Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta Ixa taldearen adibidea

Itziar Aduriz, Iñaki Alegria, Xabier Artola, Arantza Díaz de Ilarraza, Kepa Sarasola

Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta Ixa taldearen adibidea

Iñaki Alegria, Xabier Artola,
Arantza Diaz de Ilarraza, Kepa Sarasola
Universidad del Pais Vasco -
Euskal Herriko Unibertsitatea
i.alegria@ehu.es

Itziar Aduriz
Universitat de Barcelona
itziar.aduriz@ub.edu

Resumen

El artículo comienza presentando varios datos que muestran la situación de la lengua vasca, y a continuación proponiendo una clasificación para las lenguas del mundo según sea su presencia en Internet y en la tecnología de la lengua. El cuerpo del artículo presenta el trabajo hecho por el grupo Ixa en el campo del procesamiento automático del euskara, identificando sus siete hitos principales y describiendo la estrategia que ha guiado este desarrollo. Se plantea que esta estrategia puede servir como referencia para 190 lenguas que según la clasificación propuesta no poseen recursos de tecnología de la lengua pero si poseen una mínima presencia significativa en Internet.

Laburpena

Euskararen egoeraren inguruan hainbat datu ematen dira labur-labur, eta horrekin batera munduko hizkuntzak sailkatzeko proposamen bat aurkezten da Interneten eta hizkuntz teknologian duten egoeren araberakoa. Euskararen prozesaketa automatikoan Ixa taldeak izan duen bilakaeraren nondik norakoak zehazten dira gero, hainbat mugarri azpimarratuz eta ibilbide hori jarraitzeko erabili den estrategia deskribatuz. Munduko 190 hizkuntzentzat erreferentzia izan daiteke estrategia hori, hain zuen, Interneten presentzia minimo eduki bai baina oraindik hizkuntza-teknologia mota hau landu ez duten hizkuntzentzat

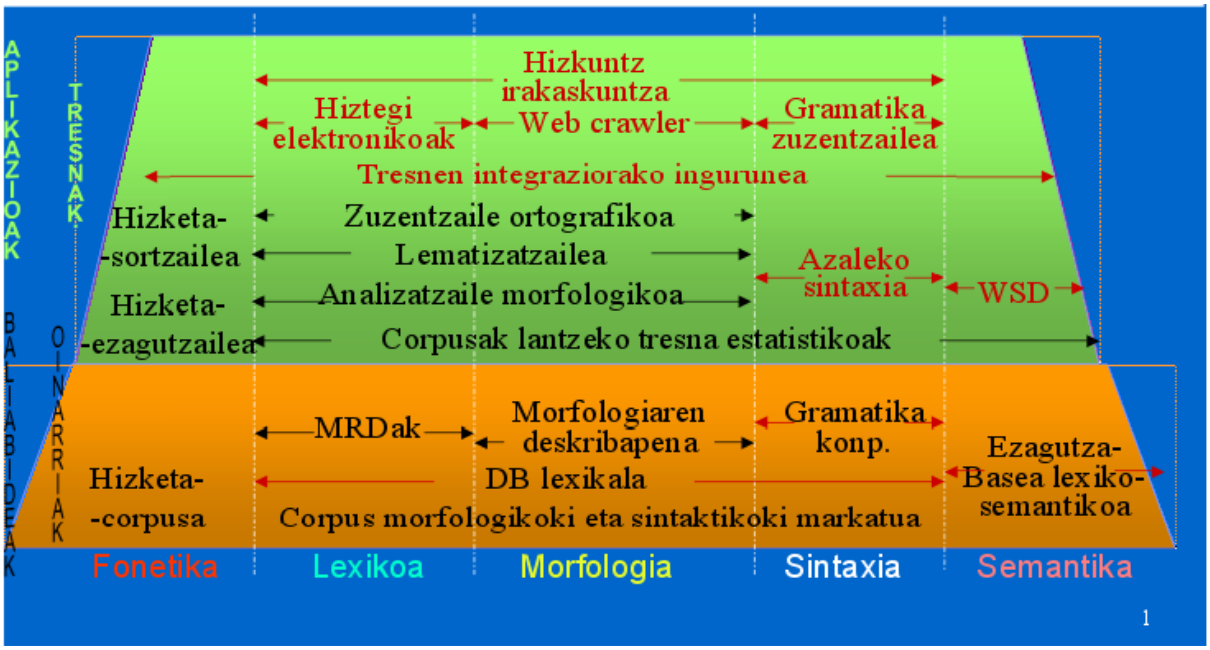
1 Sarrera

1988an Ixa taldea[1] sortu zenean Hizkuntzaren Prozesaketa eta Hizkuntz Ingeniaritza mundu akademikotik kanpo kontzeptu erabat ezezagunak ziren. Hala ere jakintza-arlo horretan oinarritutako hainbat produktu baziren merkatuan, hizkuntza gutxi batzuetarako. Adibidez, ortografia-zuzentzaileak ibiltzen ziren garaiko *MS-Word* eta *WordPerfect* testu-prozesadore ezagunetan, eta lehenengo itzultzaile automatikoak martxan zeuden erakunde handi batzuetan. Informatikaren sorreraren garaitik ametsa izan zena errealitate bihurtzen hasia zen. Edozein kasutan ere, argi zegoen bide luzea geratzen zela egiteko, are luzeago hedadura urriko hizkuntzetarako. Geroago etorriko zen Interneteko zabalkundeak areagotu egin zuen hizkuntza-teknologiaren beharra.

UPV/EHUko Informatika Fakultateko irakasle batzuek egoera horretan aukera ezin hobea ikusi genuen arnas luzeko ikerketa-lerro bat zabaltzeko. Idatzi gabe bazeuden ere, buruan argi izan genituen hainbat funts metodologiko hasiera-hasieratik:

- Euskara izango zen gure ikerkuntzaren zutabeetako bat. Guk heltzen ez bagenion, seguruen urte luzeetan beste inork ez zion helduko. Gainera euskararen ezaugarri linguistiko eta soziolinguistikoek aztergai berezia eta interesgarria eskaintzen zuten zientziaren ikuspuntutik.
- Anbizioa eta nazioarteko erreferentzia. Euskara erreferentzia izateak ez zuen ekarri beharko isolamendurik edo txokokeriarik. Nazioarteko aldizkari eta kongresuak izan behar ziren gure lanerako inspirazioa eta bertan argitaratu nahi genituen gure emaitzak.
- Berrerabilpena. Ikerketa eta aplikazioa uztartu nahi genuen, eta uztarketa horretan arrakasta izateko berrerabilpena funtsezkoa izango zen eman beharreko urrats bakoitzean.

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



1. irudia. Euskararako dauden zenbait baliabide linguistiko, tresna eta aplikazioak.

- Lankidetza-giroa eta diziplinarteko ekinbidea. Aurrekoarekin lotuz, ezin genuen aparte utzi Euskal Herrian gure lankide izan zitezkeen ikerlariak eta erakundeak: UZEI, Elhuyar, Euskaltzaindia, EHUko *Euskal Filologia* saila, eta abar luze batekin saiatu behar ginen koordinatzen. Horrez gain informazio eta Komunikazioetako Teknologien (IKT) arloko hainbat enpresa ere bidelagun gisa hartu nahi genituen.

Hau guztia azaleratu zen gure lehen proiektuan. Itzulpen automatikoaz egindako azterketa bat izan zen lan hura, eta bere ondorioa hau izan zen: *egunen batean heldu beharreko eginkizuna izango zen itzulpen automatikoa, baina hori baino lehen oinarri sendoak behar ziren, artean euskararako oinarrizko baliabiderik gabe itzulpen automatikoaren proiektua ariketa akademiko hutsean geratuko baitzen asmoa*. Eta hortik irten zen gure estrategia, 1. irudian azaltzen den piramideko produktuak urratsez-urrats eraikitzeko bidea markatu diguna (Aduriz et al., 1998). Produktu bakoitza, produktu berrien garapenean ahalik eta modu zabalenean berrerabiltzea izan da helburua.Artikulu honetan 23 urtetan zehar Ixa taldean egin den lanaren laburpena egiten da, antzeko egoera soziolinguistikoan dauden hainbat hizkuntzatarako baliagarria izango delakoan.

2. atalean euskararen egoeraren inguruan hainbat datu emango dira labur-labur, eta horrekin batera Internet eta hizkuntz teknologien arloan hizkuntzek duten egoera desberdinak tipifikatzen saio bat egingo dugu.

3. atalean sarrera honetan perfilatutako estrategiaz eztabaidatuko dugu, gure ideiez eta bideaz gain azken urteetan arlo honetan nazioartekoan egin diren ekarpenak hona ekarriz.

4. kapituluan taldearen bilakaeraren nondik norakoak zehazten dira, hainbat mugarri azpimarratuz. Bukatzeko, baliabide urriko hizkuntza baten ikuspuntutik teknologia garatzeko estrategiari buruz hainbat ondorio azpimarratu nahi izan ditugu.

2 Euskararen egoera soziolinguistikoa eta bere posizioa teknologia linguistikoan

2.1 Euskararen egoera

Mendeetako erregresio-prozesu batean sartuta ibili da euskara. Amorrorturen (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 daude (EITB telebista, Berria egunkaria...)
- Euskara estandarraren oinarria definitu zuen Euskaltzaindiak 1966an. Morfologia guztiz definituta dago orain, baina lexikoa oraindik ez. Euskara batua da gaur egun irakaskuntzan eta komunikabideetan erabiltzen dena.
- Egun 700.000 hiztun ditu euskarak, biztanleagoaren %25 gutxi gora-behera.

