BerbaTek: euskararako hizkuntza teknologien garapena itzulpengintza, edukien kudeaketa eta irakaskuntza arloetan

Igor Leturia
Elhuyar Fundazioa
i.leturia@elhuyar.com

Kepa Sarasola, Xabier Arregi, Arantza Diaz de Ilarraza
IXA Taldea, Euskal Herriko Unibertsitatea
[kepa.sarasola,xabier.arregi,a.diazdeilarraza]@ehu.es

Eva Navas, Iñaki Sainz
Aholab Taldea, Euskal Herriko Unibertsitatea
[eva.navas,inaki.sainz]@ehu.es

Arantza del Pozo, Aitor Álvarez
Vicomtech-IK4
[adelpozo,aalvarez]@vicomtech.org

David Baranda, Urtza Iturraspe
Tecnalia
[david.baranda,urtza.iturraspe]@tecnalia.com

Abstract

Basque is both a minority language (only a small proportion of the population of the Basque Country speaks it) and also a less-resourced language. Fortunately, the Basque regional government is committed to its recovery, and has adopted policies for funding, among other things, language technologies, a feld which a language aiming to survive cannot dispense with. BerbaTek was a 3-year (2009-2011) strategic research project on language, speech and multimedia technologies for Basque carried out by a consortium of five members, all prominent local organizations dedicated to research in the above-mentioned areas, and partially funded by the Departments for Industry and Culture of the Basque Government. Collaboration in BerbaTek allowed to carry out a great amount of both basic and applied research. In addition, various prototypes were developed to show the potential of integrating the developed technologies to the language industry sector.

Laburpena

Euskararen garapen eta hedapenerako guztiz beharrezkoa da ordenagailu bidezko bere tratamendua ahalbidetzea. Horretan aritu ginen hiru urtez (2009-2011) BerbaTek ikerketa estrategiko proiektuan, hizkuntza-, ahots- eta multimedia-teknologiak garatzen. Arlo horietan jakintza eta esperientzia zabaleko bost partzuergokideren artean eraman genuen aurrera BerbaTek proiektua: Elhuyar Fundazioa, Euskal Herriko Unibertsitateko IXA eta Aholab taldeak eta Vicomtech-IK4 eta Tecnalia zentro teknologikoak. Proiektuak Eusko Jaurlaritzaren Industria eta Kultura sailen laguntza ere jaso zuen. Kolaborazio honi esker, BerbaTek proiektuan ikerketa ugari egin zen, bai oinarrizkoa eta bai aplikatua; baina horretaz gain, hainbat prototipo garatu ziren teknologia hauek hizkuntzen industrian, hau da, itzulpenak, edukien kudeaketak eta irakaskuntzak osatzen duten sektorean, egin dezaketen ekarpenaren eta izan dezaketen potentzialaren erakusgarri.

Keywords: Natural Language Processing; Speech Processing; Less-Resourced Languages

Hitz gakoak: Lengoaia naturalaren prozesamendua, Ahotsaren prozesamendua, Baliabide gutxiko hizkuntzak.

1. Sarrera

Dakigunez, euskararen zabalkundea ez dago uniformeki banatuta ez geografikoki ezta sektoreka ere. Azken banaketa honi dagokionez, beharbada industria eta, bereziki, IKTen arloa izan daitezke egoera okerrenetakoan daudenetako batzuk. 2012an META-NET erakundeak (hizkuntza- eta ahots-teknologiak lantzen dituzten Europako erakunde anitz batzen dituen elkartea eta bere arloko ezagun eta garrantzitsuenetakoa) euskara jarri zuen iraungitze digitalaren arriskuan dauden hizkuntzen artean (Hernáez et al. 2012). Egungo informazioaren gizartean bizirik iraun nahi duen hizkuntzak derrigorrezkoa du alor horietan presente egotea, eta horrek hizkuntza- eta ahots-teknologiak eskura izatea eskatzen du. Euskarak, beste hizkuntza gutxituak bezala, esfortzu handia egin behar du erronka honi aurre egiteko (Williams et al. 2001).

Testuinguru honetan, BerbaTek (http://www.berbatek.com) 2009-2011 urteen artean aurrera eramandako hizkuntza-, ahots- eta multimedia-teknologien inguruko ikerketa estrategikoko proiektua izan zen. Proiektuaren kontsortzioa Elhuyar Fundazioak, Euskal Herriko Unibertsitateko IXA eta Aholab taldeek eta Vicomtech-IK4 eta Tecnalia ikerketa-zentroek osatu genuen. Proiektua parte batean diruz lagundu zuten Eusko Jaurlaritzako Industria eta Kultura sailek.

Kontsortzioko kideek 2002tik kolaboratu izan zuten antzeko bi proiektutan, Hizking (Diaz de Ilarraza et al. 2003) eta AnHitz (Arrieta et al. 2008), non euskararentzako hainbat oinarrizko baliabide, tresna eta aplikazio informatiko garatu ziren. BerbaTek horien

jarraipen naturala izan da, oraingoan sektore jakin bateko aplikazioari enfokatuagoa.

BerbaTek partzuergoko kideok uste osoa dugu baliabide urriko hizkuntzentzako teknologien ikerketa eta garapena lau puntu jarraituz eraman behar dela aurrera: (1) estandarizazio handia, (2) kodearen irekitasuna, (3) baliabide, tresna eta aplikazioen berrerabilpena, eta (4) diseinu eta garapen inkrementala. Hizkuntza- eta ahots-teknologietan euskara antzeko tamaina eta ezaugarridun beste hizkuntza batzuen aldean konparatiboki hobeto badago gidalerro hauek jarraitu direlako izan dela uste dugu (Alegria et al. 2011).

2. Partzuergoa

Vicomtech-IK4 (http://www.vicomtech.org) ikerketa aplikatuko zentro bat da. Bere ikerketa lerro nagusiak konputazio grafikoa, elkarrekintza eta multimedia dira. Bertako hiru taldek hartu zuten parte BerbaTek proiektuan: i) Ahots eta Hizkuntz Teknologien taldea, ii) 3D Animazioko taldea, eta iii) Eduki Audiobisualen Analisiko taldea.

Tecnalia (http://www.tecnalia.com) ikerketa aplikatuko zentro pribatua da, Informazio eta Komunikazio Teknologietan espezializatua.

Elhuyar Fundazioa (http://www.elhuyar.org) irabazi asmorik gabeko erakundea da, euskara eta zientzia elkartzeko asmoz sortua. Elhuyar hiztegien, irakaskuntzarako software eta bestelako materialaren, multimedia produktuen, *plugin*en eta itzulpen automatikoaren merkatuetan ezaguna eta ongi kokatua da. 2001etik hizkuntza-teknologien ikerketa eta garapeneko arloa du (http://www.elhuyar.org/hizkuntza-zerbitzuak/EU/I-G-unitatea).

IXA Taldea (http://ixa.si.ehu.es) 43 ikerlarik osatutako EHUko taldea da, Lengoaia Naturalaren Prozesamenduan edo LNPn espezializatua. Egun esku artean dituen proiektu nagusiak PATH, OpeNER eta NewsReader europar STREP proiektuak dira.

Aholab Taldea (http://aholab.ehu.es/) EHUKO ikerketa-taldea da, esperientzia zabala duena ahots-teknologietan eta seinaleen prozesamendu digitalean. Aholab-ek garatu zuen euskarazko lehen TTS sistema komertziala, AhoTTS (http://aholab.ehu.es/tts/), eta baita euskararentzako publikoki eskuragarri dauden ahots-baliabide eta -tresna gehienak ere.

