenen funtzioak eskaintzen ditu baina kalitatezko euskarazko emaitzak lortzeko bermearekin. Googlen edota Yahooren euskarazko termino bat bilatu daiteke baina normalean emaitzetan euskarazko testuak ez dira egongo lehenengo postutan.

Euskarazko bilaketak egiteko orduan bi arazo nagusi daude: alde batetik euskarak Interneten duen garrantzia erlatibo txikia eta bestetik atzizkien erabilera zabala.

Euskarak sarean duen pisua txikia dela-eta, beste hizkuntzetan grafia bera duten hitzak bilatzen badira, “*anorexia*” edo “*energia*” adibidez, euskaraz idatzitako dokumentuak inoiz ez dira lehen emaitzen artean agertzen. Eta berdin gertatzen da “*Newton*” edo “*Egipto*” bezalako izen bereziekin, eta “*katu*” edo “*esne*” bezalako hitz laburrekin ere. Elebilak arazo hau konpontzen du.

Bestalde, euskarak oso deklinabide aberatsa duenez, bilatzaile arruntek ematen dituzten emaitzak inoiz ez dira osoak: “*energia*” bilatzen bada hitz zehatz hori duten dokumentuak agertzen dira soilik, eta “*energiaren kontsumoa handitu da*” moduko esaldiak ez dira emaitzan agertuko. Elebila arazo honi aurre egiteko prestatuta dago, lehenago Miaturen kasuan aipatu den antzera lematizazioa eta sorkuntza erabiltzen dituelako.

Googlen edota Yahooren euskarazko termino bat bilatu daiteke baina normalean emaitzetan euskarazko testuak ez dira egongo lehenengo postutan.

Elhuyar Fundazioko I+G sailak garatutako Corpeusek (www.corpeus.org) interneteko euskarazko edukiak corpustzat hartzen ditu.

**Corpusekin lotuta,
Elhuyar Fundazioak eta
EHUko IXA Taldeak
Zientzia eta
Teknologiaren Corpusa
landu dute
(www.ztcorpusa.net). ZT
corpusa zientzia eta
teknologiaren alorreko
euskarazko testu-bilduma
egituratu eta etiketatua
da eta alor horietako
euskararen erabilera
ikertzeko baliabidea
izatea du helburu
nagusia.**



7. irudia: Elebilako emaitza pantaila

Elebila Elekak garatu du Elhuyar Fundazioko I+G sailak garatutako teknologian oinarrituta, CorpEusek erabiltzen duen bera (www.elhuyar.org).

INTERNET EUSKARAZKO CORPUS GISA: CORPEUS.

Elhuyar Fundazioko I+G sailak garatutako Corpeusek (www.corpeus.org) interneteko euskarazko edukiak corpustzat hartzen ditu. Hauek dira burutzen dituen pausuak:

- Erabiltzaileak aztertu nahi dituen hitza edo hitzak sartzen ditu
- Bilatzaileen APIak erabiliz, eskatutako hitzen formak dituzten euskarazko orrien zerrenda lortzen da
- Orriak jaisten dira hainbat prozesu paraleloren bidez
- Orriak iritsi ahala, eskatutako hitzen agerpenak erakusten dira
- Amaieran, eskatutako hitzen agerpen-kopuruak erakusten dira

Hori guztia egin ondoren erabiltzaileak bilatutako hitzaren forma guztiak zerrendatuta topatuko ditu testuinguruarekin batera. Hitz multzoak ere bila daitezke edota hitz baten kategoria jakin bat duten analisiak.



8. irudia: Corpeuseko emaitza

Corpusekin lotuta, Elhuyar Fundazioak eta EHUko IXA Taldeak **Zientzia eta Teknologiaren Corpora** landu dute (www.ztcorpusa.net). ZT corpora zientzia eta teknologiaren alorreko euskarazko testu-bilduma egituratu eta etiketatua da eta alor horietako euskararen erabilera ikertzeko baliabidea izatea du helburu nagusia. Bertan, Corpeusen bezala, eskuz etiketatutako edukien gainean bilaketak egin daitezke eta testuinguruan ikusi.

Zaintza Teknologikoa: Xerlok. Xerloken helburua honakoa da: enpresa edo erakunde jakin baten inguruko informazioa interneteko eguneroko prentsan (euskaraz, gaztelaniaz, eta beste edozein hizkuntzatan) non eta zertan agertzen den automatikoki aztertu eta informazio hori kudeatzeko zerbitzua ematea. Zerbitzu honek honako informazioa bilatzen du:

- Erakundearen inguruko informazioa prentsan (euskaraz, gaztelaniaz,...) bilatu, eta aurkitutako artikuluekin datu-base bat osatzen du. Informazio horretan erakunde horren inguruan agertutako artikuluguztiak jasotzen dira.
- Erakundearen sektore berdineko edo lehiakide izan daitezkeen beste enpresa edo erakundearen prentsako agerpenen jarraipena.
- Teknologiaren jarraipena. Erakundeak adierazitako hitz-gakoen zerrendaren arabera, berari dagokion teknologiaren jarraipena.