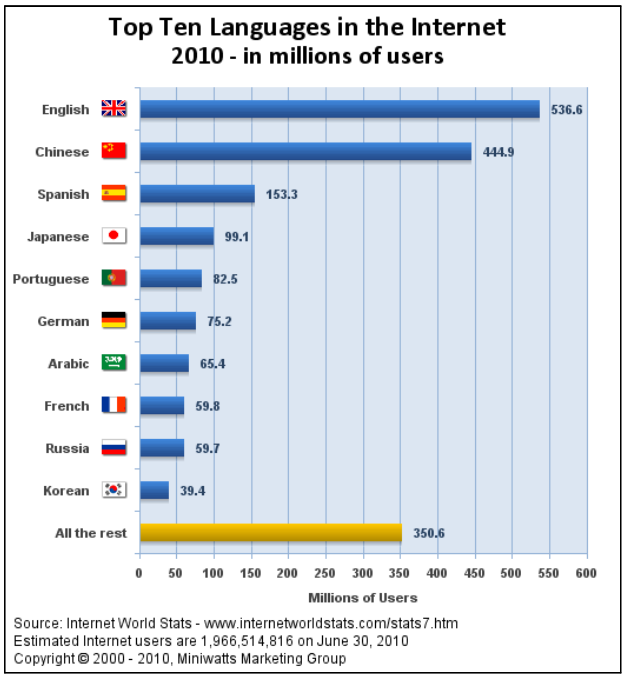
Baina ahalegin guzti horiek eginda ere euskararen etorkizuna oraindik ez dago ziurtatuta. Aipatu urrats horiek guztiz orokorrak ez izateaz gain, euskarak industri guneetatik kanpo jarraitzen du hein handi batean, baita Informazioa eta Komunikazioaren Teknologia (IKT) berriekin lotuta dauden industri guneetatik ere.

Ixa taldea hasieratik saiatu da Hizkuntzaren Teknologiaren arloa ikertzen eta produktuak gizarteratzen, beti ere IKT arloko euskararen erabileran normalizazioa sustatzeko. Adibidez, testuak errazago eta txukunago sortu ahal izateko, edo sarean edukiak zabaldu edo bilatu nahi dituenak tresna egokiak izan ditzan. Helburu horretan ere euskaldunak 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 probetxu handiena atera ahal izateko. Gogoeta horrek ekarri zuen estrategia bat definitzea eta gero hainbat urtetan horri jarraitzea.

2.2 Hizkuntzen sailkapena tresnak eta baliabideen arabera

Hizkuntza automatikoki lantzeko tresnak errealitatea dira gaur egun, hizkuntz teknologia edo HLT (Human Language Technology) izenarekin ezagutzen den arloaren barruak sortu dira. Gaur egun badira testua edo hizketa lantzeko zenbait aplikazio eskuragarri arlo honetan, 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*).

Baliabide urriko hizkuntzen artean ikusten dugu guk euskara (*less-resourced language* termino egokiena iruditzen zaigu[2]). Baina hori erlatiboa da, askoz baliabide urriagoak dituzten beste hizkuntza batzuekin konparatuta baten batek zalantzan jar dezake hori.



2. irudia. Interneteko erabileran hamar hizkuntza nagusiak (*Internet World Stats, 2010)*

Datu estatistikoak eskuratu nahi izan ditugu hizkuntzen arteko sailkapen bat zirriborratze aldera. Honakoak aurkitu ditugu IKT baliabideei buruz:

- *Internet World Stats*[3] webgunean Interneteko erabiltzaileen datuak jasotzen dira. 2010an bertan azaltzen ziren lehen 10 hizkuntzak hauek dira: ingelesa, txinera, espainiera, japoniera, portugesa, alemana, arabiera, frantsesa, errusiera eta koreera. Zoritxarrez ezin da datu zehatz gehiago jaso beste hizkuntzei buruz, baina webgune

2 Honetaz eztabaidatzen da artikulu honetan: Forcada, 2006

3 http://www.internetworldstats.com/stats7.htm

horrek ziurtatzen du gainontzeko hizkuntza guztien artean Interneteko %17,8a baino ez dutela osatzen.

- Dokumentu kopuruari dagokionez datu fidagarriak ez dira lortzen errazak. Hizkuntza erromantzeek Interneten duten hedadurari buruzko 2007ko azterketa batean[4] hauek dira lortutako datuak: dokumentuen %45 ingelesez dago, %5,9 alemanez, %3,80 espainieraz, %4,41 frantsesez, %2,66 italieraz, %1,39 portugesez, %0,28 errumanieraz eta %0,14 katalanez.
- Wikipediaren datuak[5] zehatzagoak dira. 2011ko ekaineko datuak hartuta, 281 hizkuntzetan daude artikuluak. Artikulu kopuruaren arabera lehen hamar hizkuntzak hauek dira: ingelesa, alemana, frantsesa, poloniera, japoniera, italiera, holandesa, espainiera, portugesa eta errusiera. Aurreko zerrendarekin konparatuta txinera, arabiera eta koreera desagertu dira. Beste hizkuntza iberikoei dagokienez, katalana 13. postuan agertzen da, euskara 37.ean eta galegoa 41.ean.

Hiru datu-iturri horien artean azkena da esangurasuena baina tamalez Interneten jokaera aktiboa duten hiztunen emaitza baino ez du eskaintzen.

Bestalde, IKT orokorreko datuetatik hizkuntz teknologia arloko datuetara salto eginez, hainbat webgune interesgarri kontsulta daitezke hizkuntzen egoera aztertzeko:

- **ELRA**:*European Language Resources Association*[6]. Batez ere Europakoak diren hizkuntza-baliabideak biltzen ditu (corpus eta lexikoiak). 60 hizkuntza baino gehiagoko baliabideak biltzen ditu, horien artean sei dira euskararako.
- **LDC**: *Linguistic Data Consortium*[7]. Aurrekoaren parekoa da, baina Amerikako Estatu Batuetako produktuetan espezializatua. 68 hizkuntzatako 450 bat baliabide katalogatu dira bertan, baina euskararakorik ez da agertzen.
- **ACLWiki**[8]: Hizkuntzalaritza konputazionalerako elkarteko wikia da (ACL, *Association for Computational Linguistics*). 58 hizkuntzatan dauden baliabideen berri jasotzeko gunea da. Euskararako 15 produkturen berri jaso du.
- **NLSR**: *Natural Language Soft Registry*[9]. DFKIk kudeatzen duen datu-base honetan 30 hizkuntza agertzen dira, euskararako hiru baliabide daude zerrendan, eta edozein hizkuntzatarako 59.
- ***yourdictionary.com:***[10]Hiztegi-kontsultak on-line eta itzulpen automatikoko doako zerbitzuak eskaintzen dira bertan. 307 hizkuntzarako zerbitzuak daude hor. Argi dago, baina, munduko hiztegi-zerbitzu guztiak ez daudela bertan, euskararako dagoena aztertzea aski da hori egiaztatzeko: daudenak ez dira hamarrera ailegatzen eta www.hiztegia.net gunean 50 baino gehiago bildu baitituzte. Dena dela webgune hori erreferentzia egokia izan daiteke hizkuntzen artean baliabide lexikalen azterketa konparatiboak egiteko.
- Itzulpen automatikoko sistemak eta sareko hainbat zerbitzuren berri biltzen dira gune hauetan: Translation Directory[11] eta Traduzione e computer[12]

Corpus linguistics around the world (Wilson et al., 2006) liburua ere erreferentzia interesgarria da teknologian hizkuntzek duten presentzia erlatiboa neurtzeko.

Beste adierazle interesgarri bat programen lokalizazioa da eta are gehiago oinarrizko *plug-in* linguistikoena.

- Testu-prozesadorerik hedatuena 85 hizkuntza-dialektotan dago lokalizatuta[13]. *OpenOffice*, berriz, 159tan[14] gutxienez. Euskara bietan dago.
- Bilatzaile ezagunenaren[15] interfazea 145 hizkuntzatan dago baina bilaketa aurreratuan 46 hizkuntza baino ez du bereizten du. Euskara ez da agertzen azken aukerarako.
- Itzulpen automatikoko tresna erabilienetan, BabelFish[16] eta Google[17], mugak dira hizkuntzen aldetik. Google-k ia 60 hizkuntza eskaintzen du, eta euskara alfa moduan agertzen da.

4 http://dtil.unilat.org/LI/2007/ro/resultados_ro.htm

5 http://meta.wikimedia.org/wiki/List_of_Wikipedias

6 http://www.elra.info/

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

8 http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language

9 http://registry.dfki.de/

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

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

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

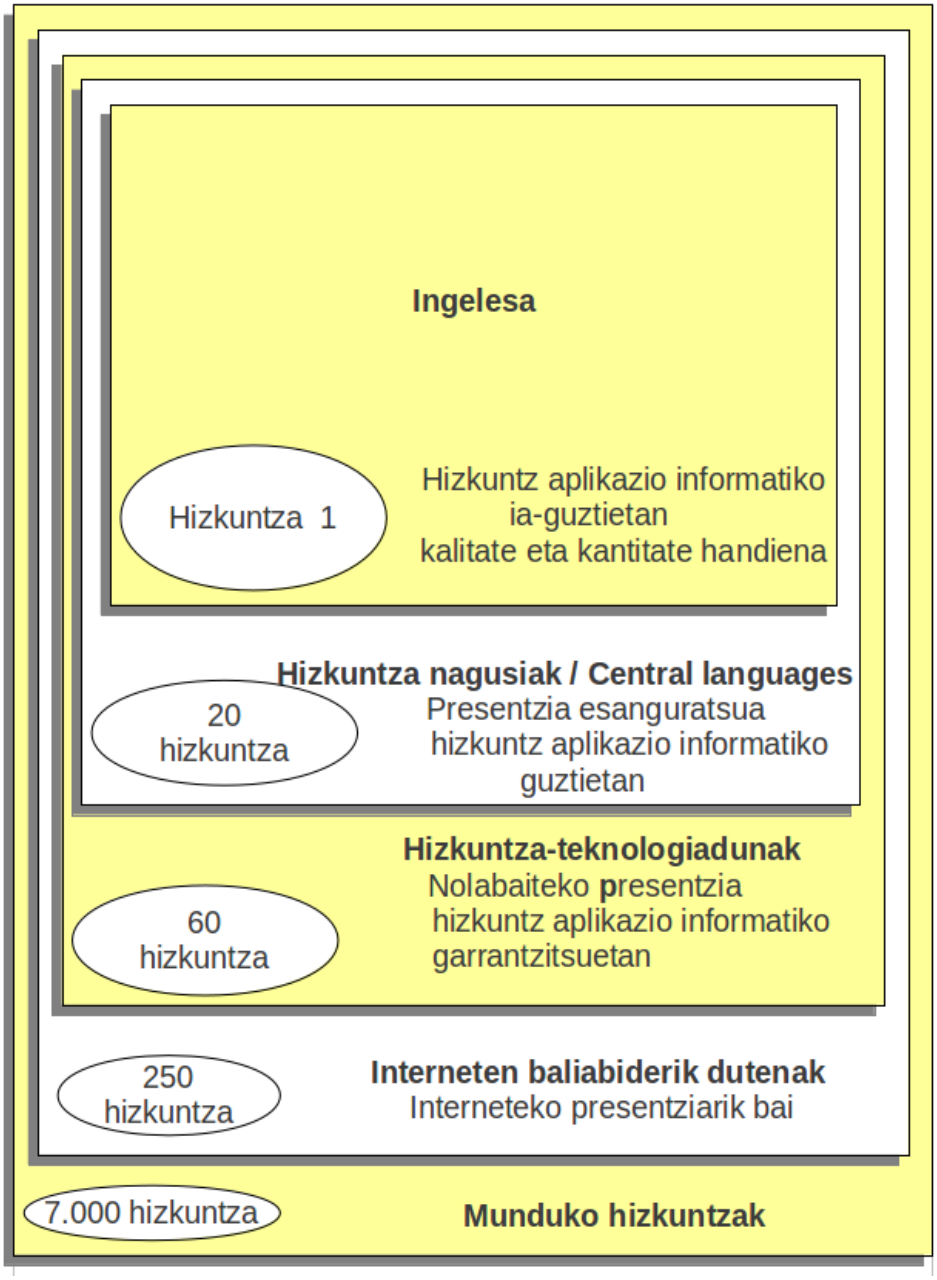
13 http://www.microsoft.com/unlimitedpotential/programs/llp.mspx

