01
TEKNOLOGI BERRIAK ETA EUSKARA: EGOERAREN AZTERKETA
>>>

HIZKUNTZA-TEKNOLOGIA IXA TALDEAN, EUSKARAREN ERABILERA ERRAZTEKO ETA SUSTATZEKO APLIKAZIOAK

Iñaki Alegria, Xabier Artola, Arantza Diaz de Ilarraza eta Kepa Sarasola
IXA taldea (UPV-EHU)
Helbide elektronikoa: kepa.sarasola@ehu.es

SARRERA

Gero eta informazio gehiago dugu eskura modu digitalean: bideo, irudia, ahotsa eta testua. Tresna eta aplikazio informatiko berriak sortu dira informazio andana hori prozesatu ahal izateko. Baina non dago kokatuta euskara mundu horretan? Lagun dezakete tresna horiek euskararen normalizazioan?

Gero eta informazio gehiago dugu eskura testu moduan

Ikaragarria da ikustea azken 20 urtean zelan aldatu diren konputagailuak eta Interneteko informazioa lortzeko eta sortzeko modua. Ikaragarria da era digitalean eskura dugun testu-masaren tamaina; Interneten ingelesez bilioi bat hitz edo dagoela estimatzen da (miloi bat miloi hitz!). Euskaraz, ingelesez baino mila aldiz edo hamar mila aldiz gutxiago omen dagoenez[1], gutxi gora behera mila miloi hitz izango dira sarean orain. Tamaina horiek zein handi diren erakusteko nahiko da jakitea liburu normal batean 100.000 hitz inguru sartzen direla, pertsona kultu batek egunean 10.000 hitz edo irakurtzen duela, urtean 3,65 milioi hitz, eta 300 milioi

bere bizitza osoan. Beraz, beldurrik gabe esan dezakegu gaur egun segundo batzuen buruan euskaraz eskura dezakegun testu guztia irakurri nahiko bagenu 3 aldiz edo bizi beharko ginatekeela, edo 3.000 aldiz bizi beharko ingelesez dagoen testua irakurri ahal izateko. Harrigarria, ez da? Horrela, noski, hainbat testu eskura izanda, errazagoa da behar dugun informazioa aurkitzea.

Bestalde, epe ertainean pertsona eta makinen arteko komunikazioa hainbat aplikaziotan geure hizkuntzan egin ahal izango dugu, ez makinaren hizkuntzan. Tresna mugatuak izango dira, eta beti errore-maila batekin, baina, hala ere, laguntza ederra emango digute.

Euskaraz, ingelesez baino mila aldiz edo hamar mila aldiz gutxiago omen dagoenez , gutxi gora behera mila miloi hitz izango dira sarean orain. Tamaina horiek zein handi diren erakusteko nahiko da jakitea liburu normal batean 100.000 hitz inguru sartzen direla, pertsona kultu batek egunean 10.000 hitz edo irakurtzen duela.

Hizkuntza automatikoki lantzeko tresnak errealitatea dira

Gaur egun badira testua edo hizketa lantzeko zenbait hizkuntza-aplikazio eskuragarri, hala nola: ortografia-zuzentzaileak, estilo-zuzentzaileak, hiztegi-kontsultak on-line, itzulpen automatikoa eta itzulpen-laguntzak, hizketa testua bihurtzen duten sistemak, testuak irakurtzen dutenak, bigarren hizkuntza ikasteko sistemak, aplikazio informatikoak gure hizkuntzan erabiltzeko interfazeak, Galderetarako erantzunak bilatzeko sistemak (Question Answering), dokumentu-bilatzaileak (IR, Information Retrieval), informazio-erauzketa dokumentuetatik (IE, Information Extraction), laburpen automatikoa (Summarization), dokumentu-sailkatzaileak, dokumentu-bideratzaileak (Routing), dokumentu-multzokatzaileak (Clustering), dokumentu-iragazleak (Filtering) edo testu-sorkuntza automatikoa.

Tamalez ez dago Interneten gune bakarra hizkuntzaren prozesamendurako produktu guztien berri biltzen duenik. Hala ere, arloka edo aplikazioaren arabera antolatuta indarrean dauden hainbat produktu honako gunetan aurki daitezke:

- Natural Language Software Registry[2] (NLSR).
- Hizkuntzalaritza konputazionaleko demoak on-line (Interactive online CL Demos[3])
- European Language Resources Association[4] (ELRA), Batez ere h Europako hizkuntza- baliabideak biltzen ditu (corpus eta lexikoiak).
- Linguistic Data Consortium[5] (LDC), Aurrekoaren parekoa, baina Amerikako Estatu Batuetako produktuetan espezializatua.
- Hizkuntzalaritza konputazionalerako elkarteko wikia[6] (ACL, Association for Computational Linguistics). Hizkuntza guztietarako baliabideen berri jasotzeko gunea zabaldu berri dute lehengo urtean.
- Yourdictionary.com[7]: Hiztegi-kontsultak on-line eta itzulpen automatikoko doako zerbitzuak. 300 hizkuntzarako zerbitzuak daude hor. Argi dago baina munduko hiztegi-zerbitzu guztiak ez daudela, euskararako daudenak ez dira hamarrera ailegatzen eta www.hiztegia.net

Tamalez ez dago Interneten gune bakarra hizkuntzaren prozesamendurako produktu guztien berri biltzen duenik.

http://liceu.uab.es/~joaquim/

gunean 50 baino gehiago bildu dituzte. Beste hizkuntza guztiekin berdin gertatzen bada, existitzen diren hiztegi-sistemen kopurua askoz handiagoa izan daiteke.

• Itzulpen automatikoko sistemak eta sareko hainbat zerbitzuren berri biltzen dira gune hauetan: Translation Directory[8] eta Traduzione e computer[9].

• Hizketako produktuen berri zabala Joaquim Llisterri[10] ikerlari katalanaren web orrietako esteketan aurkitu daiteke.

Baina hizkuntza nagusientzat dira aurrerapen gehienak

Baina horrelako aplikazio guztiak ingelesez erabili ahal badira ere, beste hizkuntzetarako ezin da berdin esan, kuantitatiboki eta kualitatiboki. Ondoko 1. taulan ikus daiteke zenbat produktu jaso diren zenbait hizkuntzarentzat, hizkuntza teknologiako produktuen berri ematen duten hiru gunetan.

	ELRA	NLSR	LDC
Ingelesa	463	196	232
Alemana	428	106	14
Frantzesa	407	99	10
Espainiera	388	85	28
Italiera	349	76	2
Arabiera	49	0	57
Nederlandera	46	69	3
Suediera	29	67	2
Daniera	18	64	1
Katalana	9	59	0
Euskara	4	61	0

1. taula. Hizkuntza-teknologiako produktuen kopurua zenbait hizkuntzatarako

Argi dago ingelesa dela nagusia. LDC elkarteko datuetan itzela da diferentzia, berau Amerikako Estatu Batuetako elkartea delako agian. ELRA (Europako Batzordearen babesa dute) eta NLSR (Alemaniako DFKI ikerketa-taldearena) biltegietan beste hizkuntzetarako produktu gehiago jasota daude, baina horietan ere ingelesa da nagusia. Beste hizkuntza nagusiek azken urteetan hainbat baliabide garatu dute eta gertutik jarraitzen diote ingelesari. Baina, beste hizkuntzek ahalegin handia egin behar dute atzean ez gelditzeko, are gehiago euskara bezalako hizkuntza txikiek.

Bigarren taulan ikus daiteke zenbat itzulpen-sistema zeuden Europako hizkuntza ofizialen artean 2005. Argi dago aldea hizkuntzen artekoa. Eta horiek guztiak ofizialak dira:

	en	de	fr	es	it	pt	du	po	lt	gr	cs	hu	sw	fn	sl	rm	dk
Ingelesa		47	41	44	30	30	10	8	2	1	4	1		1	1		
Alemana	48		24	8	10	4	2	3	1	1	2	1	1	1		1	
Frantsesa	40	23		11	13	8	4	1	1	1							
Espainiera	41	7	11		9	8	1		1	1							
Italiera	29	10	13	9		4	1		1	1							
Portugesa	29	5	7	8	4		1		1								
Holandesa	10	2	4	1	1	1			1								
Poloniera	7	2	1														
Lituaniera	2	1	1	1	1	1	1										
Greziera	3		3														
Txekiera	1	1	1		1												
Hungariarra	2	2															
Suediera	2	1															
Suomiera	2	1															
Eslovakiera																	
Errumaniera	1																
Daniera		1															

2. taula. 2005eko itzulpen-sistemen kopurua Europako hizkuntza ofizialen bikoteetarako[11]

Hizkuntza-teknologia helburu eta lagun euskara normalizatzeko bidean

Mendeetako erregresio-prozesu batean sartuta ibili da euskara. Amorrortu-ren (2002) arabera horren arrazoi nagusiak hauek izan dira: batetik hizkuntza ofiziala ez izatea, eta bestetik hezkuntza sistematik kanpo, komunikabideetatik kanpo, eta industri guneetatik kanpo egotea. Gainera hainbat euskalki diferente izateak ez zuen laguntzen euskara idatziaren zabalkuntzan.

Azken hamarkadetan, baina, urrats kualitatibo oso esanguratsuak egin ditugu egoera horri buelta emateko. Berpizkunde moduko hori honako urratsetan nabaritzen da:

Euskara hizkuntza koofiziala da Hegoaldean (Nafarroa osoan ez baina).

Hizkuntza-sisteman txertatua izan da Hegoaldean eta Nafarroako lurralde mistoan.

Euskarazko komunikabideak ditugu (EITB, Berria...)

Euskara estandarraren oinarria definitu zuen Euskaltzaindiak 1966an. Morfologia guztiz definituta dago orain, baina lexikoa oraindik ez. Eta euskara batua da irakaskuntzan eta komunikabideetan erabiltzen dena.

Egun 700.000 hiztun ditu euskarak, biztanleagoaren %25 gutxi gora behera, baina ahalegin guzti horiek eginda ere euskararen etorkizuna oraindik ez dago ziurtatuta. Aipatu urrats horiek guztiz orokorrak ez izateaz gain, euskara industri guneetatik kanpo jarraitzen du, Informazioa eta Komunikazioaren Teknologia (IKT) berriarekin lotuta dauden industri guneetatik ere bai.

Artikulu honetan aztertu nahi dugu hizkuntza-teknologiak nola lagundu dezakeen euskararen erabilera errazteko eta sustatzeko. Hasieran zenbait produktu aurkeztuko ditugu bide horretan lagungarri suertatu direnak. Gero euskarak orain informatikaren munduan duen egoera ikusiko dugu. Hirugarren puntuan hizkuntza-teknologia lantzeko IXA taldean dugun estrategiaren lehentasunak azalduko ditugu. Eta bukatzeko epe erdirako zenbait helburu finkatuko ditugu.

XUXEN zuzentzaile ortografikoak laguntza paregabea eskaintzen dio erabiltzaileari testuaren kalitatea hobetzeko eta forma estandarrekin ohitzen joateko apurka-apurka. Horrela esan dezakegu euskararen estandarizazio-prozesuaren aliatu indartsua da XUXEN programa.

IXA taldean, gure ibilbidearen hasieratik, saiatu izan gara Hizkuntzaren Teknologiaren arloan ikertzen eta ahal izan denean produktuak gizarteratzen, IKT arloan euskararen erabileran normalizazioa sustatzeko. Testuak errazago eta txukunago sortu ahal izateko, sarean edukiak zabaldu edo bilatu nahi dituenak tresna egokiak izan ditzan. Helburu horretan ere bere hizkuntzarekin errazago lan egin dezan eta norberaren hizkuntzarekin ere gozatu ahal izateko. Bide horretan gogoeta ugari egin ditugu taldean, gure indar mugatuei ahalik eta mozkin handiena atera ahal izateko. Gogoeta horren ondorioz praktikan ildo bati, estrategia bati, jarraitu izan diogu urteetan zehar.

Artikulu honetan aztertu nahi dugu hizkuntza-teknologiak nola lagundu dezakeen euskararen erabilera errazteko eta sustatzeko. Hasieran zenbait produktu aurkeztuko ditugu bide horretan lagungarri suertatu direnak. Gero euskarak orain informatikaren munduan duen egoera ikusiko dugu. Hirugarren puntuan hizkuntza-teknologia lantzeko IXA taldean dugun estrategiaren lehentasunak azalduko ditugu. Eta bukatzeko epe erdirako zenbait helburu finkatuko ditugu.

2. EUSKARAREN ERABILERA ERRAZTEKO ETA SUSTATZEKO PRODUKTUAK

Atal honetan Ixa taldearen uztako zenbait produktu aurkeztuko ditugu. Euskaldunari mundu digitalera hurbiltzeko laguntza izan nahi dute horiek, eta publiko orokorrean eskuetan daude egun, erabiltzeko edo kontsultatzeko. Euskararen morfologia gure hizkuntza nagusienarekin konparatuz hain diferente izanik, soluzio landuagoak asmatu behar izan ditugu zenbait aplikazio gure hizkuntzan erabili ahal izateko. Bada, aplikazio horiexek dira hemen jaso nahi izan ditugunak.

2.1 Zuzentzaile ortografikoa

Euskaldunak bere hizkuntzaz idatzi gura duenean zalantza ugari aurkitzen ditu. Batetik, eskolak toki guztietan oraindik idazteko gaitasuna bermatzen ez duelako, edo bestetik, belaunaldi zaharragoek euskaraz ikasteko aukerarik izan ez dutelako, sarritan euskaldunak badaki esaten hitz bat baina ez daki nola idatzi behar den batuaz; esate baterako, ondoko hitzen artean zein da batuaz erabili behar dudana arbola adierazteko? zuhaitz? zuhatz? zugaitz? zugatz? zuhaits? sugatx? Bestetik, euskara estandarraren definizioa berri xamarra denez, lexikoaren estandarizazioa oraindik bukatzear dagoenez, eta batzuetan estandarizazioan aldaketan sortzen direnez (esate baterako, hasieran eritzi, eta iharduera hitzak erabili behar ziren; gaur egun iritzi eta jarduera) beste hainbat duda sortzen dira.

Horrelakoetan XUXEN zuzentzaile ortografikoak (Aduriz et al., 1997) laguntza paregabea eskaintzen dio erabiltzaileari testuaren kalitatea hobetzeko eta forma estandarrekin ohitzen joateko apurka-apurka.

Horrela esan dezakegu euskararen estandarizazio-prozesuaren aliatu indartsua da XUXEN programa.