Zerbitzu honen bidez, egunero jasotako artikuluguztiak behar bezala sailkatu eta datu-base batean gordetzen dira. Ondoren, datu-base hori honela ustiatzen da:

- Eguneroko txosten bat egiten da prentsako agerpenekin eta txosten hau erakundearen arduradunari bidaltzen zaio.
- Web gune batean eguneroko agerpenak kontsulta daitezke, eta baita prentsako agerpen historikoak ere. Web gune honetan informazioa

Xerloken helburua honakoa da: enpresa edo erakunde jakin baten inguruko informazioa interneteko eguneroko prentsan (euskaraz, gaztelaniaz, eta beste edozein hizkuntzatan) non eta zertan agertzen den automatikoki aztertu eta informazio hori kudeatzeko zerbitzua ematea.

Dokusare: Interneteko berriak lotzeko baliabidea.
Era horretan, irakurleak albisteak irakurri ahala bere aurrean duen albistearekin erlazionatutako beste albiste batzuk eskuragarri lituzke hizkuntza eta egunkari desberdinetakoak direla ere.

bilatu ahal izango da ondoko eremuen arabera (eremu hauek konbinatu ahal izaten dira, bilaketa aurreratuak eginez):

- **Dataren arabera.** Data jakin bateko artikulua, edo data bate-tik beste data batera bitartekoak, etab.
- **Medioen arabera.** Medio jakin batean izandako agerpenak.
- **Hitz-gakoen arabera.** Hitz gakoak erabili daitezke, hauen inguruko artikulua zehatzak aurkitzeko.
- **Testu librea.** Bilaketen emaitza hobetzeko lematizazioa erabil-tzen da. Lematizazioak horrelako sistema baterako duen helburua zera da, hitzaren agerpenak deklinatuta daudenean ere aurkitzea. Adibidez, euskaraz “*eduki*” bilatzean “*genuen*” ere aurkitzea.

Dokusare: Interneteko berriak lotzeko baliabidea. Gaur egun, Interneten argitaratzen den informazioa, blogetako sarrerak, albisteak, eztabaidaguneak eta abarrek ez daukate inolako elkarren arteko erlazio-rik, hau da, erabilera askori begira ez dago ondo antolatuta. Horren ondorioz, informazioa ikuspegi semantiko orokor batetik atzitu nahi dugunean ezinezkoa gertatzen zaigu.

Dokusarek problema horri aurre egiten dio, egunkari digitalen espar-ruan. Esparru zabal eta emankorra non informazioaren nolabaiteko antolaketak prentsa irakurtzeko modu berriak eta aberatsagoak eskaini ditzaken.

Sistemaren helburu zehatza egunkari digital batean argitaratutako albisteak beste egunkari digital batzuetako albisteekin (elkarren antzeko-tasun mailaren arabera) lotzea da. Era horretan, irakurleak albisteak ira-kurri ahala bere aurrean duen albistearekin erlazionatutako beste albiste batzuk eskuragarri lituzke hizkuntza eta egunkari desberdinetakoak direla ere. Funtzio honen abantailak ugariak dira: informazioa kontrasta-tzeko aukera, albiste bateko gai batean sakontzeko posibilitatea...

Sistema lau hizkuntza desberdinetako albisteak prozesatzeko gai da: Euskara, gaztelania, frantsesa eta ingelesa.

Arakutzen dituen agerkari digitalen artean honakoak daude: *Sud-ouest, Gara, DV, Berria, Argia, Noticias de Gipuzkoa*, ... eta urruneko beste batzuk *Le Monde, The Guardian*, ...

Sistemak administratzaileak ezarritako zerrenda bateko egunkari digitaletara periodikoki konektatu eta albisteentzakoak identifikatzen ditu. Elkarren antzekoak sistemaren datu-basean gordetzen dira, aplika-zioa instalatuta dagoen web-gunetik kontsultagai egon daitezen momen-tu oro SOAP bidez.

Internet alde batera utzi eta itzulpengintzarekin zerikusia duten tres-nen azterketara pasata, ondorengo adibideak topa daitezke:

ENTITATEAK ERAUZTEKO TRESNAK: ENTIGARBI.

Zenbait kasutan dokumentuak interneten denen eskura jartzeko beharra sentitzen da. Hala ere, ez da hain erraza, dokumentu hauetan sarean barreiatu nahi ez den informazioa egon baitaiteke. Bestetik, administrazio publikoetan itzulpenak egiterakoan erabiltzen diren itzulpen-memoriak ere publiko jartzeko asmoa agertu izan da. Itzultzen diren dokumentu hauetan ere informazio konprometitua ageri ei da eta beraz, komenigarria izan daiteke sarean jarri aurretik itzulpen-memoria horietatik ere informazio pribatu hori kentzea.

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

9. irudia: Ingeleseko entitateak detektatzen dituen tresna baten emaitza

Aipatu den informazio konprometitu hori, hau da entitatea, izen berezia, NAN/NIF zenbakia, gizarte segurantzaren zenbakia...izan daiteke. Entitateak ezabatzeko modu bat testu guztia irakurri eta eskuz ezabatzea litzateke. Hau modu neketsua da eta arrazoi honengatik eta baliabide faltagatik, askotan ez dira erakunde publikoetako itzulpen-memoriak publikatzen. Elekak tresna linguistikoak erabiliz entitateen detekzio automatikoan laguntzen duen tresna bat garatu du: Entigarbi. Elhuyar Fundazioko I+G atalean garatutako teknologietan oinarrituta, itzulpen-memoretan ager daitezkeen pertsonen izenak, NAN zenbakiak eta bestelako informazio pribatua ezkutatze bali du. Datuen babes-legea indarrean dagoenetik tentu handiz ibili beharra dago pertsonen datu pribatuak publiko ez egiteko. Entigarbi dokumentu elebarkar nahiz elebidunen entitateak markatzen laguntzen du, ondoren landu eta dokumentu garbi bat modu erosoan lortzeko.

Terminologia erauzketa automatiko eleanitza. Goiko atalean aipatu denaren ildoan, bada itzulpen-memoretatik atera daitezkeen informazio erabilgarria. Itzulpen-memoriak esaldi bikote elebidunen bilguneak direla esan liteke. Askotan erakunde baten terminologia gordailutzat ere har litezke memoria horiek. Hizkuntza-teknologiaren laguntzarekin, esaldi bikoteak diren itzulpen-unitate elebidun horiek aztertu eta termino bikoiteak identifikatu daitezke, glosarioak sortzeko adibidez. Eskuz egiteak sekulako lana eskatuko luke baina tresna informatiko baten laguntzaz

Elekak tresna linguistikoak erabiliz entitateen detekzio automatikoan laguntzen duen tresna bat garatu du: Entigarbi. Elhuyar Fundazioko I+G atalean garatutako teknologietan oinarrituta, itzulpen-memoretan ager daitezkeen pertsonen izenak, NAN zenbakiak eta bestelako informazio pribatua ezkutatze bali du.

*Azken urteetan
ezagunagoak bihurtu
diren tresnak dira
itzultzaile automatikoak.*

*Opentrad sistemarekin
adibidez, gaztelania eta
galegoaren artean egiten
diren itzulpenek ez dute
eskuzko errebisiorik
beharko. Opentrad kode
irekiko sistema
eleaniztuna, librea eta
doakoa da.
(www.opentrad.org)*

erauzketa automatikoa egin daiteke eta web aplikazio batean termino zerrenda kudeatzeko aukera izango du erabiltzaileak modu eroso batean. Glosario hauek erakundeetako terminologia zaindu eta idatzietako koherentzia bermatzen laguntzen dute.

Itzultzaile automatikoak. Azken urteetan ezagunagoak bihurtu diren tresnak dira itzultzaile automatikoak. Lantzen duten hizkuntza parearen arabera, lortu daitezkeen emaitzak onargarriak edo oso onak izan daitezke. Opentrad sistemarekin adibidez, gaztelania eta galegoaren artean egiten diren itzulpenek ez dute eskuzko errebisiorik beharko. Opentrad kode irekiko sistema eleantiztuna, librea eta doakoa da. (www.opentrad.org)

Estatuko hizkuntza guztietarako itzulpen automatikoko sistema bat da Opentrad, kode irekikoa.

Gaur egun, bi noranzkoko itzulpenak egin daitezke gaztelania-galegoa eta gaztelania-katalana hizkuntza-pareetan, baita gaztelania-euskara pareko itzulpenak ere.

Bestalde, erabiltzaileen beharren arabera egokitzeko eta konfiguratzeko diseinatuta dago eta beste sistema batzuekin integratzeko aukera ematen du.

Gaur egun, milaka erabiltzaile ari dira Opentrad erabiltzen testuak, dokumentuak edo web orriak itzultzeko, Internet bidez nabigatzean orriak itzuliz. Orri pertsonalak, enpresakoak, blogak eta abar galegoz, gaztelaniaz, katalanez eta euskaraz eskaintzeko erabiltzen ari dira sistema beste erabiltzaile batzuk.

Opentraden diseinuari esker, sistema egokitu edo pertsonalizatu egin daiteke, bai hizkuntzaren ikuspegitik erakunde baten estilo-liburua kontuan hartuta, bai teknikaren ikuspegitik IP sareetan integratzen utzita edo beste sistema batzuekin erabat integratuta.

imaxin|softwarek, Galiziako enpresak (<http://www.imaxin.com/>) imaxin|context hizkuntza-ingeniaritzako sailaren bidez, balio erantsiko zenbait zerbitzu eskaintzen ditu, Opentrad norberaren beharren arabera pertsonalizatzeko eta integratzeko.

Kode irekiko itzulpen automatikoko sistema bat da Opentrad, eta, horri esker, partzuergoak berak emandako bertsioez eta boluntarioek egindako ekarpenez balia daitezke erabiltzaileak.

Gaztelania-galegoa eta gaztelania-katalana pareetako itzulpenak Apertium sistemak egiten ditu gaur egun, eta erlazionatuta dauden beste hizkuntza-pare batzuetako itzultzaile automatikoak sortzeko erabil daiteke (daniera-suediera, txekiera-eslovakiera, etab.). Gaztelania-euskararako, berriz, IXA Taldeak sortutako Matxin motorek erabiltzen du.

Es-ca eta es-gl pareentzat orain arte izan diren gainerako itzulpen automatikoko sistema gehienak komertzialak dira, eta teknologia pribatiboa erabiltzen dute; beraz, oso zaila da erabilera berriei egokitzea. Gainera, teknologia desberdina erabiltzen dute hizkuntza-parearen arabera;

hori dela eta, oso zaila da eduki eleaniztunak kudeatzeko sistema bakar batean integratuta edukitzea.

Opentradek eskaintzen duen arkitekturaren nobedade garrantzitsuenetako bat kode irekiko lizentzia baten arabera liberatua izatea (GNU GPL lizentzia programetarako, eta Creative Commons, datuetarako) eta doakoa izatea da. Beraz, behar adina hizkuntza- eta informatika-ezagutza dituen pertsona orok egokitu edo zabaldu ahal izango du produktua, itzulpen automatikoko sistema berri bat sortzeko, baita erlazionatutako hizkuntza-pare berrietan ere.

Ordenagailuz Lagundutako Itzulpen-tresnak (OLI tresnak): Wordfast, Trados, OmegaT...

Itzultzaile automatikoko tresnek lehenago aipatu den bezala, hizkuntza pare batzuetan lortu ditzaketen emaitzak oso onak dira. Beste hizkuntza pareetan, berriz, ez dira hain onak. Hori gertatzen denean eta itzulpen lan baten aurrean egonda, itzultzaileak OLI tresnen laguntzara jo dezake. OLI tresnek itzultzaileak lantzen dituen esaldi pareak gordezten dituzte, hurrengo batean antzeko eduki bat itzultzea suertatzen zaie-nean berrerabili ahal izateko. Horrela, esaldi bikoteek (itzulpen-unitate-ek) itzulpen-memoriak sortzen dituzte eta esan bezala, dagoeneko itzuli diren esaldi bikoteen gordailuak dira.

Normalean OLI tresnak erabiltzailearen hizkuntzatik independente-ak dira eta edozein hizkuntza-pare itzultzeko balio dute. Horren adibide dira Wordfast, SDL Trados edota OmegaT (<http://www.wordfast.net/index.php?lang=eu01>, <http://www.sdl.com/>, <http://www.omegat.org/es/omegat.html>).

OLI tresnek beste funtzionalitate batzuk ere eskaintzen dituzte. Glo-sarioak kudeatzeko aukera, testuinguru bilaketak, zuzenketa ortografi-koa, etab.

Artikulu honetan aipatu diren teknologia batzuk egunero erabiltzen ditu euskaraz lan egiten duen edozein pertsonak eta askotan, ohartu gabe, lana pixka bat arinago egiten laguntzen diote.

Artikulu honen helburua hizkuntzaren tratamendurako tresna batzuk aurkeztea izan da euskarari lehentasuna emanez. Lan asko dago oraindik egiteko alor honetan eta jende asko dihardu horretan hain zuzen, denon bizitza pixka bat erosoagoa egiteko, betiere teknologia berrien laguntzaz.■

Gaur egun, milaka erabiltzaile ari dira Opentrad erabiltzen testuak, dokumentuak edo web orriak itzultzeko, Internet bidez nabigatzean orriak itzuliz. Orri pertsonalak, enpresakoak, blogak eta abar galegoz, gaztelaniaz, katalanez eta euskaraz eskaintzeko erabiltzen ari dira sistema beste erabiltzaile batzuk.