Igor Leturia Azkarate
Informatikaria eta ikertzailea

berbatek euskarazko hizkuntza-teknologiak martxan

Azken hiru urteotan, euskararentzako hizkuntza-teknologiak ikertzen eta garatzen aritu gara Euskal Herriko hainbat erakunde BerbaTek proiektuan. Eta, proiektuaren helburuetako bat izanik ikerketa erabilera praktikora bideratzea, hiru demo ere sortu ditugu: zientzia eta teknologiako bilatzaile semantiko multimedia bat, dokumentalen bikoizketa automatikorako demo bat, eta hizkuntzen irakaskuntzarako tutore pertsonal bat.

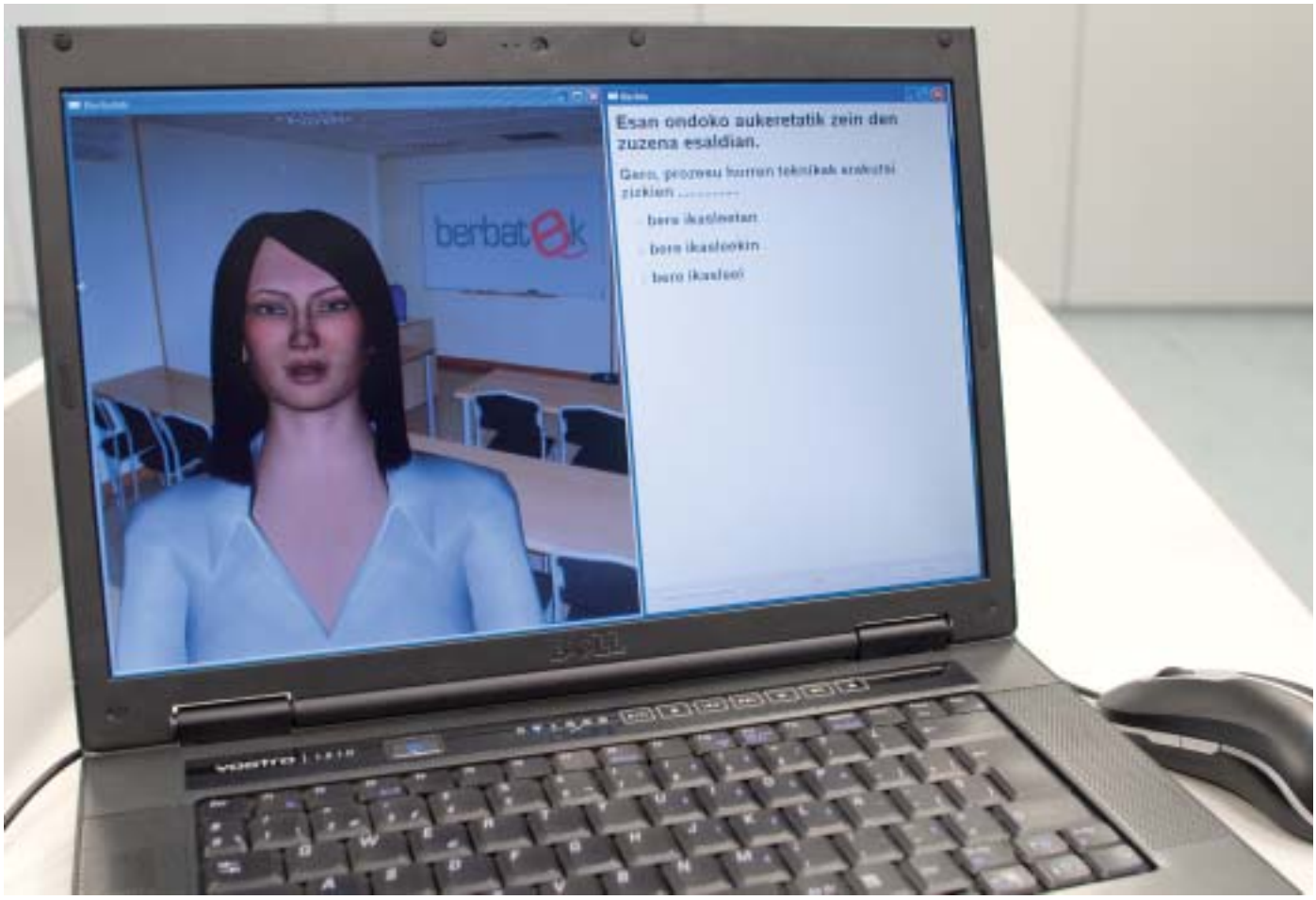
Azken ia hiru urte hauetan "Mundu digitala" atal honen jarraipena egin baduzu, konturatuta egongo zara hizkuntza-teknologiak gero eta garrantzitsuagoak izango direla gailu mugikor eta beti konektatutakoen munduan. Hitz egin dizuegu mundu berri horretan presentzia nabarmena duten (eta gero eta handiagoa izango duten) hainbat teknologiaz: web semantikoaz eta teknologia semantikoez, itzulpen automatikoaz eta corpusez, galderak erantzuteko sistemez, elkarrizketa-agenteez, bilatzaile adimendunez... Teknologia horiek oraindik bidea dute egiteko, baina zenbait kasutan erabilgarri izateko adina aurreratuta badaude eta gailu askok eta zerbitzu askok badituzte integratuta, hemen kontatu dizuegunez.