3. Helburuak

BerbaTek proiektuaren helburu nagusia hizkuntza-, ahots- eta multimedia-teknologien ikerketa zen, Euskal Herriko hizkuntzen industriaren sektore ekonomikoari laguntzeko oinarri teknologikoa jartzeko.

Erronka nagusia euskara prozesatzeko teknologiek sektore horretako produktu eta zerbitzuen errendimendua, inpaktu soziala eta lehiakortasuna hobetzeko gai zirela frogatzea zen. Erronka honek eskatzen zuen proiektuko kideek ikerketatik haratagoko aurrerapauso esanguratsu garrantzitsu bat ematea eta proiektuaren emaitzak aplikazio errealetan integratzea. Puntu hau bereziki aipagarria da, ikerketan garatutako oinarrizko baliabide eta tresnei sendotasuna ematea eskatzen baitu.

Enpresa gehienei kosta egiten zaie pauso hau ematea, garestia eta komertzialki ez errentagarria deritzotelako. BerbaTek-eko kideok gure gain hartu genuen, inbertsio sozialtzat kontsideratuz. Euskara bezalako hizkuntzentzat unibertsitateetan eta ikerketa zentroetan garatzen dira tresnak, eta hauek industriaren eszenario errealetara egokitzea berebizikoa da gure ustez.

BerbaTek, beraz, aplikazioetara lerratuta zegoen. Oinarrizko ikerketa ahaztu gabe, asmoa zen aplikazio esperimentalak ere aurkeztea, geroago are garatu zitezkeenak eta enpresek produktu edota zerbitzu bihurtu. Bereziki, ahots- eta hizkuntza-teknologiok hizkuntzen industrian izan ditzaketen aplikazioa interesatzen zitzaigun, hau da, ondoko sektoreok osatutako industrian:

- Itzulpena: interpretazioa, bikoizketa, lokalizazioa, giza-itzulpena...
- Edukien industria: Internet, ikus-entzunezkoak, komunikabideak, offline eta online argitaletxeak, multimedia enpresak...
- Irakaskuntza: hizkuntzen irakaskuntza, oinarrizko hezkuntza, lanbide hezkuntza, unibertsitate hezkuntza, helduen heziketa iraunkorra...

4. Garatutako baliabide, tresna eta aplikazioak

BerbaTek partzuergoa osatzen duten kideek 90eko hamarkadatik ari dira lanean euskararentzako teknologiak lantzen. Tresna eta baliabide oinarrizkoenak (lematizatzaileak, etiketatzaileak, datu-base lexikalak, ahots datu-baseak, hiztegi elektronikoak eta abar) aurretik eginak zeuden, baina horietako gehienak BerbaTek proiektuan zehar hobetu ziren, eta beste berri asko ere sortu ziren. Egindako lan nagusiak honela sailka ditzakegu:

- Webetik corpusak osatzeko tresnak (elebakarrak eta eleaniztunak, orokorrak eta espezializatuak, konparagarriak eta paraleloak) eta hauek erabiliz eraikitako hainbat corpus.
- Dependentzia sintaktikoen analizatzaileak, analizatzaile semantikoak eta esaldi eta sintagmen mugak identifikatzeko sistemak.
- Terminologia erauzketa corpusetatik eta hiztegien eraikitze automatikoa.
- Ontologia orokor eta espezializatuak eta bilaketa semantikoa.

- Bilaketa eleaniztuna eta galderen erantzute automatikoa.
- Itzulpen automatikoko sistemak (erregeletan oinarrituak, estatistikoak eta hibridoak).
- Bideotan ahots segmentuak detektatzeko teknikak eta testuaren denbora-lerrokatzea.
- Ahots ezagutzako motorrak eta testu-ahots bihurtzaileak.
- Avatar hizlariak.
- Idazteko laguntzak eta ariketen sorkuntza automatikoa.

1., 2. eta 3. taulek zerrendatzen dituzte sortutako baliabide, tresna eta aplikazio hauek modu zehatzagoan.

Corpus baliabideak
Euskararen Prozesamendurako Erreferentzia Corpusa (EPEC), 300.000 hitzeko corpusa.
EusPropbank eta bere garapenerako tresnak (Aldezabal et al. 2010).
AhoSyn, euskarazko ahots datu-base handia (6 ordu hizlariko) (Sainz et al. 2012).
AhoSpeakers, ahots-bihurketarako diseinatutako ahots datu-basea (Sainz et al. 2012).
AhoEmo3, emoziodun ahots sorkuntzarako diseinatutako ahots datu-basea (Sainz et al. 2012).
Euskarazko corpus orokor handi bat (100 milioi hitzetik gora) webetik automatikoki bildua (Leturia 2012).
Ontologia baliabideak
Euskal WordNet (Pociello et al. 2011), WordNet-en euskarazko bertsioa.
WNTerm (Pociello et al. 2008), WordNet + zientzia eta teknologiako 25.000 termino.
Termide, corpusetatik automatikoki ontologiak eraikitzea.
QAWS, Linked Data gainean galderak erantzutea.
Hiztegi baliabideak
Hainbat hiztegi elebidun automatikoki sortuak pibote hizkuntza bat erabiliz (Saralegi et al. 2012).

1. taula: BerbaTek-en sortu edo hobetutako baliabideak.

Analisi tresnak
Dependentzia analizatzailea (Bengoetxea and Gojenola 2010; Agirre et al. 2011).
UKB (Agirre and Soroa 2009), grafikoetan oinarritutako adieren desanbiguazioa.
ArikIturri (Aldabe 2010), corpusetatik ariketen sorkuntza automatikoa.
Web as corpus tresnak
Co3 (Leturia et al. 2009), corpus eleaniztun konparagarriak automatikoki biltzea.
PaCo2 (San Vicente and Manterola 2012), corpus paraleloak automatikoki biltzea.

2. taula: BerbaTek-en sortu edo hobetutako tresnak.

Hiztegien sorkuntza automatikoa
AzerHitz (Saralegi et al. 2008), corpus konparagarrietatik termino-ekibalentzien erauzketa.
PiboLex (Saralegi et al. 2012), pibote-hizkuntza bat erabiliz hiztegi berriak eraikitzea.
Fraseologia eta esapide idiomatikoak erauztea (Gurrutxaga and Alegria 2011).
Informazio bilaketa
Ihardetsi (Ansa et al. 2008; Agirre et al., 2009), galderak erantzuteko sistema.
Elezkari (Saralegi and López de Lacalle 2009), bilaketa eleaniztuna (euskara, gaztelania eta ingelesa).
Itzulpen automatikoa
Opentrad-Matxin (Alegria et al. 2007; Alegria et al. 2008; Mayor et al. 2011), kode irekiko erregeletan oinarritutako gaztelania-euskara itzulpen automatikoko sistema.
EUSMT, gaztelania-euskara itzulpen automatiko estatistikoa (Labaka 2010).
Ahots sintesia
AhoT2P, euskara estandarrerako letra-alofono transkribatzailea.
AhoTTS_Mod1, ahots sintesirako prozesatzaile linguistikoa.
AhoTTS, testu-ahots bihurtzaile modularra euskara, gaztelania eta ingeleserako.
HTS-n oinarritutako TTS sistema (Erro et al. 2010), vocoder propioarekin (Erro et al. 2011a).
AhoTTS hibridoa, ahots sintesi estatistikoaren eta unitateen aukeraketaren abantailak konbinatuta (Sainz et al., 2010).
Ahots ezagutza
AhoSR (Odriozola et al. 2012), ahots ezagutzako motorea.

3. taula: BerbaTek-en sortu edo hobetutako aplikazioak.

