

HIZKING21: Gaitasun linguistikoak dituzten sistemetan hizkuntza baliabide eta tresnak barneratzen

Inmaculada Hernaez¹, Eva Navas¹, Arantza Diaz de Ilarraza², Rafael Saiz³, Nora Mendoza⁴,
Imanol Madariaga¹

¹Electronika eta Telekomunikazioak Saila

²Lengoaia eta Sistema Informatikoak Saila

Universidad del País Vasco/Euskal Herriko Unibertsitatea

³Elhuyar

⁴VicomTech

jtpnacoe@bi.ehu.es

Abstract

In this paper, we present the main lines of the HIZKING21 project. Its main objective is to promote basic research in language engineering, orienting this investigation towards the requirements of the globalized environment of the present day. The project has a duration of three years and it will finish in April 2005. Our scope of work is the development of language technologies for the Basque language, as well as the integration of resources and tools for the language industry, both already existing resources and resources to be developed in this project, into different devices (PCs, PDAs, electrical household appliances, car equipment and so on). Our goal is to contribute to the easy and user-friendly interaction with all kind of devices, using language as the natural means of communication. The starting up of this project has been possible thanks to the advances and developments in the Natural Language Processing and Language Engineering for Basque language made by the participants of this project in the last fifteen years, as well as to the fact of sharing a vision of the more adequate strategy for the development of these technologies in the case of a minority language like standard Basque.

Laburpena

Artikulu honetan HIZKING21 proiektuaren ildo nagusiak azalduko ditugu. Bere helburu nagusia hizkuntza ingeniarietan oinarrituko ikerkuntza bultzatzea da, ikerkuntza gaur eguneko ingurune globalizatuaren beharrezanetara bideratuz. Hiru urtetako iraupeneko proiektua da, 2005eko apirilean bukatuz. Gure lanaren eremua Euskararako hizkuntza teknologien garapenean dago eta hizkuntza industriarentzako baliabide eta tresnak integratzea dispositibo ezberdinetan (PCak, PDAk, etxeko elektratresnak, automobileko ekipoetan eta abar), bai jadanik badauden baliabideak bai eta proiektu honetan garatutako baliabideak. Gure helburua mota guztietako tresnekin hartu-eman erraza eta eroso lortzen laguntzea da, hizkuntza komunikatzeko bide naturala bezala erabiliz. Proiektu hau abiarazi ahal izan da proiektu honen partaideek asken hamabost urteetan Euskararako Lengoaia Naturalaren Prozesaketa eta Hizkuntza Ingeniaritzan egindako aurrerapen eta garapenei esker, eta Euskara batua bezalako hizkuntza gutxitu batentzako teknologia hauek garatzeko estrategia egokiaren ikuspuntu bera edukitzeagatik.

Keywords: linguistic resources, corpora, speech recognition, text to speech conversion

Hitz gakoak: Hizkuntza baliabideak, korpusak, ahots ezagupena, testu ahots bihurtzea.

1. Sarrera

HIZKING21 (www.hizking21.org) osatzen duten taldeak gai bitaz arduratuta daude. Alde batetik, gizakien eta era guztietako tresnen arteko hartu-eman erraza eta intuitiboa egingo duten interfazeak garatzeko beharra. Hartu-eman hau betetzeko hizkuntzaren garrantzi handia argia da. Beste aldetik, jendeak ez luke bere ama hizkuntza alde batera utzi beharko hori egiteko.

Gaur egun, nahiz eta hizkuntza industriako produktu asko egon Ingeleserako, produktu hauek ez dira hainbeste hedatu beste hizkuntzetarako. Europan gizarte eleanitzean bizi garela kontuan izanik, guk, lurralde honetan lan egiten dugun ikerlari bezala, produktu eleanitzak garatzeko prestakuntza berezia daukagu. Eta Euskal ikerlari bezala, erantzukizun bat daukagu

Euskara informazioaren Gizartean integratzeko. Aldi berean, produktuak beste hizkuntzetara egokitu daitezkeen frogatzeko Euskara erabil daiteke, batez ere era bereko eskasia pairatzen duten hizkuntza gutxituetarako.