Dohainik jaitsi daiteke www.euskara.euskadi.net webgunetik. Bere erabilera orokortuz doala erakusteko esan daiteke gune horretatik 20.000 erabiltzailek jaso duela honezkero. Gainera azken urtean atera diren egokitzapen berriei esker XUXEN eskuragarriago dago; lehen Word editorearekin bakarrik erabil zitekeen, orain erraz jar dezakegu martxan Mozilla Thunderbird-ekin Interneten bidez edozein mezu edo inprimaki betetzen ari garela, Openoffice-ekin edo beste edozein aplikaziorekin testua zuzentzeko www.xuxen.com zerbitzarira jotzen badugu.

Espainiera, frantsesa edo ingeleserako zuzentzaileak baino dezente konplexuagoa da XUXEN, hitz posibleak askoz gehiago direlako, eta ondorioz, hitzen analisi morfologikoa egin behar delako.

Oraindik lexiko eta morfologiako erroreak baino ez ditu harrapatzen, baina hitz maila horretan oso praktikoa da. Sintaxiko edo estiloko zenbait errore harrapatzeko ikerketak egiten ari gara orain, eta epe motzean lehenengo bertsio bat zabalduko da.

Horrelakoekin erabiltzaileak erraz kontsulta dezake hitz bat hiztegi batean, nahikoa da kontsultatu nahi den hitzaren gainean klikatzea bere informazioa pantailan ikusteko leiho dinamiko baten bitartez.

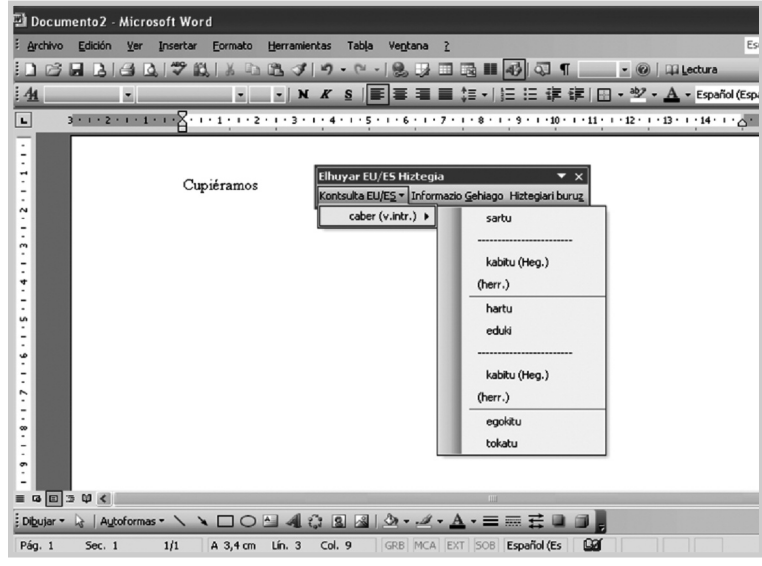
2.2 Lematizazioan oinarritutako hiztegien kontsulta edizioarekin integratua

Aplikazio mota honetan sortu diren produktuak Word editorerako plugin-ak dira. Horrelakoekin erabiltzaileak erraz kontsulta dezake hitz bat hiztegi batean, nahikoa da kontsultatu nahi den hitzaren gainean klikatzea bere informazioa pantailan ikusteko leiho dinamiko baten bitartez. Noski, morfologia aberatsa duen euskara bezalako hizkuntza baterako kontsulta askoz erosoagoa da lematizazio automatikoa egiten bada, bestela atzizkiren bat duen hitz bat bilatzerakoan ez zen ezer aurkituko. Lematizazioa espainieraz ere lagungarria izaten da; hori ikus daiteke ondoko irudian, esate baterako, erabiltzaileak *cupiéramos* hitza aztertu nahi duela. Hitz hori ezin daiteke aurkitu paperezko hiztegi batean, adizki jokatu bat duelako, baina plugin-ari esker programak *caber* aditzaren forma bat dela dakienez, erakutsiko dizkigu caber hitzaren euskarazko ordainak (kabitu, sartu...). Noski, euskarazko hitzak kontsultatzerakoan ere lematizazioa oso lagungarria da, eta adibidez tentelenei hitzaren sinonimoak bilatzen ditugunean kaikuenei hitza lortzeko gauza da; kasu horretan analisi morfologikoaren emaitza ez da bakarrik erabiltzen lema zein den jakiteko, hitz sinonimoa atzizki berdinekin eskaini ahal izateko ere erabiltzen baita. Beraz, hitz bat bere sinonimo batekin ordeztu nahi badugu, erraztasun bikoitza izango dugu plugin hau erabiliz gero, batetik ez dugu zehaztu behar zein den hiztegian bilatu behar den lema, eta bestetik atzizki berdinekin lortzen dugu sinonimoa automatikoki.

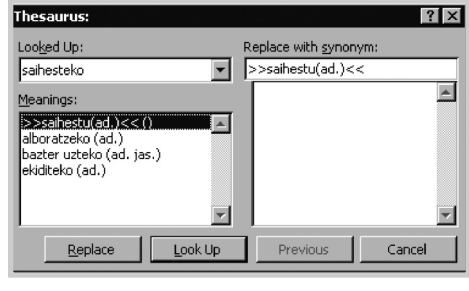
Sistema hau lau hiztegirekin erabil daiteke: UZEIren eta Elhuyarren sinonimo-hiztegiak, Elhuyarren espainiera-euskara, eta Elhuyarren fran-

Euskarazko testuetan hitz osoak bilatzea ez da bide oso zehatza, sarritan hitzetan atzizkiak azaltzen baitira; eta hitz-hasierak bakarrik bilatzen baditugu, horrelaxe hasten diren beste hitz luzeagoei dagozkien emaitzak ere azalduko zaizkigu, emaitzen kalitatea zapuztuz.

tses-euskara. Laster Elhuyarren ingeles-euskara ere erabili ahal izango da. Interneten bidez hiztegiak eta corpusak kontsultatzeko Euskalbar[12] aplikazioan integratu beharko zen lematizazio bidezko bilatzeko aukera hau epe motzean. Hain praktiko eta erabili bihurtu den bilaketa-barra hori are sendoagoa litzateke horrela.



1. irudia- Lematizazio bidezko hiztegi elebidunaren kontsulta.



2. irudia- Lematizazio bidezko sinonimo-hiztegiaren kontsulta.

2.3 Lematizazioan oinarritutako dokumentu-bilaketa

Dokumentuen berreskurapena deritzon aplikazio motan helburua da hainbat eta hainbat dokumenturen artean bakar bat (edo batzuk) hautatzea, kontzeptu bat edo informazio bat daukana. Noski, adibide tipikoena Interneterako bilatzaileena da, esaterako, Google (www.google.com).

Euskarazko testuetan hitz osoak bilatzea ez da bide oso zehatza, sarritan hitzetan atzizkiak azaltzen baitira; eta hitz-hasierak bakarrik bilatzen baditugu, horrelaxe hasten diren beste hitz luzeagoei dagozkien emaitzak ere azalduko zaizkigu, emaitzen kalitatea zapuztuz. Adibidez,

ero hitza duten dokumentuak bilatu nahi baditugu, *eroari*, *eroekin*, *eroengana* hitzak dituzten dokumentuak ere detektatu nahi ditugu; konponketa bat litzateke "ero" letrekin hasten diren hitz guztiak detektatzea (*ero** bilatzea), baina horrelakoetan *erosotasun*, *erosi*, *erosten*, *eroale*…hitzen aipamenak dituzten dokumentuak ere jasoko ditugu, eta horrelakorik ez dugu nahi, azken horien erreferentziak agertzen badira, benetan bilatzen ditugunekin nahastatuta agertuko direlako. Beraz, ahal dela, lematizazioan oinarritutako bilaketak egin beharko ditugu euskarazko dokumentuak atzitzeko.

Elebila, euskaraz moldatzen den web-bilatzailea

Elebila beste dokumentu-bilatzaile bat da, baina bere esparrua zabalagoa da, Internet osoan bilatzen du eta. Web-nabigatzaile oso bat eraikitzea lan erraldoia denez, bide berri bat urratu behar izan da euskaraz ondo moldatzen den web-bilatzaile bat sortzeko. Internet nabigatzaileek barruan bi modulu edukitzen dute: bata, *modulu indexatzailea,* eskura dituen dokumentuak aztertzen dituena barruko hitz edo kontzeptuekin indizeak sortzeko; eta bestea, *modulu bilatzailea,* bilaketak azkarrago egitea ahalbidetzen duena. Interneteko bilatzaileen kasuan, modulu bi horiek etengabe daude martxan, web gune berriak detektatzen, analizatzen eta indizeak eguneratzen. Baina modulu bi horiek etengabe martxan edukitzeak konputagailu asko eta handiak behar ditu; gainera Interneten dauden testu guztiak lematizatuta edukitzea ere lan mardula izan daiteke. Zailtasun horiek direla eta oso konplexu litzateke web bilatzaile oso bat eraikitzea euskararako. Elebila-k indexazioa eta hitz osoen bilaketa egiten dituen Web bilatzaile estandar bat erabiltzen du (MSN), baina horren gainean programa konplexuago dago. MSNk, Googlek, Yahook eta beste bilatzaileek ez bezala, Elebilak badaki bereizten euskaraz idatzita dauden orriak, eta gainera hitz bat bilatzeko eskatzen diogunean atzizkiak gehituta lortu daitezkeen beste hainbat hitz zuzen ere bilatzeko eskatuko dio bilatzaile estandarrari. Eskaera guzti horien emaitzak aztertuta osatuko da Elebilak eskainiko digun emaitza.

Emaitzan euskarazko web orriak bakarrik lortzen direla bermatzeko hainbat iragazki erabiltzen ditu Elebilak. Iragazki hau zenbat eta zorrotzagoa izan, orduan eta zailagoa da euskaraz ez dauden emaitzak lortzea. Bestalde, iragazkia oso zorrotza bada posible da euskaraz dauden zenbait emaitza kanpoan geratzea. Egokiena iragazkia maila altuenean mantentzea da (oso fidagarria da) eta emaitza gutxi lortzen diren kasuetan iragazkia ahulduz probatzea.

Elebilak barruan lematizatzaile bat duenez, beste laguntza gehigarri batzuk ere eskaintzen dio bilatzaile euskaldunei. Euskaraz bilaketa bat egiten den bakoitzean Horrela bilaketarekin aurrera jarraitzeko beste proposamen integratiboak ematen ditu hainbatetan. Proposamen hauetako

Gizakiontzat oso erraza da geure hizkuntza ulertzea, konputagailuari asko kostatzen zaio ordea. Adibidez, testu bateko hitzak irakurtzen dituguneanguk ez ditugu kontuan hartzen ezohiko diren interpretazio bitxiak, baina konputagailuak bai, denak aztertu behar ditu eta.

Yourdictionary.com gunean azaltzen den bezala, munduan 6.800 hizkuntza omen dago. Horien artean 2.261 hizkuntzek baino ez dute adierazpen idatzia, eta 300 hizkuntzetan bakarrik kontsultatu daitezke hiztegi elektronikorik. Euskara 300 hizkuntza horien tropelean dabil, eta areago, IKT eta hizkuntza-teknologiako produktuak aztertuta ziur euskara sartuko litzakeela lehenengo 100 hizkuntzen artean, agian lehenengo 50 hizkuntzen artean ere bai.

batean klik egitea besterik ez dago proposatutako bilaketa horietan aurrera jarraitzeko. Adibidez:

• bilatzeko hitza "*gorri*" bada, aukera hauek eskaintzen ditu:
Erabilitako analisia: gorri izena.
Beste analisiak: gorri adjektiboa, gor izena, gorritu aditza, gor adjektiboa

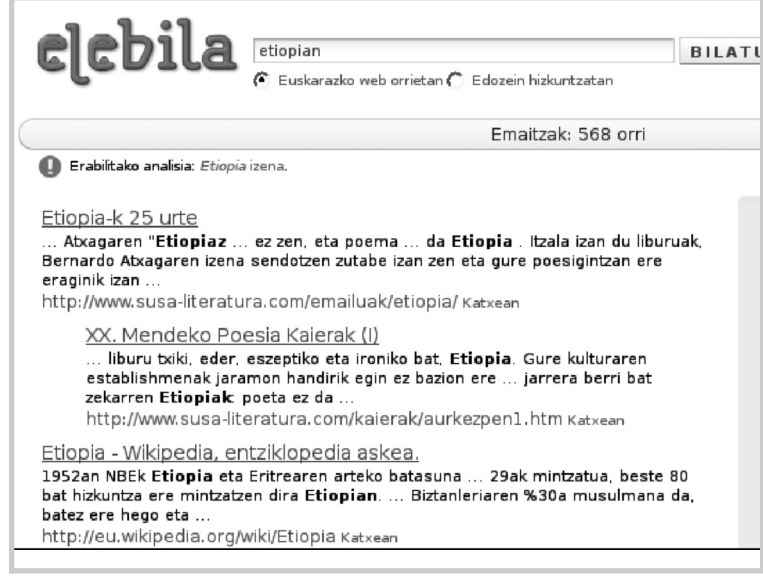
Normalean izenak bilatzen ditugu, eta izenari dagozkion atzizkiekin egin ditu bilaketak Elebilak, baina agian interesatzen zaiguna aditza da (gorritu) edo agian gor hitza izena edo adjektibo moduan

• bilatzeko hitza "*eritzi*" bada, aukera hauek eskaintzen ditu:
Erabilitako analisia: eritzi izena.
Sartutako testuan aldaerak daude: iritzi

Eritzi gaur egunean ez da euskara batuaren forma estandarra. Euskaltzaindiak "iritzi" erabaki zuen orain dela urte batzuk. Bilatzaileak aldaera horren berri duenez, forma estandarrarekin bilatzeko aukera ematen du (klik bakarra eginez)

• bilatzeko hitza "*Etiopian*" bada, aukera hauek eskaintzen ditu:
Erabilitako analisia: Etiopia izena.

Bilatzaileak Etiopian hitza Etiopia izan propio ezagutu duenez, eta beste analisi posiblerik ez dagoenez, zuzenean Etiopia lemarekin egin du bilaketa. Dokumentu batzuk aurkitu dira "Etiopian" ez diren "Etiopiak", "Etiopiaz" hitzak aurkitutakoan (ikus. irudia). Hor ikusten da argi lematizazioa sartzearen abantaila.



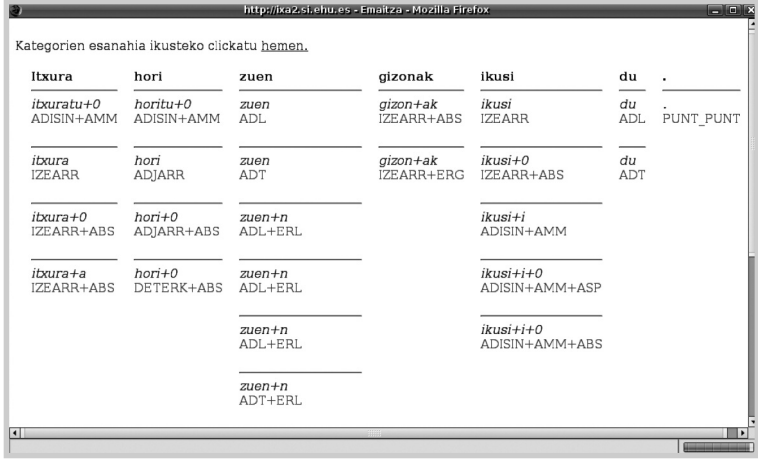
3. irudia- Elebila: lematizazio bidezko bilaketa Interneten.

2.4 Beste aplikazio eta tresna publiko batzuk

Ixa taldean, beti hizkuntzaren teknologiaren barruan, beste aplikazio informatiko, tresna eta hizkuntza-baliabide definitu dira. Aplikazio informatikoen artean aipagarri dira Matxin itzulpen-sistema (Alegria et al., 2007; Mayor, 2007) eta Zientzia eta Teknologiaren Corpusa (Areta et al., 2007). Baina bi aplikazio horiekin beste bi artikulu definitu dira BAT aldizkariko ale honetarako, beraz, jo beza irakurleak artikulu horietara berri zehatzago lortzeko.

Tresna linguistikoak gehiago dira. Erabilera publikokoak direlako eta aurreko atalean aurkeztu ditugun aplikazioetako oinarri direlako merezi du ekartzea hona analizatzailea morfologikoa eta lematizatzailea. IXA taldeko Demoak orrira joz (ixa.si.ehu.es/Ixa/Demoak), praktikan ikus dezakegu nolakoa den esaldi bateko hitzen analisi morfologikoa (Alegria et al., 96), eta programa lematizatzaileak nola murrizten dituen gero analisi-aukerak.

Gizakiontzat oso erraza da geure hizkuntza ulertzea, konputagailuari asko kostatzen zaio ordea. Adibidez, testu bateko hitzak irakurtzen ditugunean guk ez ditugu kontuan hartzen ezohiko diren interpretazio bitxiak, baina konputagailuak bai, denak aztertu behar ditu eta. Programa lematizatzaileak laguntzen dio konputagailuari interpretazio morfologikoen artean egokia aukeratzen testuinguruaren arabera.



4. irudia- Analizatzaile morfologikoaren emaitza.

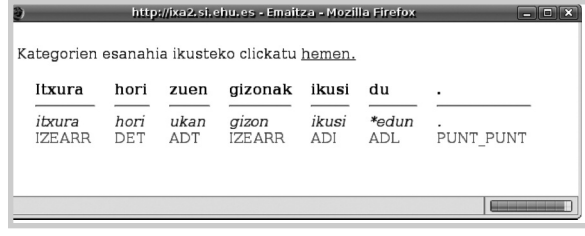
Hori erraz ikus dezakegu IXA taldeko Demoak orrira joz. Batetik analizatu morfologikoki ondoko esaldia : *Itxura hori zuen gizonak ikusi du.*

Argi dago analizatzaile morfologikoak hitz bakoitza testuingurua kontuan hartu gabe analizatzen duela (ikus 4. irudia). *Itxura* hitza aditza ere izan daitekeela dio; *hori* hitza aditza eta adjektibo ere izan daitekeela; edo *ikusi* hitza izena. Beste esaldi batzuetan agian gerta litezke, baina gure esaldi horretan ez.

Orain dela 20 urte euskara lantzeko gure lehenengo proiektua itzulpen-sistema bat sortzeko izan zen, bere bideragarritasuna aztertzekoa. Orduan lau irakasle baino ez ginen eta konturatu ginen gure orduko indarrekin askoz zentzuzkoagoa zela ez egitea itzulpen-sistema oso mugatu bat, jostailuzko lexikoia eta gramatikak izango zituena, baizik eta gure indar guztiak oinarri sendoak eraikitzen inbertitzea.

Euskararen erronka honi aurre egiteko pertsona trebatuak behar zirela jakinda, hasieratik ere saiatu gara heziketa egokia zabaltzen eta teknologia honen protagonistak izango diren teknikari eta ikerlariak trebatzen, beti ere alde informatikaria eta alde linguistikoa uztartuz.

Orduan gero analizatu esaldi bera lematizatzailearekin. Lematizatzaileak analisi morfologikoa egiten du baina gero hitzaren testuingurua aztertuta hitz bakoitzerako analisi bakarra aukeratzen du (ikus 5. irudia).



5. irudia- Lematizatzailearen emaitza.

Morfeus analizatzaile morfologikoak batez beste euskarazko hitz bakoitzerako 2,81 analisi diferente sortzen ditu. Kategoria eta azpikategoria sintaktikoa bakarrik kontuan hartuta 1,5 analisi ematen du hitz bakoitzeko. Lematizatzaileak ordea, testuingurua aztertu ondoren lema eta kategoria bakarra hautatzen du hitz bakoitzerako. Hanka sartzen du, baina %1 edo %2an baino ez. Oso tresna erabilgarria da hizkuntza-teknologian.

Bukatzeko, esan IXA Taldeko Demoak orriaren bidez beste tresna batzuk ikus daitezkeela martxan: Xuxen zuzentzaile ortografikoa, Eihera entitateen ezagutzailea (pertsonen izenak, tokiak eta erakundeak detektatzen ditu testuan; Aranzabe eta al., 2004) eta Izati sintagma-banatzailea (azaleko sintaxia; Alegria et al., 2004)).

Baliabide linguistikoen aldetik Demoen orrian bertan Zientzia eta Teknologia Corpusa eta Euskararen Datu-Base Lexikala (EDBL, gure garapen guztien oinarri lexikala) kontsultatu daitezke. Azkenik, merezi du aipatzea hemen EusWN baliabide lexikala (Agirre et al, 2002), euskarazko WordNet. Hitzen esanahiak erlazio lexiko-semantikoen inguruan egituratzen dituen Ezagutza-Base Lexikala da WordNet. Euskarazko WordNet-ek EuroWordNet-eko zehaztapenak jarraitzen ditu, eta beraz hizkuntzen arteko hitzen esanahiak *Hizkuntza_Arteko_Indize*aren bidez daude lotuta. Edukiak arakatzeko interfaze bat dago (ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl), Euskarazko, Catalanerazko, Gaztelerazko eta Ingelesezko WordNeten edukiak atzitzen dituena. Euskarazko WordNet oraindik garapen bidean dago. Datu batzuk automatikoki sartu dira, eta beraz errore eta hutsuneak topatu ditzakezu. Etengabe ari gara eguneratzen, eta kontsulta dezakezun datubase bera da gure hizkuntzalariak denbora errealean lantzen ari direna. Egungo egoera ezagutzeko eta datu-basea ikusi ahal izateko jarraitu beheko esteka.

Atal honetan aipatu ditugun produktu gehienak hizkuntza teknologiako nazio arteko erreferentzia den *Association for Computational Linguistics (ACL)* elkartearen wikian[13] ere sartuta daude.