14 http://blogs.sun.com/GullFOSS/entry/locale_data_for_159_locales

15 http://www.google.com

16 http://babelfish.yahoo.com

17 http://translate.google.com



3. irudia. Hizkuntzen sailkapena hizkuntza-teknologian eta Interneten duten presentziaren arabera.

Aurreko datuak eta webguneak aztertuta ondoko sailkapena egitera ausartzen gara, beti ere hizkuntz teknologiari begira:

- Lehen maila: **Ingelesa**. Aipatutako zerrenda gehienetan %50a edo gehiago hartzen du. Berarekin alderatuta gainontzeko hizkuntza guztiak baliabide urrikoak dira.
- Bigarren maila: **10-20 bat hizkuntza**, aipatutako zerrendetan goialdean daudenak. Streiter et al. (2006) "central languages" deritze.
- Hirugarren maila: **60 bat hizkuntza gehiago**, HLT baliabideak dituztenak. Kopuru horien bueltan dabiltza ELRA, LDC eta ACLWiki.
- Laugarren maila: **250 hizkuntza inguru**, on-line baliabideren bat dutenak. Wikipedia ingurua horretan dabil.

Sailkapen hau ikusita, atera dezakegun ondorio begi-bistakoena hau da: hizkuntza asko geratzen direla sailkapen honetatik kanpo Interneten presentziarik ez dutelako. Aipatu izaten da

munduan 7.000 inguru hizkuntza daudela, eta horietatik 250 bat baino ez direla "elektronikoki alfabetatuta" daudenak. Ikus 3. irudia.

Kontuan hartu behar da, bestetik, sailkapena ez dela guztiz zehatza. Adibidez, katalana hirugarren mailan koka genezake baina aipatu adierazleren baten arabera (Wikipediako artikulu kopuruaren arabera) bigarren mailako *hizkuntza nagusi* moduan ere ikus daiteke. Euskara berriz, hirugarren mailan dago sailkapen honen arabera, hiztun kopuruaren arabera dagokiona baino gorago; hori horrela da, Ixa taldeak eta beste eragile batzuek alor honetan egindako lanari esker. Dena den bi hizkuntza horietan egindako lanak aurkezten dira SALTMIL workshop-era[18] (*HLT for minority language*s).

Oso baliagarria litzateke honen inguruan behatoki bat egitea.

3 Hizkuntza-Teknologiak lantzeko estrategiak

3.1 Estrategia taldearen esperientzian oinarrituta

Aurreko atalean ikusi dugunez argi dago ingelesa dela nagusia teknologia berri hauetan. 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 (Petek, 2000; Williams et al., 2001). Zer egin daiteke atzean ez geratzeko? Nola ekin erronka honi? Ixa taldean urteetan jarraitu izan dugu estrategia bat, urrats-kate bat hizkuntzaren teknologiari metodologia batekin ekiteko.

Sarreran esan dugun bezala orain dela 23 urte euskara lantzeko gure lehenengo proiektua itzulpen-sistema bat sortzearen ingurukoa izan zen, bere bideragarritasuna aztertzekoa. Orduan lau irakasle baino ez ginen eta konturatu ginen gure orduko indarrekin askoz zentzuzkoagoa zela eta ez itzulpen-sistema oso mugatu bat egitea, jostailuzko lexikoia eta gramatikak izango zituena. Euskararen morfologia lantzen hasi ginen konturatu ginen zein diferentea zen euskara inguru erdaretatik, baita konturatu ere, beste hizkuntzetarako produktuak gurera egokitzerakoan arazo larriak aurkituko genituela beti. Gauzak horrela, ondorioztatu genuen hoberena zela lehenbailehen sakonki ekitea morfologiaren azterketari. Beraz, itzulpen automatikoaren inguruko kontuak gerorako utzi eta lexikoa eta morfologiari ekin genien modu sakonean, eta tresna horiek geroago itzulpen automatikorako 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.

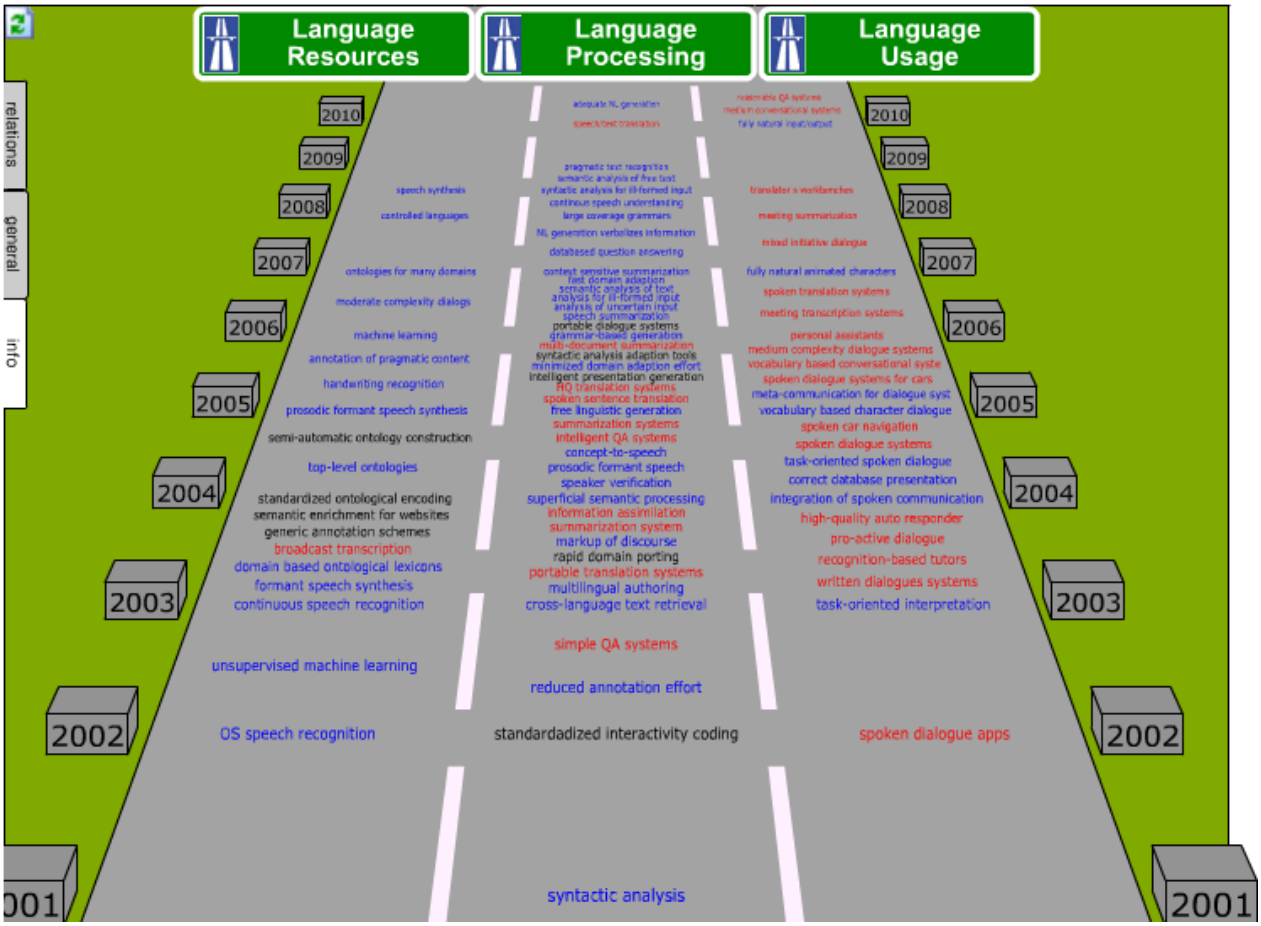
Taldearen 23 urteko ibilbidea estrategia horren arabera egin dugu. Nazioarteko forotan ere aurkeztu eta kontrastatu dugu beste ikerlari batzuekin (Sarasola, 2007). Ideia nagusiak ondokoak dira:

- **Hasieran oinarrizko baliabide eta tresna sendoak sortu behar dira**, eta geroago sortu merkatu-aplikazioak. Alderantziz ez dela egin behar. Produktuen artean bereizi 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-teknologiako teknikariek erabil ditzaten definitu dira. Baliabide, tresna edo aplikazio bakoitza noiz egin behar den aurreikusi izan dugu, ekoizpen prozesu hori optimizatu nahian.
- Horrez gain 1. irudian islatzen den **lexiko-morfologia-sintaxia-semantika progresioa aplikatu behar da,** hein handi batean behintzat.
- **Formatu estandarrak erabili behar dira**. Produktu bakoitza geroagoko produktu berrien garapenean ahalik eta modu zabalenean berrerabilia izatea da gure helburua. Sortzen diren produktuak formatu estandarren arabera[19] definitu behar ditugu, bai hartuko dituzten datuetan, bai itzuliko dituzten emaitzetan. Horrela berrerabilgarriak izango dira beste hainbat produktutan eta haien garapena modu inkrementalean egin ahal izango da.
- **Ahal den guztietan saiatu behar da software librea erabiltzen eta sortzen.** 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.

18 http://ixa2.si.ehu.es/saltmil/

19 XML and TEI dira estandar egokienak.



4. irudia. Baliabideen garapenari eta ebaluazioari begirako errepide-mapa (Busemann & Uszkoreit, 2004).

Egun euskarak hizkuntza-teknologiako lehenengo 60-80 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.

3.2 Nazioarteko erreferentziak

Atal honetan hizkuntz teknologien garapenaren inguruan argitaratu diren hainbat hausnarketaren berri eman nahi dugu. Kasu batzuetan baliabide urriko hizkuntzekin ere lotuta daude.

Hizkuntza bat gaurko behar teknologikoetan egokikuta egoteko estrategia edo jarraibide bat baino gehiago proposatu izan da nazioartean.

Gure taldean hasierako proposamen bat egin zen hizkuntza teknologiarako oinarrizko tresneria defintzeko (Agirre et al., 2002), *Basic toolkit for HLT* deitu genuena.

Krauwer-ek (2003) BLARK kontzeptua (*Basic LAnguage Resource Kit*) proposatu zuen, arrakasta lortuz, aurretik ELSNET (European Network of Excellence in Language and Speech) eta ELRA (European Language Resources Association) elkarteek proposatutakoari helduz. Ildo honetatik Maegaard-en taldeak (2004) BLARK bat proposatzen dute arabierarako eta Simov-ek eta lankideek (2004) beste bat bulgarierarako. Geroztik BLARK terminoa maiz erabili izan da literatura zientifikoan.

Streiter-ek beste batzuen lankidetzarekin 2006an argitaratutako lan batean (Streiter et al., 2006) hizkuntza ez-zentraletarako hainbat HLT

proiekturen berri ematen da. Testuinguru horretan, hainbat gomendio ematen dira erakundeek hizkuntza teknologia garatzeko banatzen dituzten laguntzak bideratzeko. *Non-central* terminoaren erabilerarekin batera azpimarratzekoa da software librearen alde egiten duen hautua. Forcada-k (2006) ere azpimarratzen du kode irekia erabiltzearen egokitasuna hizkuntza hauetarako itzulpen-sistemak garatzean.

ELSNET sareak, 2004an, baliabideen garapenari eta ebaluazioari begirako errepide-mapa[20] bat ere eskaini zuen (Busemann & Uszkoreit, 2004). Ikus

Testuinguru honetan mapa anitz argitaratu dira azken urteotan[21]. 2002ko gure proposamenean bezala diagramako elementuak hiru multzotan banatzen dira: *Language Resources / Language Processing / Language Usage* aipatutako mapetan eta *Language resources / Language Tools / Language Applications*. Gure estrategia hizkuntza baten garapen teknologikorako lehen urratsak modu sendoan emateko bideratuta dagoen bitartean ELSNETeko ekarpenean zerrenda askoz zehatzagoa da eta hizkuntza zentraletan jartzen da fokua (Europako hizkuntza ofizialak).

Borin-ek (2006 eta 2009) HLTek eremu urriko hizkuntzei irekitzen dizkieten aukerak aztertzen ditu, informazioaren gizartean hizkuntza aniztasunak duen garrantzia azpimarratuz. Ostler-en aipamen hau ere "a *language will not get by in the world of today unless it is equipped with a parser and a multi-million-word corpus of text*" ekartzen du.

Azken urteetako Europako proiektu berri batzuek, Clarin[22] eta Flarenet[23] esaterako, Europako hizkuntzetarako baliabideen eta tresnen garapena eta ustiapena lankidetzan eta koordinazio handiagoz egitea bultzatu nahi dute.

Eta azkenik aipatzeko, SALTMIL (*Speech And Language Technology for Minority Languages*) elkarteak hainbat batzar[24] antolatzen ditu HLT eta baliabide urriko hizkuntzak uztartuz.

4 Mugarriak eta etapak Ixa taldearen jardunbidean

Ixa taldearen 23 urteko ibilibidea errazago aurkeztearren etapaka banatu dugu denbora tarte hori. Ibilbide horietako mugarri nabarmenenak aukeratu ditugu etapak bereizteko, eta etapa bakoitzaren deskribapenean garatu diren produktu eta landu diren ikerketa-proiektu garrantzitsuenak aipatuko ditugu. Hauek izan dira aukeratu ditugun mugarriak:

- 1993: Morfologia eta Xuxen zuzentzaile ortografikoa
- 1996: Lexikoa eta EDBL datu-base lexikala
- 1999: Lematizatzailea
- 2002: Sintaxia
- 2005: Lexiko-semantika eta EuskalWornet
- 2009: Eleaniztasuna eta Matxin itzultzaile automatikoa
- 2010: Aplikazio aurretatuak: *Ihardetsi* galderak erantzuteko sistema

4.1 1988-1993: Morfologia eta Xuxen zuzentzaile ortografikoa

Esan bezala, hasieran gure lehenengo proiektuaren helburua euskararako itzulpen-sistema bat sortzea izan zen. Baina gure ahaleginei probetxu handiagoa ateratzearren laster itzulpen-kontuak geroago egiteko utzi eta momentuan morfologiari ekin genion modu sakonean. Ahalegin horretatik sortu zen urte gutxiren buruan zuzentzaile ortografikoa.

	1988-1993
Proiektuak Gipuzkoan (GFA)	Itzulpena Xuxen
Produktuak Morfologia	**Xuxen**

1. taula: Proiektu eta produktuak, 1988-1993.

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, hau izan daiteke duda bat idazterakoan: “Ondoko hitzen artean zein da batuaz erabili behar dudana *arbola* adierazteko*? Zuhaitz? Zuhatz? Zugaitz? Zugatz? Zuhaits? Sugatx?”*. Bestetik, euskara estandarraren definizioa berri samarra denez (beste inguruko erdaren estandarrekin konparatuta),

20 http://elsnet.dfki.de/roadmap.php

21 http://elsnet.dfki.de/roadmap.phpversion=LREC_2004

22 http://www.clarin.eu/

23 http://www.flarenet.eu

24 http://ixa2.si.ehu.es/saltmil/eu/activities/workshops/workshops.html

lexikoaren estandarizazioa oraindik bukatzear dagoenez, eta batzuetan estandarizazioan aldaketak gertatzen direnez (esate baterako, hasieran *eritzi*, eta *iharduera* hitzak erabili behar zirenak gaur egun *iritzi* eta *jarduera* idatzi behar dira) 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 dela XUXEN programa. Eta gainera 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 urteetan atera diren egokitzapen berriei esker XUXEN eskuragarriago dago. Lehen Word editorearekin bakarrik erabil zitekeen, orain erraz jar dezakegu martxan Mozilla Thunderbird-ekin, nabigatzailearekin Interneten bidez edozein mezu edo inprimaki betetzen ari garenean, edo Openoffice-ekin. Posible da beste edozein aplikaziorekin ere 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.

Xuxen-en inplementazioa hasieran programa propio bat izan zen. Geroago exekuzio azkarragoa lortzearren Xeroxeko tresnetara egokitu zen[25], eta zken urteetan software librera jauzi ahal izateko *hunspell* eta *foma* tresnak erabili izan dira (Alegria et al., 2009).

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 lehen bertsio bat integratu da XuxenIV bertsioan.

Xuxen programaren erabilera guztiz hedatuta dago gaur egun, datu hauetan ikus daitekeenez:

- 1998z geroztik Microsoft Officeko banaketa ofizial guztiek barruan daukate.
- www.euskara.euskadi.net webgunetik egin diren deskargak 20.000 baino gehiago izan dira.
- Firefoxerako deskargak 120.000 baino gehiago 2007-2011 tartean.
- OpenOffice-rako deskargak 7.000 baino gehiago izan ziren 2010. urtean.

25 www.stanford.edu/~laurik/fsmbook

4.2 1993-1996: Lexikoa eta EDBL datu-base lexikala

Xuxen zuzentzaile ortografikoaren garapenean erabili ziren lexikoa eta analizatzaile morfologikoa ondo antolatu ziren geroago bere mantentze-lanak errazteko eta beste aplikazioetan erabili ahal izateko.

	1993-1996
Proiektuak Eusko Jaurlaritzan	Xuxen
Proiektuak Gipuzkoan (GFA)	HAIN
Produktuak Lexikoa	**EDBL**
Produktuak Morfologia	Xuxen... Morfeus

2. taula. Proiektu eta produktuak, 1993-1996.

Hizkuntzaren lexikoaren biltegi orokorra da datu-base lexikala. Hiztegi elektroniko moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahiak dituen eskakizunak kontuan harturik antolatua. Horrek lexiko-deskribapenaren sistematizazio bat eskatzen du: sarreren kategoria-sistema bateratu eta homogeneoa, kategoria bakoitzeko elementuak behar den bezala deskribatzeko beharrezko diren ezaugarriak zehaztea, etab. *EDBLk* (euskararen datu-base lexikala) lehen bertsioan 60.000 sarrera zituen eta 147.700 inguru biltzen ditu egun —120.000 hiztegi-sarrera, 20.000 adizki eta 700 morfema ez-independente—, eta Ixa taldea arduratzen da egunean mantentzeaz (Aldezabal et al., 2001). Internet bidez kontsulta daiteke[26]. Hasieran zuzentzailearen oinarri lexikal gisa pentsatu bazen ere, gaur egun oinarri lexiko orokorra da eta hainbat tresna elikatzen ditu: analizatzaile morfologikoa, lematizatzailea, hitz anitzeko espresioen errekonizatzailea, entitate-espresioen errekonizatzailea, etab.

Informazio ez-estandarra ere gehitu zaio morfemetan eta hitzetan, hala nola forma dialektalak, errore tipikoak etab., beti ere dagozkien erabilpen estandar eta zuzenarekin lotuta. Gainera,

26 http://ixa2.si.ehu.es:7777/forms/frmservlet?config=lbdbl
Kontsulta sinplifikatua: http://ixa2.si.ehu.es/edbl/

hedapen dialektala eraman da aurrera, aldaerak integratuz, eta horixe izan da abiapuntua beste produktu berri batzuk sortzeko, esate baterako, Xuxen-B[27] bizkaieraren zuzentzaile ortografikoan, eta batua-bizkaiera bihurtzaile automatikoan[28].

Informazio sintaktikoa eta semantikoa (azpikategorizazioa, atributu semantikoak, etab.) barneratzeko ere diseinatuta dago, eta kasu askotan informazio hori osatuta dago.

Azken bi urteetan Euskaltzaindiak lideratutako proiektu baten barruan dago taldea (Lexikoaren Behatokia1) eta UZEI eta Elhuyar erakundeekin elkarlanean datu-basea aberastu egin da.

Aurreko baliabideetan oinarrituta analizatzaile morfologikoa eraiki genuen. Analizatzaile morfologikoa hizkuntza guztietan beharrezkoa izanda euskara bezalako hizkuntza eranskarien kasuan ezinbestekoa gertatzen da. Analizatzaile (eta sintetizatzaile) morfologikoaren zeregina hitz-forma osatzen duten morfemak ezagutzea (eta konposatzea) da, eta morfema bakoitzari dagokion informazio morfologiko-lexikala ematea. Erreminta hau oinarri da hainbat aplikaziotan, hala nola, zuzentzaile ortografikoan, karaktere-ezagutze optikoan (OCR), eta aplikazio sofistikatuago guztietan —itzulpen automatikoa, adib.—. Interneten erabil daiteke demo[29] bat.

4.3 1996-1999: Lematizatzailea

Lematizatzaile/etiketatzailea analizatzaile morfologikotik eratortzen da, eta hitz-forma baten lema eta kategoria ematen ditu, anbiguotasuna saihestu edo gutxitzearren testuingurua aintzat hartuz (Ezeiza et al., 1998). Garaian berritasun handia izan zen desabiguaziorako teknika estatistikoekin batera murriztapen-gramatikaren formalismoa erabiltzea (Constraint Grammar ingelesez), sistema konbinatua garatuz. Zeregin nagusia desanbiguazioa bada ere, beste egitekorik ere badu halako tresna batek, esate baterako, hitz anitzeko unitate lexikalen identifikazioa (lokuzioak, hitz-elkarketak, pertsonaizenak, etab.). Oso aplikazio interesgarriak dituzte lematizatzaileek, esate baterako, dokumentubilatzaileak, informazio-eskurapena, terminologia, lexikografia, etab.

	1996-1999
Proiektuak Europan	
Proiektuak Madrilen (MEC)	Item
Proiektuak Jaurlaritzan	Xuxen, EDBL, Item, Lematizatzailea
Proiektuak Gipuzkoan	Xuxen Idazkide
Produktuak.Apl. orokorra	Multimeteo
Produktuak Semantika	
Produktuak Sintaxia	
Produktuak Lexikoa	EDBL
Produktuak Morfologia	Xuxen Eustagger

3. taula: Proiektu eta produktuak, 1996-1999.

Geroko hainbat proiekturen atea zabaldu zuen programa honek, adibidez Espainiako zenbait talderekin lankidetzan burutu ziren Item eta Hermes proiektuak. Demo[30] bat erabil daiteke.

4.4 1999-2002: Sintaxia

Morfologia eta lematizazioa bideratuta hurrengo urratsa perpaus sinpleak sintaktikoki analizatzeko tresna izan zen. Baterakuntza gramatikako erregelatan oinarrituta zegoen Patr-Ixa (Aldezabal et al., 2003). Hitzen barruko analisi morfosintaktikoa ere egiten zuen.

Geroago sintaxirako beste tresna landuago batzuk garatu dira perpaus konplexuetatik ere informazio sintaktikoa hobeto atera ahal izateko:

- *Zatiak* (edo *Ixati* ere deitua) azaleko analizatzaile sintaktikoa. Esaldiko sintagmak edo chunkak bereizten dituena.
- *EDGK:* Dependentzia Gramatika. Esaldiko buruak (aditzak izaten dira normalean) mendeko elementuekin dituzten erlazioak markatzen dira dependentziazko gramatiketan.
- *Maltixa*[31]*:* Analizatzaile sintaktiko estatistikoa

27 http://www.azkuefundazioa.org/lan-tresnak/xuxen-bizkaieraz

28 http://www.eleka.net/berriak/berria.php?id=eu&a=1&b=1303197461

29 http://ixa2.si.ehu.es/demo/analisianali.jsp

30 http://ixa2.si.ehu.es/demo/analisimorf.jsp

31 http://sisx04.si.ehu.es:8080/maltixa/index.jsp

- Eihera[32]: Testuetan entitateak ezagutzeko tresna (pertsonak, tokiak, erakundeak).
- EPEC[33] eta *Ancora*[34] corpusak: EPEC zuhaitz sintaktikoen bankua da. Prozedura erdiautomatiko bat erabili zen etiketatzeko. Guztira 50.000 hitz dauzka. Ancora ingurunean espainiera, katalanera eta euskarazko treebank-ak biltzen dira.
- Erreus corpusa[35]: Ikasleen idazketa-erroreen korpusa.

	1999-2002
Proiektuak Europan	
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Hermes
Proiektuak Jaurlaritzan	Xuxen, Sintaxi lexikoa, Ixa taldea UZEI sinon-hizt.
Proiektuak Gipuzkoan	Berbasare, Gainternet
Produktuak. Apl. orokorra	
Produktuak Semantika	
Produktuak Sintaxia	Zatiak-Ixati
Produktuak Lexikoa	Elhuyar-Word
Produktuak Morf.	Xuxen Elhuyar-Word

4. taula: Proiektu eta produktuak, 1999-2002.

Urte horietan Elhuyar-Word hiztegi-kontsultarako programa sortu zen. Word testu-prozesadorean plugin gisa integratu zena. Tresna honek euskarazko edo gaztelaniazko edozein hitz hartuta, bere lemari dagozkion itzulpenak eskaintzen dizkio erabiltzaileari; Elhuyar Hiztegi Txikia (Euskara-Gaztelania/Castellano-Vasco) elebidunean agertzen diren itzulpenak hain zuzen. Kontsultatzen den hitzaren lema eta kategoria konbinazio posible guztiak erakutsiko ditu, eta bikote bakoitzari beste hizkuntzan dagozkion ordainak ere. Ildo horretatik, geroago euskara-frantserako bertsioa ere sortu du Elhuyarrek, eta UZEIk Word-erako plugin bat garatu zuen sinonimo-hiztegia erabiltzeko.

Garai hartako Hermes eta Gainternet proiektuetan lematizazioaren erabileraren aukerak aztertzen hasi ziren (Hermes, hemeroteka elektronikoak: bilaketa eleanitza eta erauzketa semantikoa).

4.5 2002-2005: Lexiko-semantika eta EuskalWornet

Hizkuntza ulertzea xede denean, eta morfologia eta sintaxiari buruzko informazioarekin aski ez denean, semantikari buruzkoarekin aberastu behar da programa. Anbiguotasun linguistikoa ebatzi ezina da, askotan, semantikaz baliatu ezean. Hizkuntza baten tratamendurako azpiegituran, osagai semantikoak ere behar du bere lekua, beraz. Eta semantika lexikala da, beharbada, osagai horren prestakuntzan landu beharreko lehenengo alderdia. Semantika lexikalak lexikoko elementuen artean dauden erlazio lexiko-semantikoak biltzen ditu: sinonimia, antonimia, hiperonimia/hiponimia (klase/azpiklase erlazioak), eta beste.

	2002-2005
Proiektuak Europan	Meaning
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	EuropenTrad, Hiztegia2002, Hizking21, Bilatzailea, RICOTERM2
Proiektuak Jaurlaritzan	Ixa taldea Hizking21-ETORTEK
Proiektuak Gipuzkoan	Hermes
Produktuak Semantika	**EuskalWordnet**
Produktuak Sintaxia	Erreus corpusa
Produktuak Lexikoa	UZEI sinon-hizt
Produktuak Morf.	Xuxen Eihera

5. taula: Proiektu eta produktuak, 2002-2005.

Erlazio lexiko-semantiko horiek sare semantiko moduko batean adierazten dira esplizituki. Ingelesezko sare semantikoen artean ezagunena-edo

32 http://ixa2.si.ehu.es/demo/entitateak.jsp

33 http://ixa.si.ehu.es/Ixa/resources/Treebank

34 http://clic.ub.edu/ancora/

35 http://ixa.si.ehu.es/Erreus

WordNet izenekoa dugu, eta haren euskararako egokitzapenari *Euskal WordNet* edo *EusWN*[36] deitzen diogu (Agirre et al., 2006).

EusWN hori *EuroWordNet*-en markoan garatu zen, euskal hitzak ingelesezko *WordNet*-era metodo erdiautomatikoz lotuz (ezagutzaren eskurapen automatikoa). Ingelesaren gehiegizko eragina saihesteko eta kalitate linguistikoa babesteko euskal *synset*-ak eskuz orraztu ziren geroago.

EusWN garatzeko taldeak egin zituen ahaleginak indartu egin ziren *MEANING* europar proiektuan parte hartzearekin (*MEANING*: Amaraun mailako hizkuntza-teknologia eleanitzen garapena). Lau hizkuntzatarako wordnet diferenteak lotu egin ziren *MCR* egitura lexiko-semantiko eleanitza[37] sortzeko.

Behin hitzen adierak zein diren definituta zegoela, esaldien ulerkuntzari ekin ahal izateko garrantzitsua zen jakitea bereizten esaldi konkretu batean zein den hitz bakoitzerako erabili den adiera. Horregatik Hitz-Adieren Desanbiguatze (HAD) sistema bat[38] garatu zen geroago, *Support Vector Machine* metodo ezagunean oinarrituta. Hasieran ingeleseko corpusen gainean bakarrik aplikatu zen, baina gaur egunean, EuSemCor semantikoki etikatuta dagoen corpusa[39] sortu denez gero, posible izan da euskararako ere entrenatzea.

2002an ELEKA enpresa sortu zen, spin-off moduan, Ixa eta Elhuyar fundazioaren lankidetzaren fruitua da. Aurretik sortutako produktuen kudeaketa informatikoa eta merkatu-garapenaz arduratuko zena, ixakideen jarduera ikerketan hobeto zentratzearren. Aldi berean Elhuyar fundazioak bere I+G atala antolatu zuen.

2002.ean ETORTEK deialdiko *Hizking21* proiektua abiatu zen. Eusko Jaurlaritzako ikerketa estrategikorako proiektu horretan lankidetzan hasi ginen arloko beste ikerketa zentro batzuekin: Elhuyar Fundazioa, EHUko Aholab ikerketa-taldeak eta Vicomtech eta Robotiker teknologia-zentroak. Hizkuntza-, hizketa- eta multimedia-teknologien alorrean ezagutza eta eskarmenturik handiena duen euskal taldea osatzen dugu elkarrekin. Hizking21 proiektuaren jarraipena izan dira gero *Anhitz* (2006-2008) eta *Berbatek* (2010-2012) proiektuak.

Hizking21 proiektuko emaitzen artean *ZT corpusa*[40] azpimarratu behar da. Zientzia eta teknologiaren alorreko euskarazko testu-bilduma egituratu eta etiketatua da, eta alor horietako euskararen erabilera ikertzeko baliabidea izatea du helburu nagusia. Corpus berezi edo espezializatua da, eta Ixa taldeak eta Elhuyar Fundazioak elkarlanean eratu dute. Corpus etiketatua da, bai testuaren egiturari eta formatuari dagokionez, baita linguistikoki ere. Testuko hitz bakoitzaren lema eta kategoria/azpikategoria etiketatu dira. Corpusaren lehen bertsio honetan, 8,5 milioi hitz daude, eta horietatik 1,9 milioi hitz eskuz berrikusi, desanbiguatu eta zuzendu dira.

4.6 2006-2009: Eleaniztasuna eta Matxin itzultzaile automatikoa

2006. urtean Matxin sortu zen, euskararako lehen itzultzaile automatikoa (Alegria et al., 2007). Aurreko etapako *Opentrad* eta *Europentrad* proiektuen barruan garatu zen, estatu espainiarreko lau hizkuntza ofizialen arteko itzulpena landu baitzen proiektu horietan. Lau unibertsitateren eta hainbat enpresaren arteko elkarlanaren emaitza izan ziren proiektu horiek. Parte hartu zuten unibertsitateak ondoko hauek izan ziren: Euskal Herriko Unibertsitateko Ixa taldea, Alacanteko Unibertsitateko Transducens taldea, Vigoko Unibertsitateko Linguistika Informatikoko Mintegia eta Kataluniako Unibertsitate Politeknikoko TALP taldea. Enpresa arduraduna Eleka Ingeniaritza Linguistikoa izan zen, Elhuyar Fundazioaren zein Galiziako Imaxin Software enpresaren laguntzarekin. Alacanten *Prompsit* izeneko enpresa bat sortu zen, proiektuaren emaitzak Herrialde Katalanetan zabaltzeko asmoz. Espainiako Industria, Turismo eta Merkataritza Ministerioaren laguntzaz garatu zen proiektua.

	2006-2009
Proiektuak Europan	Kyoto
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Know, OpenMT, IMLT, Praxem, Avivavoz
Proiektuak Jaurlaritzan	Ixa taldea Anhitz-ETORTEK

36 http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl

37 MCR, kontsulta on-line: http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl

38 HAD/WSD demoa: http://ixa3.si.ehu.es/wsd-demo

39 Eusemcor corpusaren demoa: http://sisx04.si.ehu.es:8080/eusemcor/

40 http://www.ztcorpusa.net/aurkezpena.htm

Proiektuak Gipuzkoan	Remixee, Prest
Produktuak. Aplikazio orokorra	Anhitz **Matxin**
Produktuak Semantika	MCR, WSD-Ixa
Produktuak Sintaxia	Ancora corp.
Produktuak Lexikoa	EDBL
Produktuak Morfologia	ZT corpusa Eulia

6. taula: Proiektu eta produktuak, 2006-2009.

Proiektu horien barruan bi teknologia sortu ziren, bat A*pertium* izenekoa, antz handia duten hizkuntzen artean itzultzeko; eta bestea M*atxin* izenekoa, egitura desberdineko hizkuntzen artean itzultzeko.

Matxin erabilera publikoko programa bihurtzen zen bitartean Itzulpengintza automatikoko ikerketak hurbilketa estatistikoan kokatzen ziren. Horrela OpenMT (2006-2009) proiektuaren barruan EUSMT itzultzaile estatistikoa sortu zen eta lehenengo sistema hibridoak (erregelak eta estatistika batuz) proposatu ziren. Proiektu horren segida den OpenMT2 (2009-2012) proiektu berrian sistema hibridoetan sakondu nahi da, ebaluazio-metodoetan, aurre-edizioan eta postedizio automatikoan.

Etapa honetan, 2006. urtean, europar mailan *Kyoto* proiektua abiatu zen. Aurreko etapako *Meaning* proiektuan sortutako hizkuntza prozesatzaileak hobetu ziren proiektu berri honetan eta domeinu espezifikoetako dokumentuetan kontzeptuak eta gertakizunak automatikoki erauzteko erabili ziren gero. Proiektuaren bukaeran informazioa eta dokumentuak bilatzeko teknikak garatu ziren horrela lortutako ezagutza-baseen gainean, eta, helburu horrekin erabilita, emaitza onak eman ditu hitzen adieren artean desanbiguatzeko balio duen UKB algoritmoak[41]. Espainiako MICINN ministerioak finantzaturiko *KNOW* (2006-2009) proiektu koordinatuan ere antzeko helburuak landu ziren, baita honen jarraipena den *KNOW2* (2009-2012) proiektuan, non eleaniztasunaren ikuspuntua ere lantzen den. Bestalde, MICINN ministerioko *IMLT* proiektuaren barruan sortutako tresna eta baliabide linguistiko horiek guztiak batera integratzeko eredu orokor bat eskaintzeko izan da. XML estandarrean oinarritutako proposamen sendo bat lortu da eta berau inplementatzeko balio duen *LibXml* programa-liburutegia sortu da. Proiektu hauen arteko elkarlana ikus daiteke puntu honetan: *IMLT* proiektuko XML eredu horren sinplifikazio bat onartua izan da Kyoto proiektuan ezagutza adierazpide gisa, *KAF eredu* deritzoguna.

Proiektu horietan alde batetik itzulpen automatikoan eta bestetik informazio-bilaketan lortutako sistemak prototipo batean erabili ahal izan ziren etapa honen bukaeran. ETORTEK deialdiko *AnHitz* proiektuan euskaraz hitz egiten duen 3D *avatar* bat sortu zen prototipo mailan. Zientzia eta teknologian aditua denez gai horien inguruko galderak erantzun ditzake, edo gai horietako termino bilaketa eleanitza egin eta emaitzak automatikoki euskarara itzuli. Ixa taldearen ekarpena batez ere galderak erantzuteko sisteman eta itzulpen automatikoan egon da. Baina prototipoan beste modulu batzuk integratu dira: 3D avatarra (VICOMTech), testu-ahots bihurgailu eleanitza (Aholab), euskarazko ahots-ezagutza (Robotiker, Aholab) termino-bilaketa eleanitza (Elhuyar), zientzia eta teknologiazko corpus eleanitzak (Elhuyar), eta azkenik modulu guztiak integratzeko sistema (Elhuyar).

4.7 2009tik gaurdaino: Aplikazio aurretatuak: *Ihardetsi*, galderak erantzuteko sistema

Anhitz proiektuan euskarazko galderak erantzuteko erabili zen modulua *Ihardetsi* sistema bihurtu da azken garai honetan, alegia, euskaraz egindako galderen erantzun zehatzak testu-bildumatan aurkitzen dituena. Aplikazio konplexu honek ohiko "Question Answering" (QA) sistemen ezaugarriak ditu (Ansa, 2006).

	2009...
Proiektuak Europan	Paths
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)	Know2, OpenMT2, Hybridosint, RTTH, TIMM, Ancora-corpus
Proiektuak Eusko Jaurlaritzan	Ixa taldea Berbatek-ETORTEK
Proiektuak Gipuzkoan	Langune
Produktuak	**Ihardetsi**

41 UKB algoritmoa deskargatu: http://ixa2.si.ehu.es/ukb/

Aplikazio orokorra	BASYQUE EUSMT
Produktuak Semantika	Eusemcor UKB
Produktuak Sintaxia	Maltixa EDGK
Produktuak Lexikoa	Lexkit Dicc. Escolar
Produktuak Morfologia	BertsolariXa LibiXaml

9. taula: Proiektu eta produktuak, 2009. urteaz geroztik.

Paraleloan beste aplikazio aurreratu baten eraikuntzan parte hartuko du taldeak Europa mailako PATHS proiektu berrian, Meaning eta Kyoto proiektuen ondorioa dena. Aurreko proiektuetan sortutako hizkuntza prozesatzaileak erabiliko dira gero domeinu espezifikoetako dokumentuetan kontzeptuak eta gertakizunak automatikoki erauzteko. Eta horrela lortutako ezagutza-baseetan informazioa eta dokumentuak bilatzeko teknikak garatuko dira, Europeana liburutegiaren esparruan.

Europeana Europako eduki digitalen liburutegi erraldoia da. Hainbat museo, liburutegi, agiri eta ikus-entzunezko bildumatarako sarbide irekia da. Bere helburu nagusia Europaren aniztasun kultural eta zientifikoaren zabalkundea erraztea da. Liburutegiak 15 milioi ale biltzen ditu hainbat formatutan (irudia, testua, audioa eta bideoa).

Interneteko liburutegi digitalei esker kultur-ondare diren material ugari daude eskuragarri gaur egun. Hala ere, kopuru erraldoi horiek nahasgarriak ere izan daitezke erabiltzaile arruntarentzat, zailtasunak izan baititzake aurkitutako informazio guztia interpretatzen. PATHS proiektuak pertsonalizatutako bisita gidatu interaktiboak eskaini nahi ditu, eta horri esker erabiltzaile arruntak ere eroso mugituko dira liburutegi digital horien barruan. Erabiltzaileari maiz proposatuko zaizkio berarentzat interesgarri izan daitezkeen antzeko edukiak eta, gainera, aurkitutako informazioa erraz interpretatzeko laguntza emango zaio.

PATHSek proposatzen duen nabigazio gidatu berri honek kontuan hartuko ditu bilduma digitalean zehar egin daitezkeen hainbat *ibilbide* ("path", ingelesez). Edozein gairi buruzkoa izan daiteke ibilbidea, adibidez, artista eta bere medioei buruz ("Picassoren margolanak"), garai historikoei buruz ("Gerra hotza"), lekuei buruz ("Venezia"), edota pertsonaia ezagunei buruzkoa ("Muhammad Ali"). Ibilbideak sortzeko eta jarraitzeko moduak hainbat izango dira, hala nola, aldez aurretik adituek definitutakoak, PATHS sistemak berak automatikoki proposatuak, edo, nahi izanez gero, erabiltzaileak sortutakoak ere. Beraz, eduki digitaletara iristeko era berritzailea eskainiko dio PATHSek erabiltzaileari, eta gainera erabiltzaileekin izandako esperientzia baliagarri izango zaio sistemari liburutegi digitala bera ere aberasteko.

Beste aplikazio bat hiztegiak editatzeko leXkit tresna[42] da (Alegria et al., 2001), bezero-zerbitzari arkitektura darabilena. Ezagutza teknikoren beharrik gabe, edizio-lana errazten dio lexikografoari. Sarrerei buruzko meta-informazioa baliatzen du funtzionalitate aurreratua eskaintzeko, esate baterako, testuinguruaren araberako atazak. Kubako *Diccionario Escolar*[43] hiztegi-aplikazioa (Miyares et al., 2010) tresna honekin sortu izan da.

BertsolariXa[44] aplikazioaren helburua (Arrieta et al., 2001) xumeagoa da, baina oso praktikoa: bukaera bat emanda, hitz errimatuak aurkitzen ditu. Lemak ez ezik, BertsolariXa gai da hitz deklinatuak eta aditz-formak ere eskaintzeko. Arloka iragaz daitezke emaitzak. Arau fonetikoak aplikatzeko aukera ere ematen du taldeko webgunean erabil daitekeen aplikazio honek.

Azken garai honetan nabarmena da taldea egiten ari den ahalegina nazioarteko sareetan parte hartzeko. Horrela sare hauetan integratu da:

- Clarin[45], Bere helburua hizkuntza-baliabide konputazionalak zabaltzea da giza zientzietako ikerketetan erabiliak izan daitezen.
- Flarenet[46] Fostering Language Resources Network.
- RTTH[47],Red Temática en Tecnologías del Habla.
- TIMM[48], Red Temática en Tratamiento de la Información Multilingüe y Multimodal.

42 http://sourceforge.net/projects/lexkit/

43 http://www.unibertsitatea.net/blogak/ixa/aaa

44 http://ixa3.si.ehu.es/tresnak/bertso/nagusia.html

45 http://www.clarin.eu/external/

46 http://www.flarenet.eu/

47 http://lorien.die.upm.es/~lapiz/rtth/

48 http://ararat.ujaen.es/timm

Bukatzeko, azpimarragarria da *Langune*[49], Hizkuntzen Industriaren alorreko Euskal Herriko enpresen elkartea sortu egin dela eta gure taldea tartean egon dela. Elkarte hau 2010an sortu da eta itzulpengintza, edukiak, irakaskuntza eta hizkuntzen teknologiaren alorreko 30 enpresatik gora elkartzen ditu.

5 Ondorioak

5.1 Teknologia garatzeko estrategia bat baliabide urriko hizkuntzetarako

Ixa taldearen jardunbidea oinarri hartuta baliabide urriko hizkuntzetarako baliagarri izan daitekeen estrategia azaldu dugu teknologia garatzeko. Orain dela 23 urte hasi ginen definitzen estrategia hori, eta beti izan da gure iparra jardunbidea planifikatzeko. Hori hobeto azaltzearren taldearen ibilbidearen traza erakutsi nahi izan dugu bi tauletan (ikus 10. eta 11. taulak). Taula horiek taldeak sortu dituen produktu eta proiektu nagusiak biltzen dituzte urteen eta ezagutza linguistikoen arabera ordenatuta. Ikus daitekeenez, etapa horien edukia eta ordena guztiz bat dator orain dela aspaldi definitu genuen estrategiarekin:

- **Hasieran oinarrizko baliabide eta tresna sendoak sortu ditugu**, eta geroago merkatu-aplikazioak. 10. taulan argi ikus daiteke hori.
- **Produktuen garapenean lexiko-morfologia-sintaxia-semantika progresioa erabili dugu**: hasierako produktuak oinarri-oinarri den morfologiaren gainean sortu ziren, geroago morfologiako produktuak hobetzen joan ziren bitartean lexikoan oinarritutakoak sortu ziren, geroago sintaxikoak, semantikakoak eta azkenik aplikazio aurreratuekin lotuta daudenak (itzulpen automatikoa, eta galderak erantzuteko sistemak batez ere).
- **Formatu estandarrak erabili ditugu** produktuen berrerabilpena erraztearren, XML formatuen erabilera zabalak eta LibiXa liburutegiaren sorkuntzak frogatzen dute hori.
- **Ahal izan den guztietan software librea erabili eta sortu dugu**. Hiru produktu ditugu eskuragarri Sourceforge biltegian, eta taldeak sortu dituen hainbat aplikazio publikoki erabil daitezke.

Estrategia horri jarraitu izanaren ondorio nabarmenak dira euskararen prozesamenduan egin diren aurrerapauso esanguratsuak.

Hizkuntzak sailkatzeko irizpide batzuk definitu ditugu hizkuntza-teknologian eta Interneten duten presentziaren arabera. Sailkapen horren arabera, duen hiztun kopuruagatik eta bere egoera soziolinguistikoarengatik, logikoena euskara laugarren mailan egotea litzateke, azkenaurreko mailan alegia.. Baina Ixa taldeak (beste eragile batzuen laguntzarekin) alor honetan egindako lanari esker euskara ez dago laugarren mailan, hirugarren mailan baizik. Estrategia horri jarraitu izanak aukera eman du horretarako. Gaur egun euskarak nolabaiteko presentzia du hizkuntz aplikazio informatiko gehienetan. Munduan dauden 7000 hizkuntzetatik 60 bat bakarrik dira maila horretara iritsi direnak, eta euskara horietako bat da.

Gure ustez gure estrategia eta ibilbidearen erreferentzia lagungarria izan daiteke oraindik hizkuntz teknologian sartuta ez dauden hizkuntzentzat, eta bereziki hizkuntz teknologian sartu ez baina IKTetan hasierako urratsak egin duten 190 hizkuntzentzat, alegia, sailkapeneko laugarren mailan sartzen diren hizkuntzentzat.

5.2 Aurrerapausoak alorrez-alor

Ikus ditzagun orain zein izan diren euskararen prozesamenduan egin diren aurrerapausoak alorrez alor (morfologia, lexiko, sintaxia, semantika eta pragmatika). Hizkuntzaren prozesaketaren bidean geratzen diren hutsuneak eta eginkizunak ere markatuko ditugu ildo horretan.

Euskararen **morfologia**ren azterketa ia osorik dago eginda eta inplementatuta. Morfologiaren alorrean gaur egun gure erronka inplementazio horien bertsio eraginkorragoak eta libreak eraikitzea da.

Sintaxi konputazionalaren tratamendua oraindik ikergai irekia da. Egun ezinezkoa da euskarazko edozein esaldi luze sintaktikoki ondo analizatzea. Ingelesarentzat gehiago landu da eta analizatzaile sintaktiko aurreratuagoak dituzte, baina beste hizkuntz nagusietan ere oraindik arazoak agertzen dira “esaldi errealak” osorik analizatu nahi izaterakoan. Egoera horretan, sintaxiari buruzko alorrean, hiru ikerlerro jorratzen ditugu: azaleko sintaxia (*Ixati* tresnak eramaten du aurrera eta testua zati sintaktikoetan banatzea du helburu), hitzen arteko dependentzia-erlazioak atzematea (EDGK), eta parser estatistikoa (Maltixa). Bitartean sintaktikoki etiketatuta dagoen EPEC corpusaren tamaina 10 edo 100 aldiz handiagoa egitea da helburua, prozedura semiautomatikoak erabiliz.

49 http://www.langune.com/

PRODUKTUAK	1988-1993	1993-1996	1996-1999	1999-2002	2002-2005	2006-2009	2009...
Produktuak. Aplikazio orokorra			Multimeteo			Anhitz **Matxin**	**Ihardetsi** BASYQUE EUSMT
Produktuak Semantika					**EuskalWN**	MCR WSD-Ixa	Eusemcor UKB
Produktuak Sintaxia				**Zatiak-Ixati**	Erreus corp.	Ancora corp.	Maltixa EDGK
Produktuak Lexikoa		**EDBL**	EDBL	Elhuyar-Word	UZEI sinon-hizt	EDBL	Lexkit Dicc. Escolar
Produktuak Morfologia	**Xuxen**	Xuxen... Morfeus	Xuxen **Eustagger**	Xuxen Elhuyar-Word	Xuxen Eihera	ZT corp. Eulia	BertsolariXa LibiXaml

10. taula: Produktu garrantzitsuenak urtea eta finantzazioaren arabera.