2. Lengoaia Naturalaren Prozesaketa eta hizkuntza gutxituak

Informazioaren gizarte modernoan iraun nahi duen hizkuntzak, lengoaiaren teknologiko produktuak behar ditu. Lengoiaren Teknologia oinarritzko ekarpena ematen ari dira informazio gizartearen arrakastarako, baina aplikazio gehienak ingelesez bakarrik daude prest. Honi aurre egiteko hizkuntza gutxituek ahalegin handia egin behar dute (Petek 2000) (Williams et al. 2001).

Hizkuntza gutxituen hizkuntza teknologien garapena hizkuntza hedatuen garapenaren ezberdina da alde askotan. Arrazoi bigatik batez ere:

- Alde batetik hiztunen multzoa txikia izan ohi da. Ondorioz, hauetariko hizkuntza gehienak ez daukate giza baliabide nahiko, dirulaguntza eskasak dituzte, eta etekin komertziala lortzea gehienetan helburu saila da.
- Beste aldetik, hizkuntza bakoitzaren tipologiari lotutako arazoak daude. Hizkuntza gutxituekin ez da beti posible beste hizkuntzekin garatutako teknologiak erabiltzea. Hau arau bidezko teknikekin gertatzen da, baina corpus-bidezkoetan ere bai, corpusak benetan ustiatzeko etiketatu egin behar dira, eta hau gehienetan arau bidezko metodoekin egiten da.

3. Euskararako erabilgarri dauden teknologia, baliabide eta tresnak.

Asken 15 urteetan hainbat ikertalde ibili dira Euskal Herrian arlo honetan lanean euskararako oinarritzko baliabide konputazionalak garatzeko helburuarekin. Lan honen buru Euskal Herriko Unibertsitateko talde bi izan dira:

- Aholab ikertaldea (bips.bi.ehu.es/ahoweb) ahots sintesi eta ezagupeneko teknologietan espezializatua. Euskal Herriko Unibertsitateko Elektronika eta Telekomunikazioak Saileko Seinalearen Tratamendua eta Telekomunikazioak Taldekoa da.
- IXA taldea (ixa.si.ehu.es) testu idatzia maila ezberdinetan prozesatzen espezializatua (morfologia, sintaxia, semantika; korpusak, itzulpen automatikoa, ...); taldea Euskal Herriko Unibertsitateko Informatika Fakultateko kidez osatua dago batez ere, eta UZEIko kidez.

Gaur egun erabiltzaileentzako prest dauden tresna aipagarrienak hauek dira:

- EUSLEM lematizatzaile/etiketatzailea, IXA taldeak UZEIrekin garatua, MORFEUS maila biko analizatzaile morfosintaktikoan eta EDBL datu-base lexikalean oinarritua. Orintsu “Constraint Grammar” formalismoan oinarritutako modulu desanbiguatzaile eta Multiberba Unitate Lexikalak prozesatzen dituen moduluarekin hobetu da (Ezeiza 2001).
- BasWN Euskarazko WordNet ezagutza base lexikala (ixa.si.ehu.es/Ixa/PilPilean/1022169813) IXA taldeak euskararako egokitua (Aguirre 2002).
- AhoTTS testu ahots bihurgailua (bips.bi.ehu.es/tts/tts_eu.html) Aholab taldeak garatua (Hernaiz 2001).

Tresna hauek erabiltzen dituzten hainbat aplikazio daude salgai:

- XUXEN zuzentzaile ortografikoa.
- GAIN lematizazioan oinarritutako web-bilatzailea.
- Lematizazioan oinarritutako web-hiztegi elebiduna on-line (Elhuyar Hiztegi Txikiaren plugina Microsoft Word 2000).
- MultiMeteo, eguraldi informeak sortzeko.
- AhoTTS testu ahots bihurgailu aplikazioa.