5. Prototipoak

Proiektuan zehar, demo batzuk sortu genituen ahots, hizkuntza eta multimedia teknologien erabilgarritasuna eta beraien konbinazioen potentziala erakusteko, bereziki aipatutako itzulpenen, edukien kudeaketaren eta irakaskuntzaren arloetan. Hauek dira egindako demo horiek:

- Dokumentalak automatikoki euskarara bikoizteko tresna, gaztelaniazko transkripziotik abiatuta.
- Zientzia eta teknologiazko edukiaren bi bilaketa motor semantiko multimedia eta eleaniztunak.
- Hizkuntzen irakaskuntzarako tutore pertsonalizatu bat ahotsez gidatutako avatar baten bidez, automatikoki sortutako gramatika eta ulermen ariketekin, idazteko laguntzekin (hiztegiak, zenbakien idazketa, inflexioa...) eta ahoskeraren ebaluazio automatikoa.

5.1. Dokumentalen bikoizketa automatikoa

Filmak automatikoki bikoiztea erronka zaila da oraingoz (ahots asko, hizkera kolokiala, abiadura

ezberdinak...), baina dokumental-mota batzuekin (hizlari bakarra, off-eko ahotsa, ezpainekin koordinazioa ez da beharrezkoa edo garrantzitsua...) ongi funtzionatzen duen demo bat egitea lortu genuen. Garatutako aplikazioaren egitura orokorra 1. irudian ikus daiteke. Gaztelaniaz dagoen dokumental bat eta han esaten denaren transkripzio bat emanik (transkripzio hori nahi bada automatikoki lor daiteke, merkatuan egon bai baitaude diktaketa-programak gaztelaniarako), Vicomtech-IK4ren denbora-lerrokatzearen teknologiaren bidez azpititulu-fitxategi bat lortzen da (transkripzioa, baina esaldi bakoitzaren hasierako eta bukaerako uneekin). Gero, IXA Taldearen Matxin itzultzaile automatikoak euskarara itzultzen ditu azpituluok, eta Aholab-en testu-ahots bihurketa-teknologiak euskarazko ahots sinkronizatua sortzen du. Demo hori arrakastaz aplikatu zaie Elhuyarrek egiten duen Teknopolis saioko hizlari bakarreko atalei. Demo hau online dago[1].

Ahotsaren eta testuaren lerrokatze automatikoa ahots-ezagutzako teknologian oinarrituta dago. Ahots-ezagutzako sistema bat transkripzioko testua ezagutzera behartzen da, fonema eta hitz mailako denbora-informazioak jasotzen dira. Hala, hitz bakoitzaren hasiera eta bukaerako denborak lortzen dira eta azpituluak bideoaren irudiarekin sinkronizatzeko erabiltzen dira.

Azpituluen itzulpena Opentrad-Matxin (Mayor et al. 2011) erabiliz egiten da, baina zientzia eta teknologiaren domeinura egokituta. Matxin euskarara itzultzeko erregeletan oinarritutako transferentzia sistema bat da. Testua gaztelaniatik euskarara itzultzen du, baina bere arkitekturak sistema berrien inplementazioa errazten du beste hizkuntza batzuetatik ere euskarara itzultzeko (Mayor et al. 2009). Opentrad-Matxin kode irekikoa da. Gaztelania-euskara sistemaren kodea eta lexikoi elebidunaren bertsio murriztu bat http://matxin.sourceforge.net helbidetik deskargatu daiteke, eta sistema martxan ikus daiteke http://matxin.elhuyar.org/ helbidean.

Matxin-en batezbesteko HTER ebaluazioaren emaitza 0,42 izan zen; honek esan nahi du 42 zuzenketa-edizio behar direla 100 hitz bakoitzeko. Sistemaren ezaugarri gakoetako bat hizkuntza-baliabideen berrerabilpena da: bere lexikoia estaldura handiko hiztegiak automatikoki prozesatuz egin da, XML formatu bat erabiliz interoperabilitatea lortzeko. Une honetan sistema estatistikoak ere lantzen ari gara, bai eta hiru hurbilpen ezberdin konbinatzen dituen sistema hibrido bat ere (España-Bonet et al. 2011).

Ahotsaren sorkuntzari dagokionez, AhoTTS-ren HMM-n oinarritutako sintesi-motorra erabiltzen dugu. Lehenik, euskararentzako modulu linguistikoak sarrerako testuaren ezaugarri linguistikoak erauzten ditu. Gero motor akustikoak hauek erabiltzen ditu aurrez entrenatutako eredu estatistikoak aukeratzeko eta parametro akustiko egokien sekuentzia bat sortzeko. Azkenik, AhoCoder-ek aipatutako parametroetatik ahots sintetikoko seinalea sortzen du. Azpituluen denbora-markak erabiltzen dira audio sintetikoa eta jatorrizko bideoa sinkronizatzeko, ahotsaren abiadura edo isiluneen luzapenarekin jokatuz.

5.2. Bilaketa motor semantiko multimedia eta eleaniztunak

5.2.1. Dokumentuaren hedapenean oinarritutako bilaketa semantikoa

Informazio-Berreskurapenean (aurrerantzean IB) erabiltzaileek formulatutako kontsultei erantzuten dieten dokumentuak aurkitu nahi izaten dira. Kontsulta bat sarreratzat hartuta, dokumentu (ustez) adierazgarriak berreskuratzen eta erakusten dituzte IB sistemek.

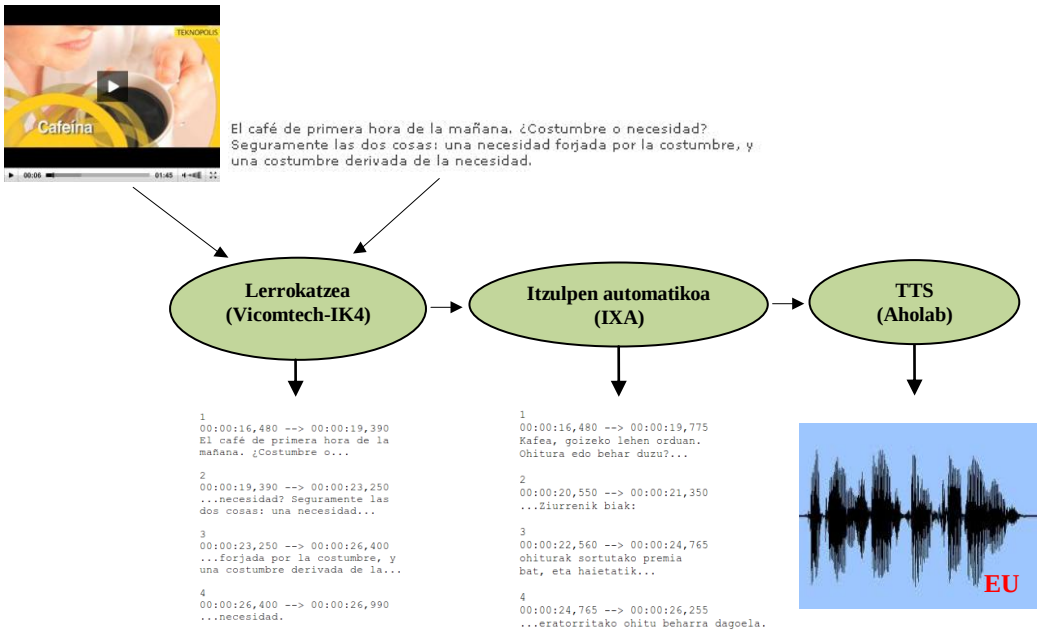
Zeregin horretan maiz gertatzen da kontsultan erabilitako terminoak eta dokumentuetakoak bat ez etortzea, dela desberdintasun morfosintaktikoengatik, dela aldaera lexikal edo semantikoengatik.