PROIEKTUAK	1988-1993	1993-1996	1996-1999	1999-2002	2002-2005	2006-2009	2009...
Proiektuak Europan					Meaning	Kyoto	Paths
Proiektuak Madrilen (MEC, MICINN Cicyt, Prontic...)			Item	Hermes	Hizking21 EuropenTrad Bilatzailea RICOTERM2 Hiztegia2002	Know OpenMT IMLT Praxem Avivavoz	Know2 OpenMT2 Hybridosint RTTH, TIMM Ancora
Proiektuak Eusko Jaurlaritzan		Xuxen	Xuxen EDBL Lematiz. Item	Xuxen Ixa taldea Sintaxi lexikoa UZEI sinon-hizt	Ixa taldea Hizking21 ETORTEK	Ixa taldea Anhitz ETORTEK	Ixa taldea Berbatek ETORTEK
Proiektuak Gipuzkoan (GFA)	Itzulpena Xuxen	HAIN	Xuxen Idazkide	Berbasare Gainternet	Hermes	Remixee Prest	Langune

11. taula: Proiektu garrantzitsuenak urtea eta finantzazioaren arabera.

Semantikaren bidean askoz ere lan gehiago gelditzen da egiteko. EuskalWornet taxonomia oinarri ikaragarria da, baina segitu behar da osatzen eta hobetzen. Halaber, Wikipedia iturburu oparoa da ezagutza semantikoa aberasteko. Ildo horretan, metodo automatikoak definitu behar ditugu hortik informazioa eraузteko. Hala ere, hitzen adieretatik harantzago joanda, perpausen interpretazio semantikoa lortu ahal izateko oraindik urrats asko eman behar dira, hasieraren hasieran baikaude oraindik. Jarraitu behar da aditzen azpikategorizazioa eta rol tematikoak aztertzen, batez ere informazio giltzarria eskaintzen dutelako gainontzeko arloen erresoluzio egokian, hala nola, desanbiguazio sintaktikoan, erreferentzia-kidetasunaren azterketan, etab. Corpus semantikoki etiketatuak behar dira gero ikasketa automatikoa erabili ahal izateko. Horrekin batera, hitz mailan adierak etiketatu behar dira, eta esaldi mailan aditzen azpikategoriak eta rol tematikoak.

Semantika konputazionalaren egoera oso hasierakoa bada ere, **pragmatika**rena askoz gordinagoa da, ohian basati landugabea dela esan daiteke. Diskurtsoaren egituraren azterketarekin hasi gara lanean berriki eta horretan bi bide nagusi definitu ditugu. Alde batetik, hizkuntzalaritzaren ikuspegitik aztertzen ari gara. Eta bestetik, saioak egiten ari gara hori bera lantzen ikasketa automatikoko sistemak erabiliz. Azkenik, bidean dagoen beste ikerlerro batek diskurtsoaren erlaziozko egitura zehazten dihardu. Elipsia, erreferentzia, hizketa-ekintzak (*speech acts*), hizketaren planifikazioa, hizlari-ereduak erabiltzea... Gai horiek guztiak zain daude.

Hizkuntzaren erabateko ulerkuntza automatikoa oraindik urruti dago. Oraingo ezagutza mugatua da, baina azken urteetan argi frogatu da teknologia ez-oso (ala partzial) hori gauza dela aplikazio praktikoak sortzen. Eta helburu horrekin jarduten dugu Ixa taldean. Hasieran lau partaide ginenak, orain 33 informatikari, 10 hizkuntzalari eta 3 teknikari gara. Euskal Herriko zazpi enpresekin lankidetzan gabiltza eta atzerriko beste bostekin. Spin-off erako bi enpresen sorkuntzan parte hartu dugu. 2002. urtetik Eusko Jaurlaritzak definitu zuen *Ingeniaritza linguistikoa* ikerlerro estrategikoa parte hartu dugu (Hizking21 eta Anhitz proiektuak) beste ikerketa-zentrorekin batera (Aholab, Elhuyar, Vicomtech eta Robotiker).

5.3 Oraingo ikerlerroak

Oinarrizko baliabide orokorrak sortzeko lerroaz gain, hauek dira gure azken proiektu estrategikoan (BerbaTek) ezarri ditugun ikerketa-lerro garrantzitsuenak:

- **Oinarrizko baliabideak**: hizkuntzen industriaren zenbait esparrutan erabil daitezkeen baliabideak eta teknologiak dira, edo gainerako alorren batean baliagarriak izan daitezkeen tresnetarako lehengai izan daitezkeenak. Esate baterako, testu-edo ahots-corpusa, lexikoiak, hiztegiak, ontologiak, gramatika konputazionalak, analizatzaile morfosintaktikoa, ahots-ezagutzea, ahots-sintesia, elkarrizketa-sistemak...
- **Itzulpengintza**: itzulpengintzaren sektorean erabil daitezkeen sistemak, itzulpen automatikoa, itzulpen-memoriak, ahots-ahots itzulpen-sistemak eta bikoizketa automatikoa kasu.
- **Edukiak**: edukien sektorea hobetzen lagundu dezaketen sistemak, hala nola, informazio-bilaketa (elebakarra, eleaniztuna, semantikoa, multimedia...), informazio-erauzketa, idazketan laguntzeko sistemak (zuzentzaileak, adibidez), ezagutzaren kudeaketa, galderei erantzuteko sistemak...
- **Irakaskuntza**: irakaskuntzaren eremuan erabiltzeko sistemak dira; adibidez, tutore pertsonalak, e-learning sistemak, ahoskera zuzentzeko sistemak, ariketen eta adibideen eraikitze automatikoa...

5.4 Giza talde baten ilusioa

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 2008tik abiatu zen izen bereko Europako master ofiziala. Unibertsitate mailako euskarazko master ofizial bakanetakoa da.

Hauek guztiak hamaika ahalegin eta ilusioren fruituak dira. Zenbat irakasle eta ikasle ibili garen, hor, lanean elkarrekin, heziketa-aukera hau euskaraz egin ahal izateko! 75 baino gehiago dira bide horietatik titulua eta heziketa berezitua jaso

duten teknikari/ikerlari berriak. 24 doktorego-tesi[50] sortu dira iturri horretatik. Ingeniaritza linguistikoan I+G horretan (Ikerketan eta Garapenean) arituko den komunitate zabal bat sortu dugu, baina aurrera egingo badugu, zabaldu egin behar dugu komunitate zientifiko hau. Hala biz!

6 Erreferentziak

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997. A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. 31-38. Oxford University Press. Oxford.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. 1998. A framework for the automatic processing of Basque. Proceedings of the Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada.

Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sarasola K., Soroa A. 2002. Towards the definition of a basic toolkit for HLT. LREC 2002.Workshop on Portability Issues in HLT. Las Palmas, Canary Islands.

Agirre E., Aldezabal I., Pociello E. 2006. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordNet. GOGOA aldizkaria ISNN 1577-9424 (pp. 237-266).

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K. 2003. Patrixa: A unification-based parser for Basque and its application to the automatic analysis of verbs. In Bernard Oyharcabal (ed.), Inquiries into the lexicon-syntax relations in Basque, Anuario de Filología Vasca "Julio de Urquijo" nº XLVI, pp 47-73. University of the Basque Country.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi M. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. IRCS Workshop on linguistic databases. Philadelphia (USA).

Alegria I., Arregi X., Artola X., Astiz M. , L. Ruiz Miyares. 2001. A Dictionary Content Management System. Proceedings EURALEX 2006 I, 105-109 (Turin, Italy). (ISBN 88-7694-918-6).

50 http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak

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. pp. 374-384. Cicling 2007.

I. Alegria, I. Etxeberria, M. Hulden, M. Maritxalar 2009. Porting Basque Morphological Grammars to foma, an Open-Source Tool. FSMNLP2009. Pretoria. South Africa.

Ansa O., Arregi X., Otegi A., Valverde A. 2006. An XML Framework for a Basque Question Answering System. 7th International Conference on Flexible Query Answering Systems. Milano, Italia.

Arrieta B., Alegria I., Arregi X. *An assistant tool for Verse-Making in Basque based on Two-Level Morphology. Literary and Linguistic Computing. Online ISSN 1477-4615 - Print ISSN 0268-1145 . Vol. 16, No. 1; pag 29-43; 2001 (Oxford University press).*

L Borin. Linguistic diversity in the information society. 2009 SALTMIL2009 Workshop: IR-IE-LRL Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country. ISBN 978-84-692-4940-6.

S. Busemann, and H. Uszkoreit (2004) Predicting the Future: Technology Roadmapping. In: ELSNews, (3) 2004.

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).

Forcada M. Open source machine translation: an opportunity for minor languages. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages. Genoa. 2006.

S. Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. International Workshop Speech and Computer. 27-29 October 2003, Moscow.

Maegaard B., Krauwer S., Choukri, K. and Jorgensen, L.D. The BLARK concept and BLARK for Arabic. Fifth International Conference on Language Resources and Evaluation, LREC. 2006.

Miyares Bermúdez, Eloína, Leonel Ruiz Miyares, Cristina Álamo Suárez, Celia Pérez Marqués, Xabier Artola Zubillaga, Iñaki Alegria Loinaz,

Xabier Arregi Iparragirre. 2010. *La segunda y tercera ediciones del Diccionario Básico Escolar*. Euralex2010. Leeuwarden (Herbehereak)

B. Petek. 2000. Funding for research into human language technologies for less prevalent languages, Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.

Sarasola K. 2007. Technology is an effective tool to promote use of Basque. ICML Colloquium on Language Revitalisation through Multimedia Technology, Pecs, Hungary.

Simov K., Osenova P., Kolkovska S., Balabanova E. and Doikoff D. A language resources infrastructure for Bulgarian. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC. 1685--1688. 2004.

O. Streiter, K.P. Scannell, M. Stuflesser (2006) Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. Machine Translation Journal. Volume 20, Number 4, pp. 267-289.

B. Williams, K. Sarasola, D. Ó´Cróinin, B. Petek. 2001. Speech and Language Technology for Minority Languages. Proceedings of Eurospeech 2001.

Wilson A., Archer D., and Rayson P. Corpus linguistics around the world. Rodopi. 2006.