Korpus baliabideei dagokionez gaur egun Euskal testuen korpus garrantzitsu bi daude:

- OEH-Orotariko Euskal Hiztegiko korpus testuala: Euskara Batuararen sorkuntzararte (~1960) idatzitako testu guztiak jasotzen dituen lematizatu gabeko korpusa. 5’5 miloi berba inguru ditu.
- XX. mendeko Euskararen Korpus Estatistikoa: Oracle datu base erlazionalean inplementaturiko korpus lematizatua SGMLrekin etiketatutako egiturarekin. On-line kontsulta daiteke (www.euskaracorpora.net/XXmendea/Konts_arrunta_fr.html). 4.658.036 berba ditu.

Ahots errekurtsoei dagokionez ELDA (www.elda.fr) konpainiaren bidez datu base hauek daude eskura garri:

- 1060 hizlariko FDB SpeechDAT datubasea. Ahots ezagutze sistema komertziala entrenatu eta testatzeko erabili izan da.
- Bizkaifon (bideoteka.ehu.es) Bizkaierako herrialde ezberdinetako hizkera espontanea daukana.

Testu-Ahots Bihurketa eta Ahots Ezagupen Automatikoko sistemak garatzeko beste datu base batzuk garatu dira.

Asken urteotan, Euskal Herriko hainbat konpainia pribatu eta teknologi zentro hasi dira arlo honetaz arduratzen eta inbertitzen. Aldi berean erakunde gehiago konturatu dira elkarlana ezinbestekoa dela hizkuntza gutxituentzat teknologiak garatzeko. Elkarlan honen ondorio bat HIZKING21 proiektua da. Goian aipatutako IXA eta Aholab taldeetaz gain erakunde hauek dira HIZKING21en partaide:

- Vicomtech: (www.vicomtech.es) ikerketa aplikaturako zentro bat da, konputagailu bidezko elkarrekintzazko grafikoak eta multimedia digitala ikertzen dituen. INI-GraphicsNet Fundazioak eta EiTB taldeak sortua da.
- Elhuyar Fundazioa: (www.elhuyar.com) Euskararen gizarte-normalizazioa bultzatzen duen irabazi-asmorik gabeko erakundea da. Arlo hauetan lan egiten du: Lexikografia eta

terminologia, hiztegien argitalpena, zientzia eta teknologiaren gizarteratzea, testu-liburuak eta multimedia serbitzuak.

- Robotiker: (www.robotiker.com) Informazio eta Telekomunikazio Teknologietan espezializatutako zentro teknologikoa. Tecnalia Korporazio Teknologikoaren barne dago.

4. HIZKING21: helburu orokorrak

Aurrean azaldutako egoera izanik, HIZKING21en bultzatzaileak beharrezkoa ikusten dute: alde batetik, indarrak batu eta koordinatzea eta ezagupena eta baliabideak partekatzea Euskaren teknologien arloan garapenak eta etekinak sakondu, azkartu eta hedatzeko; beste aldetik, epe erdira produktu komertzialen garapena bermatuko duten prototipo eta aurreaplikazioen diseinu eta inplementazioa.

Behar hauei buruzko adostasuna izan da HIZKING21 proiektua sortzera eraman gaituena.

HIZKING21 proiekturako helburu orokor hauek proposatzen ditugu:

- Estrategia bateratua diseinatzea eta hizkuntza teknologien arloko I+Gan bikaintasun sarea sortzea, mundu mailako ildoari jarraituz.
- Ahots teknologien eta linguistika konputazionalaren eremuan Euskararako baliabide eta tresnak 2006 urtean gaur egun Ingeleserako dauden parekoak izatea.
- Hizkuntza teknologiek ematen dituzten aukeren ustiapena bultzatzea:
 - Edukien kudeaketa
 - Ahotsaren sintesi eta ezagupena
 - Dokumentuen sorrera eta itzulpen teknologia
 - E-ikasketa
 - Multimedia sistemak
- Gure industria linguistikoa nazioartera ateratzea erraztuko duten oinarriak ezartzea.