3. EUSKARAREN EGOERA HIZKUNTZA-TEKNOLOGIAN ORAIN

Euskararen Softwarearen Katalogoan (www.ueu.org/softkat) hizkuntzaren prozesamenduarekin lotuta dauden aplikazioak 44 dira; honela daude sailkatuta: ediziorako laguntzak (Xuxen, Elhuyar-Word, sinonimo hiztegiak EuskalBar...), hizketaren tratamendua (Bizkaieraren Fonoteka eta AhoTTS Testu-Ahots Bihurgailua, Fonatari), Euskara ikasteko metodoak (Bai & Bye, BOGA eta HEZINET), Lematizatzailea eta informazioa bilatzeko tresna, datu-base dokumentala (Kapsula), corpus (XX. mendekoa), eta 20 baliabide lexikal (hiztegiak, esamoldeak, ...). Eusko Jaurlaritza "Euskararen IKTen inbentarioa" prestatzen ari da orain eta beste produktu batzuk ere hasi dira ikusten gune berri horretan[14], baina oraindik ez dago guztiz osatuta. Bi gune horietan bilduta dagoena aztertuta, eta kontuan hartuta beste produktu batzuk gune bi horietan jasota ez daudenak, esan dezakegu, beraz, ez gaudela basamortuan, hori ez dela hutsaren hurrengoa, baina bai oso gutxi gaztelaniarako, frantseserako edo, batez ere, ingeleserako eskuragarri dauden ehunka programa eta baliabiderekin alderatzen badugu.

Yourdictionary.com gunean azaltzen den bezala, munduan 6.800 hizkuntza omen dago. Horien artean 2.261 hizkuntzek baino ez dute adierazpen idatzia, eta 300 hizkuntzetan bakarrik kontsultatu daitezke hiztegi elektronikorik. Euskara 300 hizkuntza horien tropelean dabil, eta areago, IKT eta hizkuntza-teknologiako produktuak aztertuta ziur euskara sartuko litzakeela lehenengo 100 hizkuntzen artean, agian lehenengo 50 hizkuntzen artean ere bai. Azken 25 urteetako ahaleginen fruitua da hori, baina agian hori ez da nahikoa etorkizun hurbileko erronkei ekiteko.

Orokorrean aztertuta, ez hizkuntzaren prozesamenduarekin lotuta dauden aplikazioak bakarrik, euskarak orain informatikaren munduan duen egoera ere ez da guztiz txarra; badira hainbat aplikazio, baina honetan ere oraindik zeregin handia dago egoera normalizatu batera iristeko. Jotzen badugu berriro Euskararen Softwarearen Katalogora aplikazio-motaren arabera honako zenbaki hauek aurkituko ditugu:

- 31 bulego aplikazio (testu prozesatzaileak, kontabilitatea…)
- 30 aisialdikoak (musika,jokuak…)
- 44 hizkuntzarekin lotuta (itzultzaileak, zuzentzaileak, hiztegiak…)
- 56 interneten aritzeko (nabigatzaileak, posta elektronikoa…)
- 26 tresna orokor (sistema eragileak, interneteko datu-baseak eta bilatzaileak...)
- 74 irakaskuntzarekin lotuta edo eta joku pedagogiko (matematika, zientziak…)

Artikulu honetan morfologian egindako aplikazioak aurkeztu ditugu, baita egun aurreratu samar ditugu sintaxi eta semantika lantzeko zenbait tresna. Tresna guzti horietan metodo sinbolikoekin eta metodo empirikoekin esperimentatzen ari gara egun, eta eginkizun dugu mota bietako metodoak konbinatzea horrela emaitzak hobetu ahal izateko.

Azken 30 urtean egindako urratsei ezker euskararen egoerak hobera egin du nabarmen, baina ahalegin guzti horiek eginda ere euskararen etorkizuna oraindik ez dago ziurtatuta. Egindako urratsak horiek erabatekoak ez izateaz gain, euskara industri guneetatik kanpo jarraitzen du, Informazioa eta Komunikazioaren Teknologia (IKT) berriarekin lotuta dauden industri guneetatik ere bai.

4. HIZKUNTZA TEKNOLOGIA LANTZEKO ESTRATEGIA, LEHENTASUNAK

Gorago ikusi dugunez argi dago ingelesa dela nagusia teknologia berri honetan. Ingelesa batez ere, baina beste hizkuntza nagusiek ere, bigarren maila batean, hainbat produktu eta baliabide garatu dituzte. Argi dago beste hizkuntzek ahalegin handia egin behar dutela atzean ez gelditzeko, are gehiago euskara bezalako hizkuntza txikiek. Zer egin daiteke atzean ez geratzeko? Nola ekin erronka honi? IXA taldean urteetan jarraitu izan dugu estrategia bat, urrats-kate bat hizkuntzaren teknologiari metodo batekin ekiteko.

Orain dela 20 urte euskara lantzeko gure lehenengo proiektua itzulpen-sistema bat sortzeko izan zen, bere bideragarritasuna aztertzekoa. Orduan lau irakasle baino ez ginen eta konturatu ginen gure orduko indarrekin askoz zentzuzkoagoa zela ez egitea itzulpen-sistema oso mugatu bat, jostailuzko lexikoia eta gramatikak izango zituena, baizik eta gure indar guztiak oinarri sendoak eraikitzen inbertitzea. Euskararen morfologia hain diferentea bazen, beste hizkuntzetarako produktuak gurera egokitzeko arazo larriak aurkituko behar bagenitu beti, hoberena zen morfologiaren azterketari lehenbailehen sakonki ekitea. Beraz, itzulpen-kontuak gerora utzi eta lexikoa eta morfologiari ekin genien modu sakonean; tresna horiek geroago itzulpenerako erabili ahal izango ziren…baita beste hainbat aplikaziotarako ere! Geroago etorri ziren morfologiaren gaineko aplikazio informatikoak, gorago aipatu ditugunak, geroago etorri ziren ere beste tresna eta aplikazio konplexuagoak. Taldearen ia 20 urteko ibilbidea estrategia horren arabera egin dugu. Nazioarteko foroetan ere aurkeztu eta kontrastatu dugu beste ikerlari batzuekin (Alegria et al., 2001; Sarasola, 2007). Ideia nagusiak ondokoak dira:

- **Hasieran oinarrizko baliabide eta tresna sendoak sortu behar dira**, eta geroago sortu merkatu-aplikazioak. Alderantziz ez dela egin behar! Produktu posibleen artean bereiztu izan dugu zein diren hizkuntza-baliabideak, zein tresna, eta zein aplikazioa. Tresna eta aplikazioak bereizten ditugu, biak produktu informatikoak izan arren, tresnak ez baitira erabiltzaile arruntarentzat eta aplikazioak bai; tresnak hizkuntza-teknologian dabiltzan teknikariek erabil ditzaten definitu dira. Eta horren arabera baliabide, tresna edo aplikazio bakoitza noiz egin behar den aurreikusi izan dugu, ekoizpen prozesu hori optimizatu nahian.
- **Formatu estandarren erabilpena**. Produktu bakoitza geroagoko produktu berrien garapenean ahalik eta modu zabalenean berrerabilia izatea da gure helburua. Sortzen diren produktuak formatu estandarren arabera[15] definitu behar ditugu, bai hartuko dituzten datuetan, bai itzuliko dituzten emaitzetan. Horrela berrerabilgarriak izan-

go dira beste hainbat produktutan eta haien garapena modu inkrementalean egin ahal izango da.

• **Ahal den guztietan software librea erabili eta sortu.** Berrerabili ahal izateko, noski, oso bide interesgarria da produktuak software libre moduan plazaratzea.

Badakigu puntu horiek "oso sinpleak" diruditela, informatikako edozein aplikazio garatzeko erabili behar direnak direla, baina gure eskarmentuak dio hainbat hizkuntzatarako proiektutan ez dela horrela jokatu. Egun euskarak hizkuntza-teknologiako lehenengo 100 hizkuntzen artean baldin badago, neurri handi batean estrategia horri jarraitu izan zaiolako dela uste dugu. Estrategiaren erabileraren adibide gisa esan dezakegu itzulpen automatikoan denbora laburrean garatu ahal izan bada lehenengo prototipoa hainbat tresna berrerabili direlako izan dela, bai taldean aurretik sortutakoak (baliabide lexikalak eta morfologia), baita software libreko beste batzuk ere (erdaretarako analizatzaileak), modulu berri bakar batzuk soilik sortu behar izan ditugu eta horiek beste talde batzuekin garatu ditugu elkarlanean.

Bide horretan mugarri hauek definitu ditzakegu:

• 1993 Xuxen zuzentzaile ortografikoa.
• 1996 EDBL datu-base lexikala
• 1998 Lematizatzailea
• 2002 Elhuyar Word hiztegi kontsultarako plugina
• 2006 ZT corpusa, Matxin itzulpen sistema
• 2007 Euskal Wordnet, azaleko analizatzaile sintaktikoa

Hasieran lau partaide baginen ere, orain 28 informatikari, 13 hizkuntzalari eta 2 teknikari gara; Euskal Herriko zazpi enpresekin lankidetzan gabiltza eta atzerriko beste bostekin. Spin-off erako bi enpresen sorkuntzan parte hartu dugu (Usurbilgo Eleka eta Alacanteko Prompsit). 2002. urtetik Eusko Jaurlaritzak definitu zuen *Ingeniaritza linguistikoa* ikerlerro estrategikoa parte hartu dugu (Hizking21 eta Anhitz proiektuak) beste ikerketa zentroekin batera (Aholab, Elhuyar, Vicomtech eta Robotiker).

Euskararen erronka honi aurre egiteko pertsona trebatuak behar zirela jakinda, hasieratik ere saiatu gara heziketa egokia zabaltzen eta teknologia honen protagonistak izango diren teknikari eta ikerlariak trebatzen, beti ere alde informatikaria eta alde linguistikoa uztartuz. 1989an doktorego-ikastaroak ematen hasi ginen, 2002an Hiztek titulu propioa sortu zen UEUren lankidetzarekin, 2005ean doktorego-programa bat (Hizkuntzaren azterketa eta prozesamendua) eta aurten abiatu da izen bereko Europako master ofiziala. Unibertsitate mailako euskarazko masterrak ez dira asko. Hamaika ahalegin eta ilusioren fruituak dira. Zenbat irakasle eta ikasle ibili garen, hor, lanean elkarrekin, heziketa-aukera hau

euskaraz egin ahal izateko! 50 baino gehiago dira bide horietatik titulua eta heziketa berezitua jaso duten teknikari/ikerlari berriak. 15 doktorego tesi sortu dira iturri horretatik. Ingeniaritza linguistikoan I+G horretan (Ikerketan eta Garapenean) arituko den komunitate zabal bat sortu behar dugu.

5. GURE ETORKIZUNERAKO PLANAK

Euskararako baliabide linguistikoak sortu beharko dira

Hasieran oinarrizko baliabideak eta tresnak sendo sortu behar direla defendatzen dugu, eta hortik geroago sortuko direla aplikazio praktikoak Ixa taldean, une honetan corpus oso garrantzitsua da. Testu edo hizketa-grabaketa andana duten bildumak dira corpusak dei ditzakegunak, eta ezaugarri hauek dituzte: eredugarriak, adierazgarriak, neurri mugatukoak eta makinan irakurtzeko moduan jarriak. Erreferentzia estandarra dira hizkuntza lantzeko. Informazioaren gizartean, hizkuntza batek duen garrantzia neurtu nahi denean, aplikazioak garatzeko dituen baliabide linguistikoak aztertzen dira gaur egun. Baliabide horien artean, corpus handien garapena lehenetariko helburua izan ohi da.

Corpus idatzien adibide gisa (ikus Corpus Survey[16]) aipa daiteke British National Corpus[17]. Honek 4000 testu-zati ditu, eta denera 100 milioi hitz . Gure inguruan baditugu adibide batzuk (etiketatuak: XX. Mendeko Corpusa[18] (4,6 Mhitz), Ereduzko Prosa Gaur[19] (9,6 Mhitz) eta ZT[20] (8 Mhitz); ez-etiketatuak: Susa[21], Klasikoen Gordailua, Ibinagabeitia Proiektua eta Orotariko Euskal Hiztegia) baino oraindik asko falta zaigu beste hizkuntzetan dauden tamainetara iristeko.

100 megahitz izan daiteke helburu bat testu-bilketan (corpus elebakarra) lortzeko epe erdian, ingeleserako gaur egunean badira tamaina horretako corpusak. Eta maila askotan (morfosintaktikoan, sintaktikoan, adiera semantikoarekin, ...) etiketatu beharko dira gero.

Bestalde, corpus elebidun parekatuak oso garrantzitsuak dira itzulpengintza sustatzeko. Gutxienez 30 miloi hitzeko corpusa sortu beharko litzateke teknika estatistikoek emaitza minimoki onargarriak izan zitezen. Frogatuta dago hortik gora corpusaren tamaina bikoizketa bakoitzeko %1eko hobekuntza lortzen dela itzulpenaren kalitatean[22]. Euskaraz milioi gutxiko corpus elebidun parekaturik nekez biltzen ahal den bitartean, nazioartekoan tamaina hauek lortu dira:

- Europarl[23]: 30 milioi hitz Europako 11 hizkuntza ofizialetan
- Acquis Communitaire: 8-50 milioi hitz Europako 20 hizkuntza ofizialetan.
- Canadian Hansards: 20 milioi hitz ingelesez eta frantsesez.
- Txinera-ingelesa eta arabiera-ingelesa: 100 milioi hitz baino gehiago LDC zerbitzuan.

e-edukiak biltzea, biblioteka nazionala

Bide horretan Interneteko euskarazko edukiak (edo hobeto, modu digitalean argitaratzen den guztia) sistematikoki biltzea litzateke helburua. Testu andana hori oinarrizko baliabidea litzateke, hizkuntza-teknologiarako eta giza-zientzia guztietarako ere bai. Biblioteka Nazional zenbait, eta beste erakunde batzuk ere, ari dira alor honetan neurriak hartzen. 2002ko urtarrilean nazioarteko bilkura bat[24] egin zen horren inguruan. Europa mailan badaude momentu honetan ondare historikoa gordetzea helburu duten hainbat deialdi eta proiektu. Besteak beste, nahiko aurreratuta dago Danimarkako esperientzia[25]. Hona hemen beste erreferentzia interesgarri batzuk:

a) European Digital Library Project[26]
b) Biblioteca Nacional de España[27]
c) DELOS Network of Excellence on Digital Libraries[28]

Hizkuntza-teknologiako tresnak oso lagungarriak lirateke, biltze-prozesu erraldoi horretan.

Sintaxia, semantika eta beste aplikazio batzuk

Artikulu honetan morfologian egindako aplikazioak aurkeztu ditugu, baita egun aurreratu samar ditugu sintaxi eta semantika lantzeko zenbait tresna. Tresna guzti horietan metodo sinbolikoekin eta metodo empirikoekin esperimentatzen ari gara egun, eta eginkizun dugu mota bietako metodoak konbinatzea horrela emaitzak hobetu ahal izateko. Etorkizun hurbilean ere aplikazio berriak garatu nahi ditugu tresna horiekin, adibidez: gramatika eta estilo-zuzentzaileak, bigarren hizkuntza ikasteko sistemak, galderak erantzuteko sistemak (Question Answering), dokumentu-bilatzaileak (IR, Information Retrieval), informazio-erauzketa dokumentuetatik (IE, Information Extraction), laburpen automatikoa (Summarization) eta dokumentu-sailkatzaileak.

6. ONDORIOAK

Azken 30 urtean egindako urratsei ezker euskararen egoerak hobera egin du nabarmen, baina ahalegin guzti horiek eginda ere euskararen etorkizuna oraindik ez dago ziurtatuta. Egindako urratsak horiek erabatekoak ez izateaz gain, euskara industri guneetatik kanpo jarraitzen du, Informazioa eta Komunikazioaren Teknologia (IKT) berriarekin lotuta dauden industri guneetatik ere bai.

Hizkuntzaren Teknologiaren ekarpena funtsezkoa da IKT arloan euskararen erabileran normalizazioa sustatzeko. Testuak errazago eta txukunago sortu ahal izateko, sarean edukiak zabaldu edo bilatu nahi

dituenak tresna egokiak izan ditzan. Helburu horretan ere bere hizkuntzarekin errazago lan egin dezan eta norberaren hizkuntzarekin ere gozatu ahal izateko.

Produktuak garatzeko lehentasun-ordena, estandarizazioa, berrerabilpena, eta software librea oinarri dituen estrategia arrakastatsu izan da. Bide horretan eta estrategia bati jarraituz IXA taldetik hainbat ekarpen egin dira, ikerketa mailan eta aplikazio praktikoen aldetik ere bai. Morfologiaren azterketa sakonaren fruituak izan diren hiru aplikazio praktikoek (zuzentzailea, hiztegi-kontsulta on-line, dokumentu-bilatzailea, itzulpen-sistema...) horrela frogatzen dute.

Nazioartekoan puntako mailan mugituko den industria sendoa sortu dezakegu. Gure eskarmentua, eta tresnak beste hizkuntza batzuen prozesamenduan lagungarria izan daiteke. Ikerketa-taldeek, industriak eta erakunde ofizialek koordinatu egin behar dira helburu hori lortzeko. Argi dago hizkuntzaren industria honetan ingeleserako produktuen merkatua oso handia dela. Baina guk uste dugu produktu horiek ez direla zabaldu modu egokian beste hizkuntzetarako, badagoela espazio eta zeregin, ikerketan eta produktuen mailan ere bai, gure ekarpenak bideratu ahal izateko. Guk euskaldunok, eta orokorrean europarrok, ohituta gaude eleaniztasunean bizi izaten. Alde batetik hainbat ahalegin, gastu eta buruhauste ekartzen digu eleaniztasuna, baina, beste alde batetik, kokapen oso aurreratuan jartzen gaitu nazioartekoan ekarpen esanguratsuak egiteko. Gaitasun handiagoa dugu eleaniztasuna lantzeko. Gainera gure hizkuntza oso ezaugarri desberdinak dituenez probaleku ezin hobea izan daiteke hizkuntza-produktuen moldagarritasuna frogatzeko.▮

7. ERREFERENTZIAK

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K. 1997. A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford. 1997.

Agirre E., Ansa O., Arregi X., Arriola J., Díaz de Ilarraza A., Pociello E., Uria L. 2002. Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. Proceedings of First International WordNet Conference. pp. 32-40. Mysore (India).

Alegria I., Artola X., Sarasola K. 1996. Automatic morphological analysis of Basque. Literary & Linguistic Computing Vol. 11, No. 4, 193-203. Oxford University Press. Oxford. 1996.

Alegria I.,Artola X.,Sarasola K. (2001) Hizkuntzaren tratamendu automatikoa: aplikazioak, tresnak, baliabideak eta oinarriak Euskonews, http://suse00.su.ehu.es/euskonews/0110zbk/frgaia.htm

Alegria,I. eta M. Jesus Rodriguez. (2003) Euskarararen presentzia Interneten neurtu nahian.. *BAT soziolinguistika aldizkaria, 48, 89-100. 2003*

Alegria I., Arregi O., Ezeiza N., Fernandez I., Urizar R. 2004. Design and

Development of a Named Entity Recognizer for an Agglutinative Language. First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.

Alegria I., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2007. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS 4394. 374-384. Cicling 2007.

Amorrortu E. (2002) Bilingual Education in the Basque Country: Achievements and Challenges after Four Decades of Acquisition Planning. Journal of Iberian and Latin American Literary and Cultural Studies, Volume 2, Number 2.

Aranzabe M., Arriola J.M., Díaz de Ilarraza 2004. Towards a Dependency Parser of Basque. Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland.

Areta N., Gurrutxaga A., Leturia I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Sologaistoa A. 2007. ZT Corpus: Annotation and tools for Basque corpora Copus Linguistics. Birminghan.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages COLING-ACL'98. Pgs. 380 - 384. Vol 1. Montreal (Canada). August 10-14, 1998.

Mayor A. 2007. MATXIN: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz. Doktorego tesia. *Euskal Herriko Unibertsitateko Donostiako Informatika Fakultatea.*

Sarasola K. 2008. Technology is an effective tool to promote use of Basque. *ICML Colloquium on "Language Revitalisation through Multimedia Technology" Pecs, Hungary (Forcoming).*

OHARRAK

1. Euskararen presentzia Interneten neurtu nahian. I. Alegria eta M. Jesus Rodriguez. *BAT soziolinguistika aldizkaria, 48, 89-100. 2003*

2. http://registry.dfki.de

3. http://www.ifi.uzh.ch/CL/InteractiveCLtools/index.php

4. http://catalog.elra.info

5. http://www.ldc.upenn.edu/Catalog/catalogSearch.jsp

6. http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque

7. http://www.yourdictionary.com/languages.html

8. http://www.translation-directory.com/machine.html

9. http://www.federicozanettin.net/sslmit/cattools.htm#publications

10. http://liceu.uab.es/~joaquim/

11. J. Hutchins-en "Compendium of Translation" liburuaren eta Euromatrix proiektuaren arabera. http://www.euromatrix.net/

12. Euskalbar Firefox-erako tresnabarra da, euskera-gaztelera itzulpenak egiten lagunduko diguna, Interneten dauden hiztegi eta tresna ofizialak kontsultatuz (Euskalterm, Elhuyar, Hiztegia3000, Eus-

kaltzaindia, ItzuL posta zerrendaren artxiboa, Harluxet, Mokoroa, ZT Corpusa, Opentrad itzultzaile automatikoa eta XUXENweb zuzentzaile ortografikoa) http://www.interneteuskadi.org/euskalbar

13. http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque
14. www.euskara.euskadi.net/r59-734/eu
15. XML and TEI dira estandar egokienak
16. http://bowland-files.lancs.ac.uk/corplang/cbls/corpora.asp
17. http://info.ox.ar.uk/bnc
18. http://www.euskaracorpusa.net
19. http://www.ehu.es/euskara-orria/euskara/ereduzkoa
20. http://www.ztcorpusa.net/cgi-bin/kontsulta.py
21. http://www.susa-literatura.com
22. BLEU neurria automatikoa erabiliz.
23. http://www.statmt.org/europarl
24. http://www.nla.gov.au/ntwkpubs/gw/56/p08a01.htm
25. http://www.netarchive.dak
26. http://www.edlproject.eu
27. http://www.bne.es/esp/bne/index.htm
28. http://www.delos.info