Baina, oro har, hizkuntza hedatuenentzat soilik daude martxan horrelakoak (ingelesez besterik ez sarritan); konpainia handiek ez dute interesik euskara horietan txertatzeko. Eta izango balute ere, ez daude prest teknologia horiek euskarara egokitzeko kostua beren gain hartzeko. Euskarara egokitze hori ez baita lan hutsala; batzuetan, oinarrizko ikerketa egin beharra dago, oinarrizko baliabideak garatu...

Bada horretan aritu gara Elhuyar Fundazioa, EHUko IXA eta Aholab ikerketa-taldeak, eta Vicomtech-IK4 eta Tecnalia teknologia-zentroak, BerbaTek proiektuaren barruan, 2009. eta 2011. urteen artean, euskararentzako (nagusiki) hizkuntza-, ahots- eta multimedia-teknologiak ikertzen. Eusko Jaurlaritzako Industria eta Kultura sailek finantzatu dute BerbaTek proiektuaren aurrekontuaren zati bat, Etortek programaren bitartez.

> *Hizkuntza-teknologiak gero eta garrantzitsuagoak izango dira gailu mugikor eta beti konektatutakoen munduan.*

Ez da lehen aldia 5 erakunde horiek elkarlanean hizkuntza-teknologien ikerketan aritzen garena. Aurretik 2002-2004 aldian Hizking XXI proiektuan aritu ginen, eta 2006-2008 aldian, AnHitz proiektuan. Azken horren amaieran, zientzia-aditu birtual baten demo bat eraiki genuen, AnHitz izenekoa hori ere: ahozko interak-



BerbaTek proiektuan sortutako hizkuntzen irakaskuntzarako demoa. ARG.: DANEL SOLABARRIETA.

zioa zuen 3D avatar bat zen, zientziari buruzko galderei erantzuteko eta bilaketa eleaniztunak egiteko gai zena.

BerbaTek proiektuan, oinarrizko ikerketa handia egin dugu: oinarrizko baliabide eta tresna asko garatu edo hobetu ditugu (testu- edo ahots-corpusak, lexikoiak, hiztegiak, ontologiak, gramatika konputazionalak, analizatzaile morfosintaktikoak, ahots-ezagutza, ahots-sintesia, elkarrizketa-sistemak...), eta hainbat teknologia landu ditugu (itzulpen automatikoa, informazio-bilaketa, informazio-erauzketa, idazketan laguntzeko sistemak, ezagutzaren kudeaketa, galderei erantzuteko sistemak, tutore pertsonalak, e-learning sistemak, ahoskera zuzentzeko sistemak, ariketen eta adibideen eraikitze automatikoa...). Eta bertan landutako teknologiak zenbait proiektu eta zerbitzutan erabili dira.

HIZKUNTZEN INDUSTRIAREN ZERBITZURA

BerbaTek proiektua ikerketa-proiektua bada ere, ikerketa horren erabilera praktikoa izan da hasieratik guretzat helburu nagusietako bat. Eta praktikotasun hori hizkuntzen industriaren alorrean eman nahi izan diogu.

Hizkuntzen industriatzat ulertzen da hiru azpisektore hauek osatzen dutena: itzulpena (itzulpenak, lokalizazioak, interpretazioa, bikoizketa...), edukiak (argitaletxeak, komunikabideak...) eta irakaskuntza (hizkuntzen irakaskuntza, irakaskuntza arautua...). Euskal Herrian, duela gutxi hasi dira ematen lehen pausoak hizkuntzen industriaren sektorea egituratzeko: 2010ean, Langune hizkuntzen industriaren alorreko Euskal Herriko enpresen elkartea sortu zen; 30 bazkide baino gehiago ditu. BerbaTek-eko kideek aktiboki hartu dute parte han sorreratik, eta industriaren eta elkartearen euskarri teknologiko izateko bokazioa du BerbaTek proiektuak.