Proiektuaren helburu nagusia “gaitasun linguistikoak dituzten sistemak” deitutakoak garatzea da, erabiltzaileekin harreman eleanitza izatea bermatuko dutenak. Hau lortzeko proiektu osoa lan ezberdinetan banatu da:

- Hasteko, horrelako sistemak behar duten osagai guztiak zehaztu behar dira: ingurune bakoitzerako interfaze multimediei, sistema mota bakoitzak behar dituen baliabide eta tresnak, dauden osagaien garapen maila, eta abar. Pausu honen helburu nagusia zera da, argitzea zeintzuk diren jadanik prest dauden osagaiak, zeintzuk garatu behar diren goitik

behera Euskal Herrian eta zeintzuk moldatu daitezken Euskararako beste hizkuntza batekoetan oinarrituta.

- Bigarren pausuan aurrekoan zehaztutako lanak garatuko dira. Lanak taldeetan antolatu behar dira aipatutako garapen estrategiaren arabera eta erabili beharreko teknologiaren arabera. Zehaztutako talde teknologikoak hauek dira: Korpusak, Baliabideak, Tresnak, Oinarritzko teknologiak eta Aurre-aplikazioak. Talde tekniko hauek era koordinatuan lan egin behar dute HIZKING21en helburu bat lortzeko: hizkuntzaren bidez harreman erabilerraza duten tresnen diseinua, Informazioaren Gizarako Teknologien erabileraren hedapenari lagunduz bere erabileraren zailtasuna gutxituz.
- HIZKING21 proiektuan egin beharreko beste lan garrantzitsu bat hizkuntza teknologien arloan goi-mailako heziketa bultzatzea da, Euskal Herriko ikerketa gaitasuna areagotzeko. Hori lortzeko, Europako erreferentziazko zentro garrantzitsuekin, unibertsitate eta enprekin, jende trukaketak aurreikusi dira, graduatu ondorengo kurtso eta doktorego tesiekin batera. Aipatutako zentroekin elkarlan akordioak lortu behar dira.
- HIZKING 21en helburuak lortzeko ezagupenaren elkar-trukaketa ere beharrezkoa da, hizkuntza gutxituetan lanean dauden ikertzaile komunitate ezberdinen artan agertu daitezken elkar-indarketak eskuratu ahal izateko. Beraz, zaintza teknologikoa eta teknologiaren hedapenak proiektuan murgilduriko jendearen indar handiak hartuko ditu. Aurreikusitako ekintzak zaintzarako metodo sistematikoak, azalpen batzarrak, lortutako emaitzak konferentzia eta aldizkari espezializatuaren argitaratzea, arloko mintegi eta kurtsoak, eta abar dira. HIZKING21 proiektuaren informazio aipagarri guztia publikatuko duen web lekua dago ere (www.hizking21.org), eta proiektu osoaren garapenean zehar informazio trukatzeko toki garrantzitsua bilakatuko da.

5. HIZKING21:helburu zehatzak LNPN

HIZKING21 proiektuan arlo garrantzitsuenak hauek dira: I+G, heziketa, azpiegitura, nazioarteko elkarlana, hedapena eta behatoki teknologikoaren sorrera. Helburu zehatzak finkatu ditugu oinarritzko ikerketan, baliabideen sorreran, lanabesen garapenean eta prototipo eta aurreaplikazioen garapenean. HIZKING21en muina ikerkuntza eta garapen proiektuak burutzea da. Lehentasuna daukaten proiektuak hauek dira:

- Oinarritzko baliabide linguistikoak garatzea
- Oinarritzko lanabesak garatzea

- Baliabideak eta tresnak integratzeko teknikak eta oinarritzko metodoak
- Pertsona/Makina Interfazea eta bere integrazioa multimedia inguruneetan.
- Wireless teknologia eta gaitasun linguistikodun sistemei loturiko hardwarea.