Fenomeno hau, hiztegiaren *parekatze-arazoa* deitzen dena, IB sistemek duten zailtasun handienetakoa da; izan ere, dokumentu batzuk adierazgarriak dira kontsultarekiko nahiz erabilitako terminoak oso desberdinak izan. Eta gerta daiteke, bestalde, termino berdinak erabiliagatik, berreskuratutako dokumentuak ez izatea adierazgarriak.

Arazo honen oinarrian dago hizkuntzaren aberastasuna (kontzeptu edo ideia bera adierazteko hitz edo esamolde desberdinak erabil daitezke) eta anbiguotasuna (interpretazio bat baino gehiago eduki ditzake hitz batek testuinguruaren arabera). Horiek horrela, kontsultetako eta dokumentuetako testu-kateen parekatze soilean oinarritutako IB sistemek ezintasuna dute zenbait dokumentu adierazgarri aukeratzeko eta zenbait dokumentu ez-adierazgarri baztertzeko. Hortaz, pentsa daiteke hobe dela hitzen esanahia ere tratatzea, hitzen forma soilarekin aritzea baino.

Arazo hau IB arloaren hastapenetatik jaso izan da gaiari buruzko literaturan, baina ebazteke dago oraindik, eta ez dago argi ea Hizkuntzaren Prozesamenduko (HP) teknikek hobetu dezaketen sistemen errendimendua.

[1] http://bikoizketaautomatikoa.berbatek.com:8086/demo/eu/



1. Irudia: Dokumentalen bikoizketa automatikoaren demoaren eskema.

Online dagoen beste demo batek[2] aztertzen du ea HP teknikek hobetzen duten bilaketa-makina baten eraginkortasuna (Otegi 2012).

Nahiz eta, printzipioz, sinonimia, polisemia, hiponimia eta anafora bezalako fenomenoak kontuan hartu behar diren zehaztasun handiko dokumentu-berreskuratzeetan, nekeza gertatu da fenomeno hauen eraginen azterketa sistematikoa. Ikertzaileek eredu semantiko distribuzionaletara jo dute berreskuratutako dokumentuen adierazgarritasuna hobetzeko. Metodo gehienak kontsulta-hedapenean (ingelesez *Question Expansion*, QE) oinarritu izan dira. Kontsulta-hedapeneko metodoek erabiltzaileak egindako kontsultako terminoak aztertu eta hauekin erlazionatutako terminoak gehitzen dizkie hasierako kontsultari (Manning et al., 2009).

Dokumentu-hedapena (ingelesez *Document Expansion*, DE) alternatiba arrazoizkoa da, baina ez da gehiegi aztertu duela gutxi arte. Ikertzaile batzuek metodo distribuzionalak erabili dituzte, bildumako antzeko dokumentuetatik terminoak erauzi eta horiekin jatorrizko dokumentuak osatzeko. Berbatek-en garatutako lana osagarria da, izan ere, DE aukerak ustiatzen dira, baina WordNet erabiltzen da metodo distribuzionalen ordez (Agirre et al. 2010c).

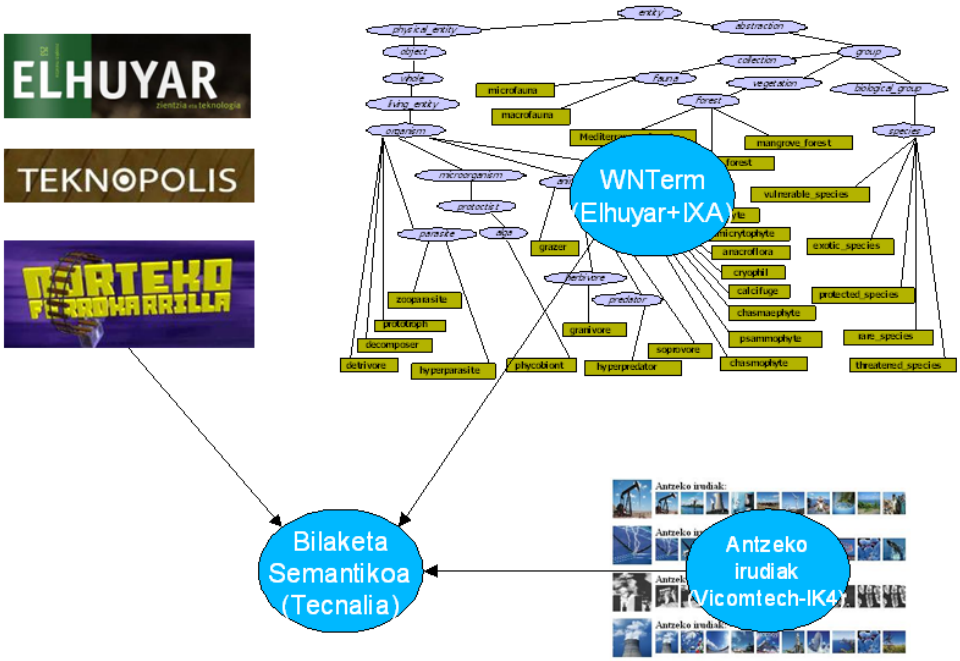
WordNet (Fellbaum 1998) ezagutza-base lexikal orokorra da. Gure ikerlanean oinarritzat hartu dugu dokumentuak hedatzeko. WordNet-en izenak, aditzak, adjektiboak eta adberbioak sinonimoen multzotan (synset) antolatzen dira, horietako multzo bakoitzak kontzeptu desberdin bat adierazten duelarik. Synset-ak erlazio lexikalen eta semantikoen bidez lotzen dira: hiperonimia, meronimia, kausalitatea, eta abar.

Beste aurreko ikerlanetan ez bezala, dokumentua osotasunean hartzen da gure lanean erlazionatutako kontzeptuak aukeratzeko. Horretarako, WordNet-eko kontzeptuen eta erlazioen grafo bidezko errepresentazioaren gaineko *ausazko bide* izeneko algoritmoa erabiltzen dugu. Algoritmo hori bera hitzen antzekotasun semantikorako eta hitzen adiera-desanbiguaziorako (Agirre eta Soroa, 2009) erabilia izan da emaitza arrakastatsuekin.

Metodo honek kontzeptu esanguratsuenak lortzen ditu, kontzeptu horiek hitz-forma esplizituekin adierazi ez badira ere dokumentuan. Behin dokumentu-hedapenerako hitzen zerrenda edukita, bi indize sortzen dira: bata, dokumentuko jatorrizko hitzekin; bestea, hedatutako terminoekin. Horrela, gai gara jatorrizko hitzak baizik ez erabiltzeko bilaketetan, edota, nahi izanez gero, hedapen-hitzak ere sartzeko.

Dokumentuen berreskurapen-sistema MG4J (Boldi and Vigna 2005) bilatzailea erabiliz inplementatu da. Bilatzaile honek artearen egoerako emaitzak lortzen ditu, eta dokumentu-bildumaren gainean indize bat baino gehiago konbinatzeko aukera ematen du. BM25 ranking-funtzioa erabili da, oso esanguratsua, sendoa eta erabilia baita (Robertson and Zaragoza 2009).

[2] http://ixa2.si.ehu.es/BerbatekDemo/bilatu



2. Irudia: Bilatzaile semantikoaren demoaren eskema.

5.2.2. Ontologia espezializatu batean oinarritutako zientzia eta teknologiako bilatzaile semantiko eleaniztuna, antzeko irudien bilaketarekin

Edukien arloan hizkuntza-teknologiek egin dezaketen ekarpenaren frogagarri, zientzia eta teknologiako bilatzaile semantiko multimedia bat egin dugu. Bilatzaile horrek Elhuyarrek eta IXA Taldeak eraikitako zientzia eta teknologiako WNTerm ontologia espezializatua (Pociello et al. 2008) du oinarri (zientzia eta teknologiaren alorreko kontzeptuak semantikoki erlazionatuta ageri diren sare bat, azpiklaseekin, sinonimoekin eta abar), eta Elhuyarren edukiaren gainean (Elhuyar aldizkariko irudi eta testuak, Teknopolis telebista-programako bideoak eta Norteko Ferrokarrilla irratsaioko audioak) funtzionatzen du. Tecnaliak garatutako teknologiaren bidez, termino bat bilatzen denean, ontologiaren bidez termino horren sinonimoak, azpiklaseak edo superklaseak dituzten edukiak ere bila daitezke. Gainera, emaitza irudi bat denean, antzeko irudiak ere ematen ditu, Vicomtech-IK4ren teknologia erabiliz. Demo hau ere online dago[3].

Edukiak deskribatzeko, Dublin Core estandarra erabiltzen da, interoperabilitatea ziurtatzeko. Formatu hori erabiliz, Elhuyarren zientzia eta teknologiako eduki guztia etiketatu da. Anotazioa editore baten laguntzarekin egiten da.

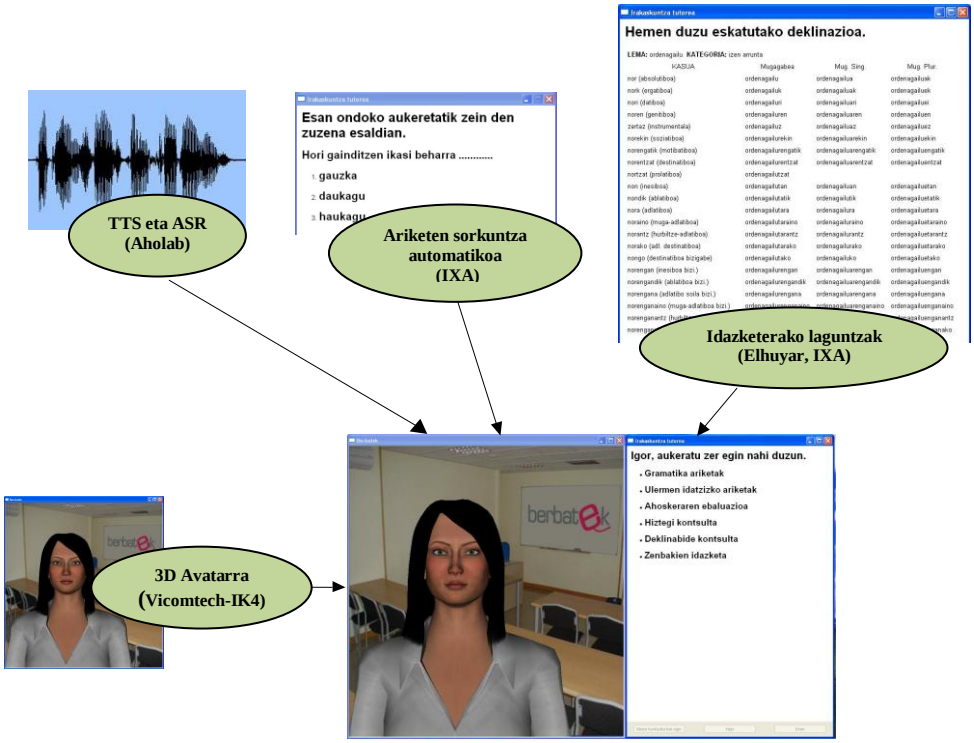
Bilatzaileari emandako bilaketa-terminoetako bat ontologiako kontzeptu bat dela detektatzen den kasuetan, erlazionatutako kontzeptuen bilaketak proposatzen dira (sinonimoak, hiponimoak, hiperonimoak eta abar). Bilatzailearen emaitzetakoren bat irudi bat den kasuetan, erabiltzaileari antzeko 10 irudiren zerrenda bat ere erakusten zaio.

5.3. Hizkuntza-irakaskuntzarako tutore pertsonala

Irakaskuntzaren alorrerako, hizkuntzen irakaskuntzako tutore pertsonal baten demoa egin dugu. Tutore hori emozioak adieraz ditzakeen 3Dko pertsonaia bat da, Vicomtech-IK4k garatutakoa, euskaraz mintzatzen dena eta euskaraz ahoz esaten zaiona ulertzen duena, Aholab-en teknologiari esker. Eta tutoreak hainbat gauzatan lagundu gaitzake: IXAren teknologiaren bidez, automatikoki sortutako gramatika-ariketak (aditzak, deklinabidea...) edo ulermen-ariketak (testu batean hutsuneak betetzea, hainbat aukera emanda) egin ditzakegu; ahoskera ebaluatzen digu, Aholab-en teknologiari esker; edo idazketarako laguntzak ematen ditu (aditzen jokabidea, zenbakien idazketa, hiztegi-kontsultak...), IXA eta Elhuyarren teknologiaren bidez. Demo honetan parte hartzen duten teknologien eskema 3. irudian ikus daiteke.

Avatarraren moduluak demoaren aurpegia den 3D pertsonaia erakutsi eta animatzeko funtzionalitateak ditu. Ezpainen animazioa TTS moduluak sintetizatutako audioarekin sinkronizatuta dago, eta beharrezkoa denean emozioak erakuts ditzake. Horrez gain,

[3] http://bilatzailesemantikoa.berbatek.com



3. Irudia: Hizkuntza-irakaskuntzarako tutore pertsonalaren demoaren eskema.

aurredefinitutako erregela batzuen arabera begiak kliskatu eta burua mugitzen du, bizirik dagoen ilusioa areagotzeko. C++ programazio-lengoaian idatzita dago, liburutegi grafiko gisa OpenSceneGraph (http://www.openscenegraph.org) duelarik.

Ariketen sorkuntza automatikorako ArikIturri (Aldabe 2010) erabiltzen da. Galdera mota ezberdinak sortzeko sistema da hau. Sarrera gisa morfologikoki eta sintaktikoki etiketatutako esaldiak hartzen ditu, XML bidez adierazita, eta sortutako galderetara bihurtzen ditu, hauek ere XMLn adierazita.

Aurreko sistemekin konparatuta (Kraift et al. 2004; Schwartz et al. 2004) gureak baditu berezitasun batzuk. Gureak bi modulu hauek bereizten ditu: *Erantzun-Fokuen Identifikatzailea*, eta *Txarto Eratutako Galderen Baztertzailea*. Sumita et al. (2005) egileek ere erabili izan dute galderak baztertzeko modulu bat, webeko oinarria ere zuena. Gure *esaldi-bilatzaileak* esaldi hautagaiak aukeratzen ditu etiketatuta dagoen iturburu-corpusetik, parametroen espezifikazioari jarraituz. Lehenengo urrats batean aukeratzen ditu landu nahi den fenomeno linguistikoa bere barnean hartzen duten esaldiak. Gero *hautagai-aukeratzaileak* aztertzen ditu hautagai horien portzentajeak, emandako parametroak betetzen dituzten esaldien kopuruaren arabera zorizko aukeraketa bat egiteko. Behin esaldiak aukeratu diren, ezagututako sintagmen (edo chunk-ak, orokorrago eta zehatzago esanda) informazio morfosintaktikoa erabiliz *Erantzun-Fokuen Identifikatzaileak* zenbait sintagma markatuko ditu erantzun-foku gisa. Ondoren *item-sortzaileak* galderak sortuko ditu zehaztutako ariketa-mota baten arabera. Modulu honetan *distraktoreak sortzeko azpimodulu* bat ere gehitu dugu. Urrats guzti horiek bete eta gero, honezkero sistemak sortu ditu galdera-instantziak. Hala ere, prozesu osoa automatikoa denez, eta sortutako galdera batzuk txarto eratuta egon daitezkeenez, sistemaren arkitekturan beste modulu bat gehitu dugu txarto eratutako galderak baztertzeko.

Idatzeko laguntzei dagokienez, sistemak hiru laguntza mota eskaintzen ditu.: hitzen inflexioa, zenbakien idazketa eta hiztegi-kontsulta.

Inflexioekin laguntzeko moduluak hitz bat eskatzen du, gero kasua (absolutiboa, datiboa eta abar) eta gero mugatasuna (singularra, plurala edo mugagabea). Azken bi hauetarako denak ere eska daitezke. Sistemak Elhuyarrek garatutako web zerbitzu bati deitzen dio, honek inflexioak sortzen ditu bi mailatako morfologian oinarritutako transduktore bat erabiliz eta taula bat itzultzen du hitzaren inflexioekin.

Zenbakiak idazteko moduluak hamarreko luzera arterainoko zifra bakarreko zenbakien segida bat eskatzen du (adibidez "*hiru zazpi lau lau bost bat*") eta erabiltzaileari esaten dio nola idatzi eta ahoskatu behar den horiekin eratzen den zenbakia (adibidean "*hirurehun eta hirurogeita hamalau mila, laurehun eta berrogeita hamaika*"), Elhuyarrek garatutako sistema bat erabiliz.

Azkenik, hiztegia kontsultatzeko moduluak euskarazko hitz bat eskatzen du eta Elhuyarren hainbat online hiztegitan begiratzen du (euskara-gaztelania, euskara-frantsesa, euskara-ingelesa eta sinonimoak), emaitza guztiak erakutsiz.

Ahots-teknologiak ere asko erabiltzen dira demo honetan. AhoSR, euskararentzako ahots-ezagutzako sistema, erabiltzailearen aukerak eta erantzunak ezagutzeko erabiltzen da, eta AhoTTS, euskarazko testu-ahots bihurtzailea, avatarraren audio-erantzunak sortzeko. AhoSr ere erabiltzen da ahoskeraren ebaluazioaren moduluan.

Demo honetan HMMetan oinarritutako AhoTTS bertsioa erabiltzen da. HTSk analisi linguistikoa ez duenez aplikatzen, AhoTTS-ren lehen moduluaren irteera erabili da. Irteera honen formatua bihurtu egin da eta informazio linguistikoa eta fonetikoa etiketa egokietan jaso da. Testuinguru etiketetan kodifikatu diren ezaugarriak (Erro et al. 2010) artikuluan ikusi daitezke xehetasun gehiagorekin. Espektroaren eta kitzikaduraren irudikapen parametrikoa lortzeko, HNM (Harmonics plus Noise Model) ereduan oinarritutako vocoderra erabiltzen da, AhoCoder deitzen dena (Erro et al. 2011b). Vocoder honek ahotsaren berreraiketa ere egiten du. AhoSyn datu-baseak (Sainz et al., 2012) dauzkan ezaugarri berberak daukan ahots datu-basea erabili da ahots sintetikoa lortzeko. Demoan emakume eta gizon ahotsak erabili dira. Emakume ahotsa prozedura estandarra jarraituz lortu da eta gizonarena ahots transformazioa aplikatuz (Erro et al., 2013). AhoTTS bertsio hau elebiduna da, gazteleraz eta euskaraz egin dezake eta online dago eskuragai: http://sourceforge.net/projects/ahotts/.

Ahoskera ebaluatzen duten sistemetan, CAPT (Computer-Assisted Pronunciation Training) helburuarekin diseinatutako datu-baseak erabiltzen dira oro har. Tamalez ez dago euskarako CAPT helburuarekin grabatu den datu-baserik. Ahots ezagupena (ASR, Automatic Speech Recognition) euskaraz egiteko datu-base batzuk egon arren, publikoki lortu daitekeen bakarra (Hernáez, et al. 2003) sare telefoniko finkoaren bitartez grabatu zen eta horregatik ez da egokia CAPT egiteko, sistema hauetan ahotsa mikrofono baten bidez grabatzen baita normalean. Horregatik euskararako CAPT sistemak beste datu eskuragarriak erabiliz eraiki behar dira.

Erabili dugun datu-basea euskarazko ahots ezagutza egiteko diseinatu zen. Speecon datu-basearen (Siemund et al., 2000) antzekoa da eta euskaldun zaharren zein berrien grabaketak dauzka. Euskaldun zaharrek euskara batua eta dagokien euskalkiarena erabiltzen dute. Euskal Herriko leku ezberdinetako 230 hizlari grabatu ziren eta bakoitzaren euskara menderatze maila ere jaso zen, beraz, datu-basea euskara mailaren arabera banatu daiteke. Datu-basean 149 euskaldun zahar daude. Euskaldun berrien artean bi maila ere bereizi daitezke: 56 hizlarik maila ona zeukaten eta 25ek maila baxua. Euskalkia eta menperatze maila ezberdinen ondorioz, fonema batzuen ahoskeran irregulartasunak agertzen dira baina ez daude datu-baseko transkripzio fonetikoan jasota. Hala ere, irregulartasun hauek hizlariari buruzko informazioa kontutan izanda ondorioztatu zitezkeen kasu batzuetan. Datu-basean audio fitxategiak eta dagozkion transkripzio ortografikoa daude. AhoT2P transkriptore fonetikoa erabili da transkripzio fonetikoak lortzeko.

Esan dugun bezala, ASR helburuarekin grabatutako datu-basea erabili da ahoskera ebaluatzeko modulua garatzeko (Odriozola et al. 2012). Erabilitako metodoak ahoskeraren egokitasuna (GOP, Goodness Of Pronunciation) automatikoki neurtzen du. Fonema bakoitzeko bi GOP banaketa lortu dira: bat ahoskera zuzenari dagokio eta bestea ahoskera okerrari. Ahoskera okerraren banaketa artifizialki kalkulatu zen, hiztegian erroreak sartuz modu kontrolatuan. Aldatzen zen fonema bakoitza multzo berberaren beste fonema batekin ordezkatu zen. Fonema multzoak klasifikazio zuhaitzak aplikatuz lortu ziren. Bi GOP banaketa lortuz gero, fonema bakoitzerako erabaki atalasea EER (Equal Error Rate) erabiliz finkatu da. Egin diren esperimentuekin konprobatu da metodo hau erabilgarria dela, CAPT helburu zehatzarekin diseinatzen direnek emaitza hobeak lortu arren. Bai ASR moduluan bai ahoskeraren zuzentasuna ebaluatzeko moduluan datu-base berbera erabili da.

6. Ondorioak

IKTetan eta, bereziki, hizkuntza-, ahots- eta multimedia-teknologietan presente egotea, gure ustez, erabat beharrezkoa du gero eta mugikorrago, digitalago eta elkarkonektatuago bihurtzen ari den mundu honetan bizirik iraun nahi duen edozein hizkuntzak.

Alegria et al. (2011) artikuluan esaten denez, euskararentzako hizkuntza-baliabide, tresna eta aplikazioen diseinua eta garapena modu inkrementalean egin behar da eta erakunde ezberdinen artean koordinatuta eta paraleloan, errendimendu handiena lortzeko. BerbaTek proiektuko esperientziak frogatu digu aipatutako arloetako agenteen elkarlana dela jarraitu beharreko bidea. Oinarrizko ikerketa egin eta gizartearen eskura oinarrizko baliabide eta tresna ugari jartzeaz gain, teknologia ezberdinen integrazioak ahalbidetu du hizkuntzen industriarako, hau da, itzulpenen, edukien kudeaketaren eta irakaskuntzaren sektoreetarako, aplikazio aurreratuen prototipoak garatzea.

META-NET erakundeak egindako Liburu Zurien Bildumarako egindako hizkuntzarteko konparazioak[4] erakutsi zuen hizkuntza- eta ahots-teknologiei dagokienean euskara gaur egun egoera hobean dagoela Europar Batasuneko hizkuntza ofizial batzuk baino (kroaziera, islandiera, gaelikoa, letoniera edo lituaniera), eta euskara 5 maila posibletatik 4.ean dagoela teknologia horien egoerari dagokionean.

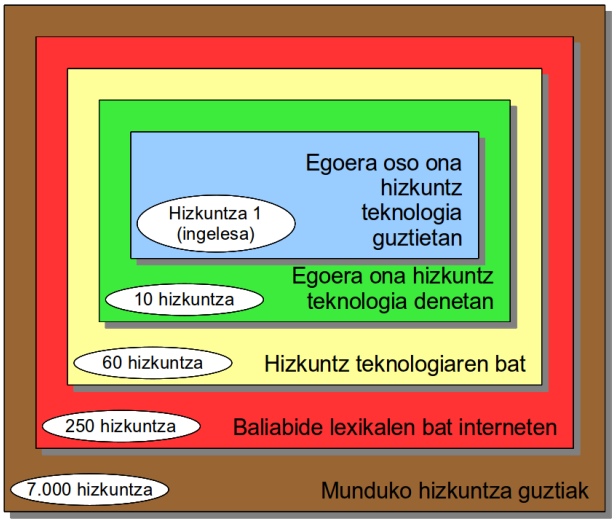
[4] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

Euskarari buruzko liburu zuriak (Hernáez et al., 2012) ondorioztatu zuen tresnak badaudela ahots sintesirako, ahots ezagutzarako eta ortografia- eta gramatika-zuzenketarako, eta itzulpen automatikorako tresnak ere badirela, bereziki gaztelania eta euskara artean.

4. irudiak grafikoki erakusten ditu Alegria et al.-ek (2011) proposatutako munduko hizkuntzen sailkapen posibleak hizkuntza-teknologien garapenari dagokionez. 2013an euskara erdiko posizioan dago, hizkuntz teknologia batzuk badituzten 60 hizkuntzaren taldean. 35. hizkuntza da Wikipediako artikulu kopuruari dagokionez[5]. Euskararako 6 produktu daude ELRAko katalogoan[6], 15 produktu ACLko wiki-an[7] eta 40 online hiztegi eta corpus baino gehiago daude Euskalbar tresnan[8].

Euskal hiztunen kopuruaren eta euskararentzako dauden baliabide eta tresna teknologikoen kopurua harrigarriki altua da. Hau guztia ikuspegi handiko elkarte produktibo batzuek hainbat proiektutan jarraian izandako esfortzu koordinatuei esker da. Euskararentzako teknologietan aritzen diren bost jokalari garrantzitsuenek BerbaTek-en izandako kolaborazioa pauso bat haratago izan da norabide horretan, baliabide eta tresna berri gehiago garatzeaz gain arlo batzuetan aplikatzeko tresna aurreratuen prototipoak egitea ahalbidetu duena.

Etorkizunean, gure arteko elkarlanarekin jarraitu nahi dugu eta aurrera jarraitu bai oinarrizko ikerketarekin bai aplikazio eta prototipoaren garapenarekin, hizkuntzan industriarako bezala beste sektore batzuetarako ere. Baina prototipoez harago joateko asmoa ere badugu eta hurrengo pauso¡u logikoari heldu, hau da, erabiltzaileentzako aplikazio errealak garatu eta merkaturatzea, hizkuntzen industriaren sektoreko enpresekin kolaborazioan. BerbaTek kontsortzioko kideak gogoz eta prest gaude erronka honi aurre egiteko.



4. Irudia: Munduko hizkuntzen sailkapen posibleak hizkuntza-teknologien garapenari dagokionez.

7. Eskerronak

Ikerketa lan honek Eusko Jaurlaritzako (BerbaTek proiektua, IE09-262) eta Espainiako Gobernuko (OpenMT2, TIN2009-14675-C03- 01; Know2, TIN2009-14715-C04-01; HibridoSint;TIN2010-20218) laguntza jaso du.

8. Erreferentziak

Agirre, E.; Soroa, A.; Alfonseca, E.; Hall, K.; Kravalova, J.; Pasca, M. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 19-27, ISBN: 978-1-932432-41-1. Boulder, USA.

Agirre, E.; Lopez de Lacalle, O.; Soroa, A. (2009). Knowledge-based WSD and specific domains: performing over supervised WSD. In Proceedigns of the International Joint Conferences on Artificial Intelligence (IJCAI),. pp. 1501-1506. ISBN 978-1-57735-429-1. Pasadena, USA.

Agirre, E.; Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). pp. 33-41. ISBN: ISBN 978-1-932432-16-9P. Athens, Greece.

Agirre, E.; Ansa, O.; Arregi, X.; Lopez de Lacalle, M.; Otegi, A.; Saralegi, X.; Zaragoza, H. (2009). Elhuyar-IXA: semantic relatedness and crosslingual passage retrieval. In Multilingual Information Access Evaluation I - Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece. Lecture Notes in Computer Science, Vol. 6241, pp. 273–280. ISSN:0302-9743 ISBN: 978-3-642-15753-0.

Agirre, E.; Soroa, A.; Stevenson, M. (2010). Graph-based Word Sense Disambiguation of Biomedical Documents. Bioinformatics 26(22): 2889-2896. Oxford University Press.

Agirre, E.; Cuadros, M.; Rigau, G.; Soroa, A. (2010). Exploring Knowledge Bases for Similarity. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA), pp. 373—377, ISBN: 2-9517408-6-7, Valletta, Malta.

Agirre, E.; Arregi, X.; Otegi, A. (2010). Document Expansion Based on WordNet for Robust IR. In Proceedings of the 23rd International Conference on

[6] http://catalog.elra.info/

[7] http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque

[8] http://euskalbar.eu/

Computational Linguistics (Coling). pp 9-17. Beijing, China.

Agirre, E.; Bengoetxea, K.; Gojenola, K.; Nivre, J. (2011). Improving Dependency Parsing with Semantic Classes. In Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics, ACL-HLT 2011 Short Paper, Portland, Oregon.

Aldabe, I.; Maritxalar, M. (2010). Automatic Distractor Generation for Domain Specific Texts. In Proceedings of the 7th International Conference on NLP, IceTAL 2010, pp. 27-38, ISBN-10 3-642-14769-0. Reykjavik, Iceland, August 2010.

Aldezabal, I.; Aranzabe, M. J.; Diaz de Ilarraza, A.; Estarrona, A.; Uria, L. (2010). EusPropBank: Integrating Semantic Information in the Basque Dependency Treebank. Lecture Notes in Computer Science 6008: 60-73.

Alegria, I.; Diaz de Ilarraza, A.; Labaka, G.; Lersundi, M.; Mayor, A.; Sarasola, K. (2007). Transfer-based MT from Spanish into Basque: reusability, standardization and open source. Lecture Notes in Computer Science 4394: 374-384.

Alegria, I.; Casillas, A.; Diaz de Ilarraza, A.; Igartua, J.; Labaka, G.; Lersundi, M.; Mayor, A.; Sarasola, K. (2008). Spanish-to-Basque MultiEngine Machine Translation for a Restricted Domain. In Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA-2008). pp.57-69. Hawai, USA.

Alegria, I.; Artola, X.; Diaz de Ilarraza, A.; Sarasola, K. (2011). Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque. In Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. pp: 42-46, November 24-27, ISBN: 978-83-932640-1-8. Poznań, Poland.

Ansa, O.; Arregi, X.; Otegi, A.; Soraluze, A. (2008). Ihardetsi question answering system at QA@CLEF 2008. In Working Notes of the Cross-Lingual Evaluation Forum, ISBN 2-912335-43-4, ISSN 1818-8044, Aarhus, Denmark.

Arrieta, K.; Diaz de Ilarraza, A.; Hernáez, I.; Iturraspe, U.; Leturia, I.; Navas, E.; Sarasola, K. (2008). AnHitz, development and integration of language, speech and visual technologies for Basque. In Proceedings of the Second International Symposium on Universal Communication OSAKA, pp. 338-344, 530-0005, JAPAN. Published by IEEE Computer Society. ISBN: 978-0-7695-3433-6 http://doi.ieeecomputersociety.org/10.1109/ISUC.2008.43.

Bengoetxea, K.; Gojenola K. (2010). Application of Different Techniques to Dependency Parsing of Basque. In Proceedings of the First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010), NAACL Workshop, Los Angeles USA.

Boldi, P.; Vigna, S. (2005). MG4J at TREC 2005. In the annual meeting for the Text REtrieval Conference (TREC), Gaithersburg, USA.

Diaz de Ilarraza, A.; Gurrutxaga, A.; Sarasola, K.; Hernáez, I.; Lopez de Gereñu, N. (2003). HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities. In Proceedings of the Workshop on NLP of Minority Languages and Small Languages. TALN 2003. Nantes.

Erro, D.;, Sainz, I.; Luengo, I.; Odriozola, I., Sánchez, J.; Saratxaga, I.; Navas, E.; Hernáez, I. (2010). HMM-based Speech Synthesis in Basque Language using HTS. In Proceedings of the Jornadas en Tecnología del Habla (JTH), Vigo, Spain.

Erro, D.; Sainz, I.; Navas, E.; Hernáez, I. (2011). HNM-Based MFCC+F0 Extractor Applied to Statistical Speech Synthesis. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czec Republic.

Erro, D.; Sainz, I.; Navas, E.; Hernáez, I. (2011). Improved HNM-based Vocoder for Statistical Synthesizers. In Proceedings of the Interspeech Conference, Florence, Italy.

Erro, D.; Navas, E.; Hernáez, I. (2013). Parametric Voice Conversion Based on Bilinear Frequency Warping plus Amplitude Scaling. IEEE Transactions on Audio, Speech and Language Processing, 21(3), 556–566.

España-Bonet, C.; Labaka, G.; Diaz de Ilarraza, A.; Màrquez, L.; Sarasola, K. (2011). Hybrid Machine Translation Guided by a Rule–Based System. In Proceedings of the Machine Translation Summit (MT Summit), Xiamen, China.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database and Some of its Applications. Cambridge: MIT Press.

Gurrutxaga, A.; Alegria, I. (2011). Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. In Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011). ACL/HLT conference, Portland, Oregon.

Haveliwala, T. H. (2002). Topic-sensitive PageRank. In Proceedings of the World Wide Web conference (WWW), Honolulu, USA.

Hernáez, I.; Luengo, I.; Navas, E.; Zubizarreta, M.; Gaminde, I.; Sánchez, J. (2003). The Basque speech_dat (II) database: a description and first test recognition results. In Proceedings of the Eurospeech conference, Geneva, Switzerland.

Hernáez, I.; Navas, E.; Odriozola, I.; Sarasola, K.; Diaz de Ilarraza, A.; Leturia, I.; Diaz de Lezana, A.; Oihartzabal, B.; Salaberria, J. (2012). The Basque language in the digital age/Euskara aro digitalean. METANET White Paper Series. Georg Rehm, Hans Uszkoreit (editors). Springer. http://www.meta-net.eu/whitepapers/e-book/basque.pdf.

Kraift, O.; Antoniadis, G.; Echinard, S.; Loiseau, M.; Lebarbé, T.; Ponton, C. (2004). NLP Tools for CALL: the Simpler, the Better. In Proceedings of the

Workshop on NLP and Speech Technologies in Advanced Language Learning Systems, Venice, Italy.
Labaka, G. (2010). EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD diss., UPV/EHU-University of the Basque Country.
Leturia, I.; San Vicente, I.; Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In Proceedings of the Workshop on Web as Corpus (WAC), Donostia, Spain.
Leturia, I. (2012). Evaluating different methods for automatically collecting large general corpora for Basque from the web. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India.
Manning, C. D., Raghavan, P.; Schütze, H. (2009). An introduction to information retrieval. Cambridge: Cambridge University Press.
Mayor, A.; Tyers, F. (2009). Matxin: Moving towards language independence. In Proceedings of the Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT), Alacant, Spain.
Mayor, A.; Alegria, I.; Diaz de Ilarraza, A.; Labaka, G.; Lersundi, M.; Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. Machine Translation Journal 25(1): 53-82.
Odriozola, I., Navas, E.; Hernáez, I.; Sainz, I.; Saratxaga, I.; Sánchez, J.; Erro, D. (2012). Using an ASR database to design a pronunciation evaluation system in Basque. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey.
Otegi, A. (2012). Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak. PhD diss., UPV/EHU-University of the Basque Country.
Pociello, E.; Gurrutxaga, A.; Agirre, E.; Aldezabal, I.; Rigau, G. (2008). WNTERM: Combining the Basque WordNet and a Terminological Dictionary. In Proceedings of the Language Resources and Evaluation Conference (LREC), Marrakesh, Morocco.
Pociello, E.; Agirre, E.; Aldezabal, I. (2011). Methodology and construction of the Basque Wordnet. Language Resources and Evaluation 45(2): 121-142.
Robertson, S.; Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval 3(4): 333-389.
Sainz, I.; Erro, D.; Navas, E.; Hernáez, I. (2011). A Hybrid TTS Approach for Prosody and Acoustic Modules. In Proceedings of the Interspeech Conference, Florence, Italy.
Sainz, I.; Erro, D.; Navas, E.; Hernáez, I.; Sanchez, J.; Saratxaga, I.; Odriozola, I. (2012). Versatile Speech Databases for High Quality Synthesis for Basque. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey.
San Vicente, I.; Manterola, I. (2012). PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey.
Saralegi, X.; Lopez de Lacalle, M. (2010). Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In Proceedings of the Language Resources and Evaluation Conference (LREC), Valletta, Malta.
Saralegi, X.; San Vicente, I.; López de Lacalle, M. (2008). Mining Term Translations from Domain Restricted Comparable Corpora. Procesamiento del Lenguaje Natural 41: 273-280.
Saralegi, X. Lopez de Lacalle, M. (2009). Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence-Based Selection. In Proceedings of the Workshop on Text-Based Information Retrieval (TIR), Linz, Austria.
Saralegi, X.; Manterola, I.; San Vicente, I. (2012). Building a Basque-Chinese Dictionary by using English as a Pivot. In Proceedings of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey.
Schwartz, L.; Aikawa, T.; Pahud, M. (2004). Dynamic Language Learning Tools. In Proceedings of the Workshop on NLP and Speech Technologies in Advanced Language Learning Systems, Venice, Italy.
Siemund, R.; Höge, H.; Kunzmann, S.; Marasek, K. (2000). SPEECON-speech data for consumer devices. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Athens, Greece.
Sumita, E.; Sugaya, F.; Yamamoto, S. (2005). "Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the Blank Questions". In Proceedings of the Workshop on Building Educational Applications Using NLP, Ann Arbor, USA.
Williams, B.; Sarasola, K.; Ó´Cróinin, D.; Nadeu, C.; Petek, B. (2001). Speech and Language Technology for Minority Languages. In Proceedings of the Eurospeech conference, Aalborg, Denmark.