BerbaTek-en landu diren teknologia askok zuzeneko aplikazioa dute hizkuntzen industriako hiru sektoretako batean, eta beste tresna, baliabide eta teknologia batzuk horietako edozeine-

BerbaTek-en landu diren teknologia askok zuzeneko aplikazioa dute hizkuntzen industriako hiru sektoretako batean.

tan aplikatzekoak dira edo beste teknologiak garatzeko oinarriak dira.

Eskemak adierazten du grafikoki hizkuntzen industria eta haren alorrak, eta BerbaTek-ek zer ekarpen egin diezaiokeen bakoitzari eta oro har.

Tutore hori emozioak adieraz ditzakeen 3Dko pertsonaia bat da, euskaraz mintzatzen da eta euskaraz ahoz esaten zaiona ulertzen du.

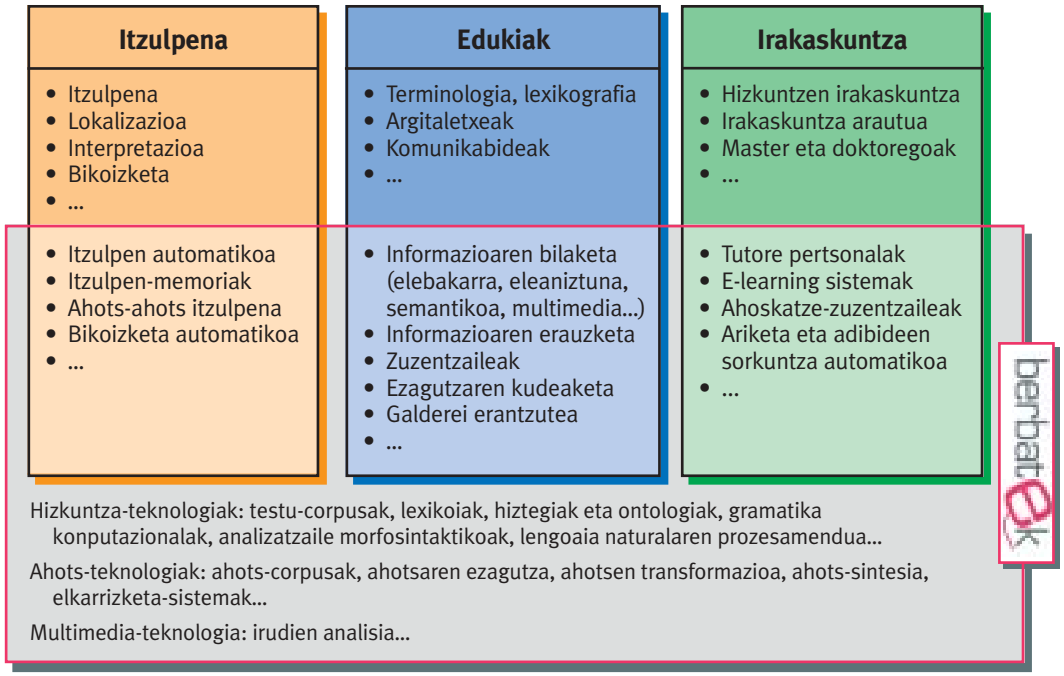
DEMOAK

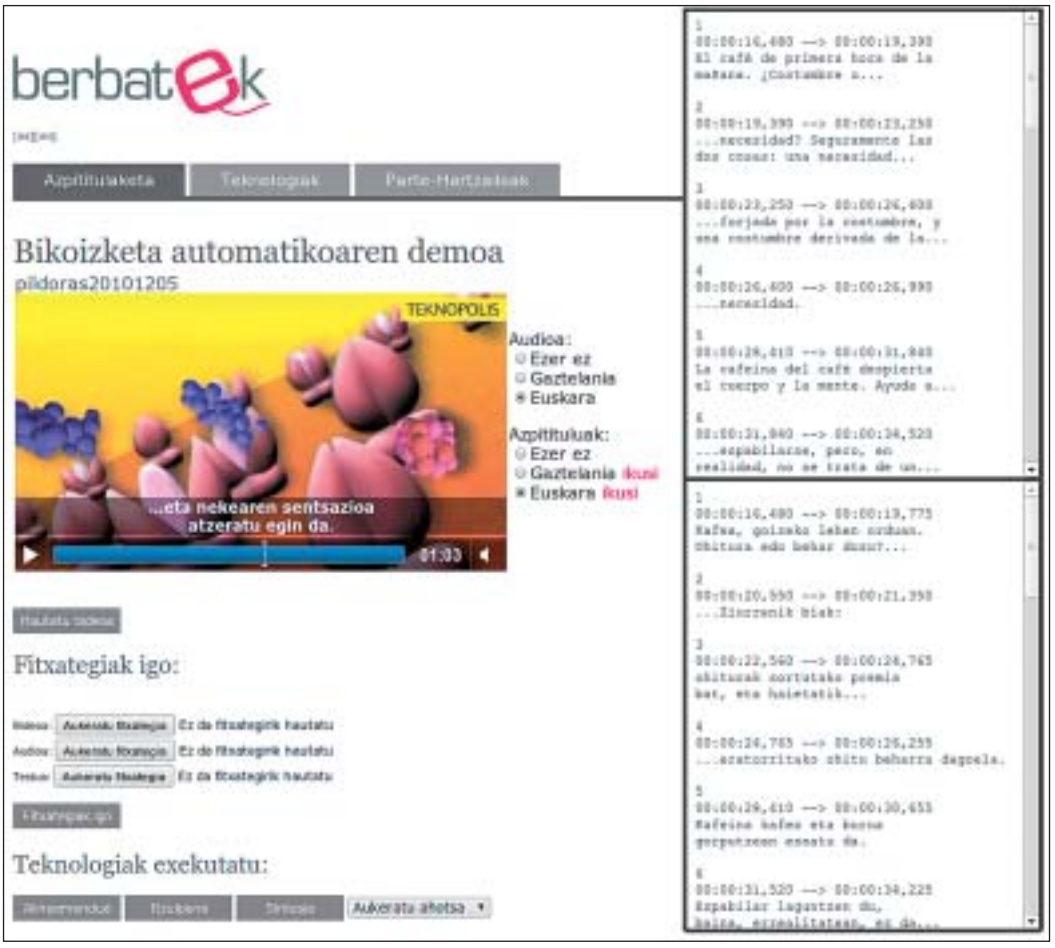
Esan bezala, BerbaTek-ek hizkuntzen industrian aplikazio praktikoa izateko bokazioa du, eta, horren erakusgarri, industria horren hiru azpisektoreentzat demo bana eraiki dugu teknologia ezberdinak konbinatuz.

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 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.

Itzulpenaren alorrerako, dokumentalen bikoizketa automatikoko demo bat egin da. 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 egin dugu. 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 azpitituluok, 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.

Azkenik, 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 eus-





Dokumentalen bikoizketa automatikorako demoa.

karaz 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.

DIBULGAZIOA

BerbaTek proiektuan garrantzia ematen diogu, oinarrizko ikerketaz eta aplikazio praktikoaz gainera, dibulgazioa egiteari. Guretzat funtsezkoa da egindako lanaren berri ikerketako foro, kongresu eta aldizkari espezializatuetan ematea, baina baita gizarte zabalari hizkuntza- eta ahots-teknologien garrantzia erakustea eta euskararentzat arlo horretan egin ditugun lorpenak ezagutaraztea ere. Azken helburu hori lortzeko, webgune bat egin dugu (http://www.berbatek.com), non, BerbaTek proiektuari buruzko informazio orokorra emateaz gainera, aldian-aldian bertan egindako aurrerapenen berri ematen baitugu. Eta horrez gainera, hizkuntza-teknologien munduan gertatzen dena ezagutarazten dugu Hizkuntza, Ahots eta Multimedia Teknologien Behatokiaren bidez (beste webgune batzuetako albisteen batzaile bat), bai eta hemengo nahiz nazioarteko ekitaldi garrantzitsuenen berri eman ere Ekitaldien egutegiaren bidez.

Oso gustura eta harro gaude BerbaTek proiektuan lortutako emaitzekin. Baina euskarak hizkuntza-teknologietan eta, beraz, mundu digital berri horretan atzean gelditu nahi ez badu, oraindik gogor lan egitea tokatuko zaigu hurrengo urteetan ere. BerbaTek proiektua aurrera eraman dugun kide guztiok prest gaude erronka horri heltzeko.