Ondoren lehenengo eginbehar biak azaltzen dira.

5.1. Oinarritzko baliabide linguistikoen garapena

Korpus multzo bat, datu-base lexikalak eta hiztegi elektronikoak burutu edo sortuko dira proiektuaren garapenean zehar.

Idatzitako korpusak:

- I korpusa. Etiketa gabeko testu landugabe bilduma (XML arina).
- II.korpusa. Adizkerak lema, POS eta azterketa morfosintaktikoarekin etiketatuak (eskuz zuzenduta).
- III korpusa. Sintaktikoki etiketatutako testua (eskuz zuzendua).
- IV korpusa. Semantikoki etiketatutako testua berben esanahia desanbiguatu ondoren (eskuz zuzendua).
- Va korpusa. Korpus eleanitza (XML arina).
- Vb korpusa. korpus eleanitz eta paraleloa.

Ahozko korpusak:

- Ahotsaren Ezagupen Automatikoko lanetarako testu korpusen garapena.
- Euskararen aldaera guztietarako ahozko korpusaren sorrera.
- Kalitate oneko testu-ahots bihurketa sistemetan erabiltzeko ahozko korpusa.
- Emoziodun hizkeraren ahozko korpus baten garapena.

Datu-base lexikalak:

- POS, informazio sintaktikoa, berba multzo unitateak, aditzen azpikategorizazioa eta kokapenaren gaineko informazioa daukan datu base lexikala.
- Informazio semantikoa eta ontologikoa lotuta daukan datu base lexikala.
- Datu base lexiko-semantikoa. Kontzeptuen taxonomia.

Hiztegi elektronikoak:

- Datu base lexikografiko eta terminologikoak datu banku baten bateratzea.

- Lan ingurune lexikografiko baten diseinu eta egitea.
- Lan ingurune terminologiko baten diseinu eta egitea.

5.2. Oinarritzko lanabesen garapena

Proiektuan hurrengo lanabesak hobetu edo sortuko dira:

- Ahotsaren prozesaketa: Hiztegi Handiko Beba Jarraiazen Ezagutza, goi-mailako Testu Ahots Bihurketa sistema, Ahots Ezagupen Automatikoko sistema.
- Sintaxia: berba multzo unitateen identifikazioa, etiketatze sintaktikoaren zehazpena, aztertzaile sintaktikoa.
- Semantika: dokumentuen sailkapena, entitateen identifikazioa eta prozesaketa, berben esanahiaren desanbiguzioa.
- Pragmatika eta Diskurtsoa: anafora pronomialen ebazpena, elkarhizketa sistema baten egitura, helburu eta elementuen zehazpena.
- Korpusak aztertzeko tresnak: XMLen etiketatutako korpusetatik informazioa jasotzea (bat-etortzeak, estatistikak, kokapenak...), etiketatutako korpusetatik terminologia ateratzeko tresna linguistiko eta estatistikoak, etiketatutako korpus eleanitzetatik baliokidetasun lexiko eta terminologikoak ateratzeko tresna linguistiko eta estatistikoak.

6. Eskerrak

Proiektu hau Eusko Jaurlaritzako Industria, Merkataritza eta Turismo saileko Etorrek programak finantzatzen du zati batean.

7. Aipamenak

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

Williams, B.; Sarasola, K.; Ó'Cróinin, D.; Petek, B. (2001). *Speech and Language Technology for Minority Languages*. Proceedings of Eurospeech.

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, Montreal (Canada). August 10-14.

Agir, 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. Mysore (India).

Hernaiz, I.; Navas, E.; Murugarren, J.L.; Etxebarria, B. (2001). *Description of the AhoTTS Conversion System for the Basque Language.* Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis.