META NET

White Paper Series Liburu Zurien Bilduma

THE BASQUE LANGUAGE IN THE DIGITAL AGE

EUSKARA ARO DIGITALEAN

Inmaculada Hernáez
Eva Navas
Igor Odriozola
Kepa Sarasola
Arantza Diaz de Ilarraza
Igor Leturia
Araceli Diaz de Lezana
Beñat Oihartzabal
Jasone Salaberria

Springer

White Paper Series Liburu Zurien Bilduma

THE BASQUE LANGUAGE IN THE DIGITAL AGE

EUSKARA ARO DIGITALEAN

Inmaculada Hernáez [1]
Eva Navas [1]
Igor Odriozola [1]
Kepa Sarasola [1]
Arantza Diaz de Ilarraza [1]
Igor Leturia [2]
Araceli Diaz de Lezana [3]
Beñat Oihartzabal [4]
Jasone Salaberria [4]

[1] Univ. del País Vasco/Euskal Herriko Unibertsitatea
[2] Elhuyar Foundation
[3] Gobierno Vasco/Eusko Jaurlaritza
[4] UMR 5478 IKER

Georg Rehm, Hans Uszkoreit
(editoreak, editors)

HITZAURREA

Liburu zuri hau hizkuntza-teknologiei eta haien potentzialei buruzko jakintza sustatzea helburu duen bildumaren atal bat da, hezitzaileei, kazetariei, politikariei eta hizkuntza-komunitateei zuzendua.

Europan, desberdina da, hizkuntza batetik bestera, hizkuntza-teknologien eskuragarritasuna eta erabilera. Horren ondorioz, desberdinak behar dute izan, halaber, hizkuntza bakoitzerako hizkuntza-teknologien ikerketa eta garapena bultzatzeko behar diren ekimenak.

Europako Batzordeak sortutako META-NET Bikaintasun Sareak gaur egungo hizkuntza-baliabideei eta -teknologiei buruzko analisi bat bideratu du liburu zurien bilduma honetan (p. 75). Analisi hori Europako 23 hizkuntza ofizialentzako eta Europako beste zenbait nazio– eta eskualde-hizkuntza garrantzitsurentzako gauzatu da. Analisiaren ondorio gisa, ondorioztatu da ikerketa-hutsune esanguratsuak daudela hizkuntza bakoitzerako. Adituen gaur egungo egoeraren analisi eta ebaluazio xeheago batez, etorkizuneko ikerketen eragina handiagotu eta arriskuak gutxiagotu litezke.

Enpresa-munduko, administrazio publikoko, industria-sektoreko, ikerketa-alorreko, software-enpresetako, teknologia-hornitzaileetako eta unibertsitate europarretako parte-hartzaileekin lanean diharduten 33 herrialdetako 54 ikerketa-zentroz (p. 71) osatuta dago META-NET. Denak elkarrekin, teknologiari buruzko ikuspegi bateratu bat ari dira sortzen, eta, aldi berean, 2020 bitartean ikerketa-hutsuneak hizkuntza-teknologien bidez betetzeko bideak zein izan daitezkeen azaltzen duen ikerketa-agenda estrategiko bat ere ari da garatzen.

PREFACE

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, journalists, politicians, language communities and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 75). This analysis focussed on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed expert analysis and assessment of the current situation will help maximise the impact of additional research and minimise any risks.

META-NET consists of 54 research centres from 33 countries (p. 71) that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

Dokumentu honen egileek beren eskerrik beroenak adierazi nahi dizkie alemanezko liburu zuriaren [1] egileei, haien dokumentuko zenbait atal, hizkuntzaren araberakoak ez direnak, berrerabiltzeko baimena emateagatik.

Liburu zuri hau Europako Batzordeko Zazpigarren Esparru Programaren eta IKTak Sustatzeko Programa Estrategikoaren diru-laguntzari esker garatu da, T4ME (249 119 Dirulaguntza Hitzarmena), CESAR (271 022 Dirulaguntza Hitzarmena), METANET4U (270 893 Dirulaguntza Hitzarmena) eta META-NORD (270 899 Dirulaguntza Hitzarmena) kontratuen baitan.

The authors of this document are grateful to the authors of the White Paper on German [1] for permission to re-use selected language-independent materials from their document.

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

AURKIBIDEA CONTENTS

EUSKARA ARO DIGITALEAN

THE BASQUE LANGUAGE IN THE DIGITAL AGE

1

LABURPENA

Hizkuntza gizakien arteko komunikazio-biderik garrantzitsuena da. Ideiak eta sentimenduak adierazteko aukera ematen digu, ikasten eta irakasten laguntzen digu, ezinbestekoa da bizitzeko, kulturaren transmisiorako tresnarik garrantzitsuena da, eta haren hiztunen identitate-ikurra da.

> Hizkuntza gizakien arteko komunikazio-biderik garrantzitsuena da.

Gaur egun, mundu globalizatu honetan, edozein lekutako pertsonekin erraz komunikatzeko tresna asko ditugu. Adibidez, informazio- eta komunikazio-teknologia berriek sare sozialak garatzeko aukera eman dute, eta, hala, ekarpen handia izan da herrialde eta kultura desberdinetako pertsonak harremanetan jartzea sustatzeko. Azken urteotan, halaber, atzerritarren mugimendua handia izan da gure herrialdeetan, dela turismoagatik, dela immigrazioagatik, eta horrek hainbat hizkuntzatan komunikatzeko beharra sortu du. Halere, hizkuntzarteko komunikazio-arazoak, maiz, lingua franca erabiliz gainditzen dira.

Europa aniztasun kultural eta linguistikoaren adibide garbia da, nahiz eta azken 60 urteetan zehar bateratzen politiko eta ekonomikoa izan duen. Hortaz, euskaratik polonierara zein italieratik islandierara, eragozpen linguistikoak gainditu behar dira, nahitaez, Europako hiritarren arteko eguneroko komunikazioan nahiz enpresaren eta politikaren esferetako komunikazioetan. Europar Batasunaren erakundeek bilioi bat euro behar dute urtean, beren eleaniztasun-politika betearazteko; alegia, testuak itzultzeko eta ahozko jarduerak interpretatzeko. Bitartean, ingelesa lingua franca bihurtzen ari da Europako hiritarren arteko komunikazioan.

Espainiar estatuan ere antzeko eszenatokia dugu. Estatu osoan, hizkuntza ofizial bakarra dago: espainiera edo gaztelania; horiez gainera, hiru hizkuntza koofizial daude: euskara, galiziera eta katalana. Estatuan eleaniztasunari eustea ez da lan erraza izan; aitzitik, espainierak gainerako hizkuntzen artean duen gailentasunean, identitate kulturalaren eta linguistikoaren babesean oinarritutako prozesu konplexu baten emaitza izan da. Europako kasuan lingua franca gisa ingelesa erabiltzen den era berean, espainiera erabiltzen da maiz espainiar estatuko hizkuntza-eremu desberdinetako hiritarren arteko komunikazio zuzenerako.

> Eleaniztasuna babestu beharreko kultura-ondarea da.

Bai Europari dagokionez, bai espainiar estatuari dagokionez, babestu beharreko kultura-ondarea da eleaniztasuna. Globalizazioak, lingua francaren erabilera gailendu eta gure hizkuntzaren erabilera murrizteko arriskua sortzen baitu, ez luke gure hizkuntza- eta kultura-ondare aberatsa alboratzen duen mekanismo bihurtu behar. Komunikazio-mundu global baten barnean, gure hizkuntza eta, harekin batera, gure identitate kulturala babesteko bideak aurkitu behar ditugu.

Gaur egungo hizkuntza-teknologiek eta ikerketa linguistikoek ekarpen handia egin dezakete eragozpen linguistiko horiek gainditzeko; izan ere, hizkuntza-

teknologiak, tresna eta aplikazio adimendunekin batera erabiliz, oso lagungarriak izango dira europarrek aise hitz egin eta salerosketak egin ditzaten, baita hizkuntza berean ari ez badira ere. Hizkuntza-teknologiek eskaintzen dituzten koponbideak hizkuntza europarren arteko zubi bikaina izan daitezke. Gaur egun merkatuan lor daitezkeen itzulpen automatikoko nahiz hizketa prozesatzeko tresnak – galderei erantzuteko sistemetatik hasi eta hizkuntza naturala darabilten interfazeetaraino, eta, besteak beste, itzulpen automatikoko sistemak eta laburpen-tresnak darabiltenak –, oraindik ere aski urrun daude asmo handiko helburu horretatik.

Hizkuntza-teknologiek eskaintzen dituzten koponbideak hizkuntza europarren arteko zubi bikaina izan daitezke.

1970eko hamarkadaren amaierarako, EB konturatua zen hizkuntza-teknologiek Europaren batasunaren gidaritzan izango zuten garrantziaz, eta, hala, lehen ikerketa-proiektua sortu zuen. Aldi berean, oso emaitza baliagarriak izan zituzten estatu-mailako proiektu asko ere jarri zituzten abian, baina inoiz ere ez europar ekintza kontzertatu baten gidaritzapean. Eremuko eragilerik garrantzitsuenak, batez ere, egoitza Amerikako Estatu Batuetan duten enpresa pribatu irabazi asmodunak dira. Gaur egungo hizkuntz-teknologia aurreratuenak hurbilketa estatistiko ez-zehatzetan oinarritzen dira eta ez dute aparteko metodo nahiz ezagutza linguistikorik erabiltzen. Esaterako, esaldiak automatikoki itzultzen dira esaldi bat gizakiek aurrez itzulitako milaka esaldirekin konparatuz. Emaitzaren kalitatea, hein handi batean, erabiliko den corpusaren tamainaren eta kalitatearen araberakoa da. Azaleko metodo estatistiko horiekin testu-material kantitate nahikoa duten hizkuntzetan esaldi sinpleak itzulita emaitza erabilgarriak lor daitezke, baina litekeena da huts egitea hizkuntza baten testu kopurua txikiagoa baldin bada edo egitura konplexuak dituzten esaldiak itzuli nahi badira. Hizkuntzen egitura-ezaugarri sakonagoa aztertuta, ondoriozta daiteke aurrerabide bakarra dela, Europako hizkuntza multzo zabaleko guztietarako ondo funtzionatuko duten aplikazioak garatuko badira.

Hortaz, hizkuntzarteko komunikazio-arazoen konponbidea da teknologia giltzarriak garatzea. Helburu hori betetzeko, baina Europako kultura- eta hizkuntza-aniztasunari eutsita, behar-beharrezkoa da lehendabizi Europako hizkuntza guztien bereizgarri linguistikoak aztertzea eta hizkuntza bakoitzerako hizkuntza-teknologiek duten gaur egungo laguntzen analisi sistematiko bat burutzea. Euskararako analisia aurkeztea da, hain zuzen, liburu honen xedea. Hala, euskarako hizkuntza-teknologien, aplikazioen eta konponbideen analisi xehatua aurkezten du ale honek.

Euskara ikerketa sustatu beharra duten EBko hizkuntzetariko bat da.

Hizkuntza-teknologien eremuan, hainbat produktu, teknologia eta baliabide daude euskararako. Badira aplikazio-tresnak hizketa sintetizatzeko, hizketa ezagutzeko nahiz ortografia zuzentzeko; badira, halaber, itzulpen automatikoko aplikazio batzuk ere, espainieratik euskarara batez ere.

Liburu zuri bilduma honetan ageri denez, ikaragarrizko aldea dago Europako estatu kideen hizkuntza-baliabideen inbentarioen artean eta ikerketa-egoeren artean. Ondoriorik nabarmenenetariko bat da ikerketa sustatu beharra duten EBko hizkuntzetariko bat dela euskara, hizkuntza-teknologietan oinarritutako aplikazio benetan eraginkorrak eta egunero jardunean erabiltzeko modukoak garatuko badira. Euskararako kalitate handiko hizkuntza teknologiaren garapena oso larria eta garrantzitsua da euskara sustatzeko.

2

ARRISKUA, GURE HIZKUNTZENTZAT ETA ERRONKA HIZKUNTZA-TEKNOLOGIARENTZAT

Komunikazioan eta gizartean izugarrizko eragina izaten ari den iraultza digital baten aurrean gaude. Komunikazio-teknologia digitalizatu eta sarekoetan izan berri diren aurrerapenak Gutenbergek inprenta asmatu zuenekoarekin alderatzen dira, batzuetan. Zer esaten digu analogia horrek Europako informazio-gizartearen eta geure hizkuntzen etorkizunari buruz?

> Iraultza digitala Gutenbergek inprenta asmatu izanarekin alderatu daiteke.

Gutenbergen asmakizunaren ondoren, benetako aurrerapausoak eman ziren komunikazioan eta ezagutzen trukaketan, hainbat lani esker; esaterako, Lutherrek egin-nikoa, Biblia hizkuntza arrunt batera itzuli zuenean. Hurrengo mendeetan, teknika kulturalak garatu dira hizkuntza-prozesamendua eta ezagutza- trukaketa hobeto egiteko:

- Hizkuntza handien ortografia eta gramatika estandarizatzeak aukera eman zuen ideia zientifiko eta intelektual berriak azkar zabaltzeko;
- Hizkuntzen ofizialtasunak aukera eman zien herritarrei muga jakin batzuen barruan (sarritan, politikoak) komunikatzeko;
- Hizkuntzen irakaskuntzari eta itzulpenari esker, hizkuntzen arteko trukaketa etorri zen;
- Kazetaritzako eta bibliografiako jarraibideak sortzeak material argitaratuaren kalitatea eta eskuragarritasuna bermatu zuten;
- Hedabide berriek – egunkariek, irratiak, telebistak, liburuek eta beste batzuek – komunikazio-beharrei erantzun zieten.

Azken hogei urteotan, informazio-teknologiak lagundu egin du prozesu horietako asko automatizatzen eta errazten:

- Autoedizioko softwareak hartu du idazmakinaren eta monotipoaren tokia;
- Microsoft PowerPoint programak hartu du proiektagailuz erakutsitako gardenkien tokia;
- Mezu elektronikoen bidez faxez baino azkarrago bidali eta jasotzen dira dokumentuak;
- Skype erabiliz, Interneteko telefono-deiak egin daitezke, eta elkargune birtualak sortu;
- Audio- eta bideo-fitxategien kodetze-formatuei esker, erraza da multimedia-fitxategiak trukatzea;
- Bilatzaileetan gako-hitzak sartuz web-orrietara sar gaitezke;
- Lineako zerbitzuek itzulpen azkar eta gutxi gorabeherakoak sortzen dituzte; hor dugu, esaterako, Google Translate;
- Gizarte-hedabideen plataformek erraztu egiten dute elkarlana eta informazioa partekatzea.

Tresna eta aplikazio horiek lagungarriak badira ere, oraindik ez dute lortu informazio-gizarte europar eleaniztun eta jasangarri bat ezartzea; gizarte moderno eta inklusibo bat, non informazioa eta produktuak askatasunez ibiltzen diren alde batetik bestera.

2.1 HIZKUNTZA-MUGEK OZTOPOAK JARTZEN DIZKIOTE EUROPAKO INFORMAZIO-GIZARTEARI

Ezin dugu jakin etorkizuneko informazio-gizartea zehazki nolakoa izango den. Europaren energia-estrategiaz edo atzerriko politika bateratuaz hitz egin behar denean, Europako atzerriko ministroak beren jatorrizko hizkuntzan mintzatzen entzun nahi izango ditugu, beharbada. Agian, plataforma bat izan nahi dugu, non hainbat hizkuntza hitz egiten dituzten eta era askotako hizkuntza-mailak dituzten pertsonak gai jakin bati buruz solasean arituko diren, teknologiak haien iritziak bildu eta laburpen txikiak egiten dituen bitartean. Baliteke, halaber, beste herrialde batean dagoen osasun-aseguruen bulego batekin hitz egin nahi izatea.

> Ekonomia- eta informazio-eremu globalak hizkuntza, hizlari eta eduki desberdinen aurrean jartzen gaitu.

Argi dago gaur egungo komunikazioak, duela urte batzuetakoaren aldean, beste kalitate batekoa izan behar duela. Ekonomia eta informazio-eremu globalean, hizkuntza, hiztun eta eduki gehiagorekin egiten dugu topo, eta hedabide mota berriekin berehala harremanetan jartzeko eskatzen digute. Gizarte-hedabideak (Wikipedia, Facebook, Twitter eta YouTube) izaten ari diren arrakasta icebergaren tontorra besterik ez da.

Gaur egun, hainbat gigabyteko testuak bidal ditzakegu mundu osora segundo gutxi batzuetan, ulertzen ez dugun hizkuntza batean dagoela ohartu baino lehen. Europako Batzordeak eskatuta duela gutxi egin den txosten baten arabera, Europako Internet-erabiltzaileen % 57k bere jatorrizko hizkuntzaz bestelako hizkuntzetan erosten ditu produktuak eta zerbitzuak (ingelesa da gehien erabiltzen den atzerriko hizkuntza, eta, haren ondoren, frantsesa, alemana eta gaztelania). Erabiltzaileen % 55ek irakurtzen ditu edukiak atzerriko hizkuntza batean, baina soilik % 35ek idazten ditu mezu elektronikoak edo sareko iruzkinak [2]. Duela urte gutxi batzuk, ingelesa izan zen sareko lingua franca – sareko edukiaren parte oso handi bat ingelesez zegoen –, baina egoera goitik behera aldatu da orain. Beste hizkuntza batzuetan idatzitako edukien kantitatea izugarri handitu da sarean (batez ere, asiar eta arabiar hizkuntzetan idatzitakoena). Hizkuntza-mugek eragin duten nonahiko banaketa digitalak ez du toki handirik hartu diskurtso publikoan. Alabaina, galdera bat sortzen du behin eta berriz: "Europako zein hizkuntzak egingo du aurrera eta iraungo du sareko informazioaren eta ezagutzaren gizartean?".

2.2 GURE HIZKUNTZAK ARRISKUAN

Inprentak informazio-trukaketa eskerga ekarri zuen Europara, baina bertako hizkuntza asko desagertzea ere eragin zuen. Eskualdeetako hizkuntzak eta hizkuntza txikiak apenas erabiltzen ziren argitalpenetarako. Horren ondorioz, hizkuntza asko ahozko transmisiora mugatu ziren – adibidez, kornubiera eta dalmaziera –, eta, beraz, mugatuta gelditu zen haien etengabeko ikaskuntza, zabalkundea eta erabilera.

Hizkuntza-aniztasuna da Europaren kultura-ondasun aberats eta garrantzitsuenetakoa (80 bat hizkuntza ditu). Europaren hizkuntza-aniztasuna haren arrakasta sozialaren ezinbesteko parte ere bada. Hiztun askoko

hizkuntzek eutsiko diote suspertzen ari den gizartean eta merkatu digitalean duten tokiari, dudarik gabe, baina baliteke Europako hizkuntza asko komunikazio digitaletatik baztertuta geratzea eta garrantzia galtzea Interneteko gizartearen begietara. Hori ez litzateke batere ona izango. Alde batetik, aukera estrategiko bat galduko litzateke, eta horrek ahuldu egingo luke Europaren posizioa munduan. Bestetik, gertaera horiek ez datoz bat Europako herritar guztien (edozein hizkuntza izanda ere) berdintasunezko parte-hartzea bermatzeko helburuarekin. UNESCOk eleaniztasunaren inguruan eginiko txostenak dioenez, hizkuntza funtsezko bitartekoa da oinarrizko eskubideez gozatzeko – adibidez, adierazpen politikoa, hezkuntza eta gizarteko parte hartzea [3].

Hizkuntza-aniztasuna da
Europako kultura-ondasun aberats
eta garrantzitsuenetarikoa.

2.3 HIZKUNTZA-TEKNOLOGIA TEKNOLOGIA BIDERATZAILE GILTZARRIA DA

Lehen, hizkuntzaren irakaskuntzara eta itzulpenetara bideratzen ziren inbertsioak. Adibidez, kalkulu batzuen arabera, itzulpen, interpretazio, software-lokalizazio eta webgune globalizazioaren merkatu europarra 8,4 mila milioi eurokoa zen 2008an, eta urtean % 10 haztea espero zen [4]. Alabaina, merkatu horren ahalmena ez da nahikoa oraingo eta geroko beharrak asetzeko.

Hizkuntza-teknologia teknologia bideratzaile giltzarria da, Europako hizkuntzak babestu eta bultzatu ditzakeena. Hizkuntza-teknologiak laguntza ematen dio jendeari elkarlanean aritzeko, negozioak egiteko, ezagutza besteekin banatzeko, eta eztabaida sozial eta politikoetan parte hartzeko, dena delako hizkuntza-mugak eta informatikako trebetasunak izanda ere. Hizkuntza-teknologiak laguntzen digu jada eguneroko lanetan, hala nola mezu elektronikoak idaztean, lineako ikerketa bat egitean edo hegaldi bat erreserbatzean. Eragiketa hauek egitean ere hizkuntza-teknologiaz baliatzen gara:

- Interneteko bilatzaile baten bidez informazioa aurkitzen dugunean.;
- Testu-prozesadore batean ortografia eta gramatika egiaztatzen dugunean;
- Lineako denda batean produktu baten gaineko gomendioak begiratzen ditugunean;
- Nabigazio-sistema baten ahozko jarraibideak entzuten ditugunean;
- Lineako zerbitzu baten bidez web-orriak itzultzen ditugunean.

Lan honetan agertzen diren hizkuntza-teknologiak etorkizuneko aplikazio berritzaileen oinarrizko osagaia dira. Hizkuntza-teknologia, normalean, teknologia bideratzailea izaten da, eta aplikazio-plataforma handiago baten barruan joan ohi da, nabigazio-sistema edo bilatzaile baten barruan adibidez. Liburu zuri honetan aztertzen da teknologia komunak hizkuntza bakoitzerako zenbateraino dauden prestatuta.

Europak hizkuntza-teknologia sendoak
eta modu onean erosteko modukoak
behar ditu hango hizkuntza guztietarako.

Laster behar izango dugu Europako hizkuntza guztietarako hizkuntza-teknologia bat, eskuragarri dagoena, modu onean eros daitekeena eta software-esparru handiagoetan ondo integratuta dagoena. Erabiltzaileak ez du funtzio interaktiboak, multimediakoak eta eleaniztunak erabiltzerik hizkuntza-teknologiarik gabe.

2.4 HIZKUNTZA-TEKNOLOGIAREN AUKERAK

Hizkuntza-teknologiak aukera eman diezaieke Europako hizkuntza guztiei itzulpen automatikoak egiteko, edukiak sortzeko, informazioa prozesatzeko eta ezagutzak kudeatzeko. Hizkuntza-teknologiak balio dezake, halaber, etxetresna elektroniko, aparailu, ibilgailu, ordenagailu eta robotentzako hizkuntzan oinarritutako interfaze intuitiboak garatzeko. Hainbat prototipo atera diren arren, aplikazio komertzial eta industrialak hasierako fasean daude oraindik. Ikerketan eta garapenean egin berri diren lorpenek benetako abagunea eman diote. Adibidez, itzulpen automatikoak (IA) zehaztasun handi samarra ematen du jada esparru jakin batzuetan, eta aplikazio esperimentalek informazio eleaniztuna eta ezagutza-kudeaketa eskaintzen dute, baita Europako hizkuntza askotan edukiak sortzeko aukera ere. Hizkuntza-aplikazioak, ahots bidezko erabiltzaile-interfazeak eta elkarrizketa-sistemak oso eremu espezializatuetan egon ohi dira eskuarki, eta errendimendu mugatua izan ohi dute maiz. Hizkuntza-teknologia hondamen eremuetako erreskate-lanetarako erabiltzea ari dira ikertzen orain. Arrisku handiko inguru horietan, itzulpenaren zehaztasuna hil ala biziko kontua izan daiteke. Antzekoa gertatzen da osasungintzan ere. Hizkuntzarteko ahalmenak dituzten robot inteligenteek biziak salba ditzakete.

Hizkuntza-teknologiek sekulako merkatu-aukerak dituzte hezkuntzan eta entretenimenduen industrian; izan ere, jokoetan, joko hezigarrietan, simulazioetan eta prestakuntza-programetan integra daitezke. Hizkuntza-teknologiak zeregin garrantzitsua izan dezake beste hainbat tokitan ere; besteak beste, mugikorretako informazio zerbitzuetan, ordenagailuz lagundutako hizkuntza-ikaskuntzarako softwarean, Internet bidezko ikaskuntzako inguruneetan, autoebaluazio-tresnetan eta plagioak aurkitzeko gailuetan. Gizarte-hedabideen aplikazioek (adib. Twitter eta Facebook) duten arrakastak iradokitzen du gero eta gehiago behar direla hizkuntza-teknologia sofistikatuak, gai direnak mezuak behatzeko, eztabaidak laburbiltzeko, iritzi-joerak iradokitzeko, erantzun emozionalak detektatzeko, copyrightaren arau-hausteak identifikatzeko eta erabilera desegokien jarraipena egiteko.

> Aniztasun linguistikoak sortzen dituen "eragozpenak" gainditzen laguntzen dute hizkuntza-teknologiek.

Hizkuntza-teknologiak aukera paregabea ematen dio Europar Batasunari, bai ekonomia aldetik, bai kultura aldetik. Eleaniztasuna arau bilakatu da Europan. Europako negozioak, erakundeak eta eskolak ere nazioartekoak eta askotarikoak dira. Herritarrek elkarrekin komunikatu nahi dute Europako Merkatu Batuan, oraindik ere hor dauden hizkuntza-mugez harago. Hizkuntza-teknologiak hor jarraitzen duten muga horiek gainditzen lagundu dezake, eta, era berean, hizkuntzaren erabilera askea eta irekia bultzatu. Gainera, Europako hizkuntzetarako hizkuntza-teknologia eleaniztun eta berritzaileak munduko beste herrialdeekin eta haietako komunitate eleaniztunekin komunikatzen lagunduko liguke. Hizkuntza-teknologiek nazioarteko ekonomia-aukera ugari ematen dituzte.

2.5 HIZKUNTZA-TEKNOLOGIAREN ERRONKAK

Azken urteotan hizkuntza-teknologiak aurrerapen handi samarra egin badu ere, aurrerapen teknologikoa eta produktuen berrikuntza erritmo motelean doaz gaur egun. Ezin dugu hamar edo hogei urtez itxaron gure inguru eleaniztuneko komunikazioa eta produktibitatea areagotuko duten hobekuntza nabariak agertu arte.

Aurrerapen teknologikoa eta produktuen berrikuntza erritmo motelean doaz gaur egun.

Erabilera handiko hizkuntza-teknologiak – hala nola testu prozesadoreetako ortografia- eta gramatika-zuzentzaileak – hizkuntza bakarrean izaten dira normalean, eta hizkuntza gutxi batzuetan bakarrik egon ohi dira eskuragarri. Komunikazio eleaniztunerako aplikazioek sofistikazio-maila bat eskatzen dute. Itzulpen automatikoa eta lineako zerbitzuak – esaterako, Google Translate edo Bing Translator – apartak dira dokumentu baten edukien gutxi gorabeherako itzulpenetarako. Baina, lineako zerbitzu eta IA aplikazio profesional horiek hainbat zailtasun izaten dituzte oso itzulpen zehatzak eta osatuak behar direnean. Okerreko itzulpen barregarri ezagunak asko dira (hor ditugu, esaterako, Bush edo Kohl izenen itzulpen literalak), eta agerian uzten dute zer-nolako erronkei egin behar dien aurre hizkuntza-teknologiak.

Aurrerabide teknologikoak arinago joan beharra du.

2.6 HIZKUNTZAREN JABETZEA GIZAKIENGAN ETA MAKINETAN

Ordenagailuek hizkuntza nola tratatzen duten eta hizkuntzaren jabetzea horren zaila zergatik den azaltzeko, ikus dezagun gizakiok nola jabetzen garen lehen eta bigarren hizkuntzez, eta, gero, itzulpen automatikoko sistemen funtzionamenduaren eskema egingo dugu – zerbaitengatik du hizkuntza-teknologiaren alorrak horren lotura estua adimen artifizialaren arloarekin.

Gizakiak bi modutan jabetzen dira hizkuntza-gaitasunez. Hasieran, hizkuntza bateko hiztunen arteko elkarreragina entzunez ikasten du umeak hizkuntza hori. Hizkuntzaren erabiltzaileek – gurasoek, anaiarrebek edo beste senide batzuek, esaterako – eratutako hizkuntza-adibide zehatzak entzuteak lehen hitzak eta esaldi laburrak esaten laguntzen die bi urte inguruko umeei. Hizkuntzak ikasteko antolaketa genetiko bereziak eman digu gizakioi gaitasun hori.

Normalean, bigarren hizkuntza ikasteak ahalegin askoz handiagoa eskatzen du, umea ez baitago bertako hiztunen hizkuntza komunitate baten barruan. Eskola garaian, atzerriko hizkuntzez jabetzeko, haien egitura gramatikala, hiztegia eta ortografia ikasten dira liburuetatik eta ikasmaterialetatik, eta, haietan, arau abstraktu, taula eta adibidezko testuen bidez azaltzen da hizkuntza. Atzerriko hizkuntza bat ikasteak denbora asko eta ahalegin handia eskatzen ditu, eta orduan eta zailagoa da adinean aurrera egin ahala.

Gizakiak bi modutan jabetzen dira hizkuntza-gaitasunez: adibideetatik ikasiz eta arau linguistikoak ikasiz.

Hizkuntza-teknologiaren bi sistema-mota nagusiak gizakien antzera jabetzen dira hizkuntza-ahalmenez. Metodo estatistikoan, hizkuntza bakarrean idatzitako adibidezko testu zehatzen bildumetatik edo bi hizkuntza edo gehiagoko testu paralelo deritzenetatik lortzen da hizkuntza-ezagutza. Ikaskuntza automatikoko algoritmoen nolabaiteko hizkuntza-gaitasunak adieraz dezake hitzak, esaldi laburrak eta esaldi osoak zuzentasunez nola erabili hizkuntza batean edo nola itzuli hizkuntza batetik bestera. Metodo estatistikoetarako behar den esaldi-kopurua ikaragarria da. Lanaren kalitatea handiagotu egiten da zenbat eta testu gehiago aztertu. Milioika esaldiko testuen gainean probatzen dituzte maiz sistema horiek. Horregatik ibiltzen dira bilatzaileen hornitzaileak ahalik eta idatzizko

material gehien bildu nahian. Testu-prozesadoreen ortografia-zuzentzaileek, linean eskuragarri dagoen informazioak eta Google Search eta Google Translate bezalako itzulpen-zerbitzuek metodo estatistikoa (datuek gidaturikoa) dute oinarrian.

Erregeletan oinarritutako sistemak dira bigarren hizkuntza-teknologia nagusia. Hizkuntzalaritzako, hizkuntzalaritza konputazionaleko eta informatikako adituek azterketa gramatikalak (itzulpen-arauak) kodetu eta hiztegi-zerrendak (lexikoiak) osatzen dituzte. Erregeletan oinarritutako sistema ezartzeko, denbora eta esku-lan asko behar da. Era berean, oso aditu espezializatuak behar dira halako sistemak sortzeko. Erregeletan oinarritutako itzulpen automatikoko sistema garrantzitsuenetako batzuek etengabeko garapena izan dute azken hogei urteotan. Erregeletan oinarritutako sistemen alde ona da adituek kontrol zehatzagoa lor dezaketela hizkuntzaren prozesamenduaren gainean. Hortaz, softwareko akatsak sistematikoki zuzendu daitezke eta feedback zehatza eman dakioke erabiltzaileari, erregeletan oinarritutako sistemok hizkuntza ikasteko erabiltzen direnean batik bat. Finantza-sarrera mugatuak direla bide, erregeletan oinarritutako hizkuntza-teknologia hizkuntza handietan bakarrik erabil daiteke.

Sistema estatistikoaren eta erregelatan oinarritutako sistemaren indarguneak eta ahuleziak osagarriak izan ohi direnez, gaur egungo ikerketen joera da bi metodologiak batera lantzen dituen metodo hibridoa erabiltzea. Hala eta guztiz ere, metodo horiek, aplikazio industrialetan ez dute izan, orain arte, ikerketa-laborategietan bezain emaitza onik.

Kapitulu honetan ikusi dugunez, gaur egungo informazio-gizartean oso erabiliak diren aplikazio askok hizkuntza-teknologiak dituzte oinarrian. Areago gertatzen da hori Europako ekonomiaren eta informazioaren esparruan, hizkuntza anitzeko komunitatea dela kontuan hartzen badugu. Hizkuntza-teknologiek urte gutxian aurrerapauso handiak eman badituzte ere, bide luzea dago, oraindik ere, hizkuntza-teknologietan oinarritutako sistemen kalitatea hobetzeko. Datozen ataletan, euskarak Europako informazio-gizartean duen eginkizuna deskribatuko dugu, eta, halaber, euskarazko hizkuntza-teknologien egungo egoera aztertuko dugu.

3

EUSKARA EUROPAKO INFORMAZIO GIZARTEAN

3.1 DATU OROKORRAK

Euskara, Nafarroako Erresumako hizkuntza nagusia zelako latinez "Lingua Navarrorum" esaten zitzaiona, mendebaldeko Europan bizirik dagoen hizkuntza preindoeuropar bakarra da. Hizkuntza bakartutzat jotzen da, ez baitzaio loturarik aurkitu beste hizkuntzekin, antzinako akitanierarekin izan ezik. Euskararen jatorria nahiz beste hizkuntzekiko duen lotura gai gatazkatsuak eta interesgarriak dira oraindik ikerlarientzat.

Euskarak 800.000 hiztun inguru ditu.

Euskara, gaur egun, eskualde txiki batean hitz egiten da, Pirinioen mendebaldean, Espainiaren eta Frantziaren arteko mugaren bi aldeetan, euskaldunek *Euskal Herria* deritzen eskualdean. Hizkuntza lurrak galduz joan da hainbat mendez, hegoaldean batez ere. Duela gutxi, Francoren diktaduraren garaian, euskara erabiltzea debekatu zutela eta, hizkuntzak berreskuratu ezinezko galera izan zuen.

Ikaragarrizko ahaleginak egin ziren hizkuntza bizibeerritzeko; batez ere, 1960ko hamarkadan, ikastolen sorrerari esker euskara hezkuntza-sisteman sartu zenean; klandestinitatean hasierako urteetan. Alabaina, euskararen berreskuratze-prozesua ez zen hasi 1980ko hamarkadara arte, autonomiak sortu eta Eusko Jaurlaritzari hizkuntzaren gaineko eskumen politikoak eman zitzaizkion arte.

Ahalegin ikaragarriak egin ziren arren, euskara hizkuntza "ahul" moduan agertu zen 2009an Unescok Munduko Arriskupeko Hizkuntzen Mapan [5] atera zuenean. Gaur egun, kalkulatzen da Euskal Herriko biztanleriaren [6] % 26 inguruk hitz egiten duela euskara, Espainiaren administraziopeko aldean nahiz Frantziaren administraziopeko aldean, baina bietan ez du estatus bera.

Alde batetik, Euskal Herriaren Espainiako partea bi eskualde politikotan banatuta dago: Euskal Autonomia Erkidegoan, euskara koofiziala da gaztelaniarekin batera, baina gaztelaniaren alderako zenbait desberdintasunekin; Nafarroako Foru Erkidegoan, hiru eremu daude, euskararen legezko estatusaren araberakoak: euskalduna, erdalduna eta mistoa. Hizkuntzarekiko atxikimendua eta hizkuntza-eskubideak ezberdinak dira hiru eremu horietan. Bestetik, Frantziako aldean, Pirinio Atlantikoetako Departamentuaren mendebaldeko partean hitz egiten da euskara, baina inoiz ez du izan inolako legezko estatusik eta ez da ofiziala inongo erakundetan. Dena dela, duela urte batzuk (2004an), erakunde publiko bat sortu zen Iparraldean euskara bultzatzeko helburuarekin.

Ahozko euskarak oso dialekto-sakabanaketa handia du. Gaur egun, onartzen da sei euskalki daudela, elkarren artean alde nabariak dituztenak. Euskara batua ez zen ofizialki ezarri 1968. urtera arte, orduan egin baitzuen Euskaltzaindiak [7] lehen estandarizazio-proposamena. Euskalkiok ezberdintasun nabariak dituzte hainbat alderditan: hiztegia, fonetika, morfofonologia eta proso-

dia, azentua eta intonazioa. Euskalkiak ez dira entitate homogeneoak; etengabe aldatzen dira batetik bestera, eta, batzuetan, ez dago hain argi bi edo hiruren arteko muga.

3.2 EUSKARAREN BEREZITASUNAK

Euskara hizkuntza eranskaria eta flexio handikoa da, eta hizkuntza ergatibo-absolutiboa izatea da bereizgarri nagusia. Horrek esan nahi du aditz iragangaitz baten subjektua absolutibozko kasuan (markatu gabekoan) joan ohi dela, eta kasu hori bera erabiltzen dela aditz iragankorren objektu zuzenarentzat; aditz iragankorraren subjektua beste era batean markatzen da, ergatibozko kasuaren bidez: *-k* atzizkia.

Euskara postposiziozko hizkuntza da; beraz, kasuzko eta postposiziozko esaldiak sintagmaren amaieran atzizki bat edo gehiago gehituta eratzen dira, eskema honen arabera:

erroa + (artikulua) + (numeroa) + [kasua(k)]

Adibidez, "mutilarenagana" honela dago osatuta: «mutil+a+Ø+r+en+gan+a», – "mutil" lema edo izenerroa da; "a", artikulua; "Ø", singularreko marka; "r", epentesi-partikula; "en", edutezko genitiboa; "gan" izaki bizidunen marka; eta "a", adlatiboa.

Ezaugarri garrantzitsua da hori, hizkuntza naturalean eta hizketa-prozesamenduan kontuan hartu beharrekoa; bada, izen sintagma bakoitzak 17 deklinabide-kasu izan ditzake, eta lau aldiz forma gehiago har ditzake zehaztasunaren eta numeroaren arabera. Hasierako 68 forma horiek are gehiago alda daitezke esaldiko beste zati batzuen arabera – hango izenen arabera ere deklinatzen direlako. Kalkulatzen da bi mailatako errekurtsioarekin euskarazko izen batek 275 deklinabide-marka izan ditzakeela, eta hori oso ohiko fenomenoa [8] da, gainera. Horrek aditzera ematen du beharrezkoa dela hainbat eratako bukaera horiek guztiak tratatzeko modu bat aurkitzea, oinarrizko hiztegi batetik abiatuta.

Aditzak dira euskara hizkuntza eranskaria dela erakusten duen beste adibide bat. Aditz laguntzailea aditz nagusi gehienekin batera joan ohi da, eta, subjektuarekin ez ezik, esaldian dagoen beste edozein objektu zuzen edo zeharkakorekin ere komunztatzen da. Europako hizkuntzen artean, pertsona askorekiko komunztadura hori euskaran, Kaukasoko hizkuntza batzuetan eta hungarieran baino ez da topatu (guztiak ez-indoeuroparrak). Euskaran, aditzen eskema honi jarraitzen zaio:

[aditz-erroa + aspektu-marka] [aditz laguntz.]

Adibidez, euskara batuan, «esaten zenizkidaten» honela dago osatuta: «esan» (aditz-erroa) + «ten» (maiztasun-aspektua) eta «zen+i+zki+da+Ø+te+n» aditz laguntzailea («zen» bigarren pertsonaren ergatibo-marka da; «i», aditz laguntzailearen erroa; «zki», hirugarren pertsona pluralaren absolutibo-marka; «da», lehen pertsona singularraren datibo-marka; «Ø», indikatibo marka; «te», pluralaren ergatibo-marka; eta «n», iraganaldiko marka). Aditza horren konplexua izanda, aditz laguntzaile bakoitza morfemetan banatu beharrean osorik tratatu ohi da hizkuntza naturalaren prozesamendurako ikerketetan.

Esaldiko hitzen hurrenkerari dagokionez, oinarrizko eraikuntza sintaktikoa subjektua-objektuak-aditza da (gaztelanian, frantsesean eta ingelesean, aldiz, subjektua-aditza-objektuak da ohikoena). Esaldi barruko sintagmen hurrenkera alda daiteke mintzagaiaren arabera, baina sintagma barruko hitzen hurrenkera zurruna da normalean. Bada, argitu behar da euskararen sintagma-hurrenkera mintzagaia-galdegaia dela; hau da, esaldi neutroetan (norbaiti gertaera baten berri emateko esaldiak, kasurako), mintzagaia ematen da lehenik, eta galdegaia ondoren. Halako esaldietan, aditz-sintagma amaieran joan ohi da. Laburbilduz, galdegaia aditz-sintagmaren aurre-aurrean doa.

Galderetan ere betetzen da arau hori; esaterako, «Zer da hau?» edo «Hau zer da?» esan daiteke, baina bietan «zer» galdetzailea «da» aditzaren aurre-aurrean doa. Arau hori horren garrantzitsua da euskaran, non beste hizkuntzetan ematen diren euskararen deskribapen gramatikaletan ere euskarazko hitza (*galdegai*) erabiltzen baita.

Euskararen ortografia fonemikoa da ia erabat: grafema bakoitza fonema bati dagokio, eta, beraz, hitz baten ahoskera erraz iragar daiteke idatzizko formatik. Dena dela, badaude salbuespen gutxi batzuk: <l> eta <n> letrek aurretik <i> letra eta atzetik bokal bat badute, bustidura egin ohi da; adib. mutila → <mutiLa>. Beste adibide bat "ez" ezezko partikularen amaierako kontsonante fonemarena da, atzetik datorren fonema ahoskabe bihurtzen baitu; adib. ez dira → <eztira>.

Euskarak sei soinu bokaliko eta hogeita hamabost soinu kontsonantiko ditu.

3.3 AZKEN GERTAERAK

Euskaltzaindiak, euskal hizkuntzaz arduratzen den erakunde ofizialak, euskararen forma estandarizatua atera zuen 1960ko hamarkadan, *Euskara Batua* izenekoa. Euskara egoera formaletan (hezkuntzan, hedabideetan, literaturan...) eta euskal hiztun guztiek ulertzeko eran erabil zedin asmatu zen euskara batua, eta gaur egun ere horretarako erabiltzen da batik bat.

Literatura klasikoko tradizioa dela eta, euskara batua erdialdeko euskalkian eta nafar-lapurteran oinarritzen da nagusiki. Ertzetako euskalkiak oso ezberdinak dira, nahiz eta mendebaldekoa hiztun gehien dituenetako bat izan erdialdekoarekin batera.

Euskara batuak oinarri sendoak ditu, eta aurrerapausoak ematen ari da sintaxiaren eta naturaltasunaren aldetik. Gaur egun, euskara ikasten duten ia guztiek euskara batua ikasten dute. Horren ondorioz, fenomeno berezi bat gertatu da Euskal Herri osoan: euskaldun zaharrak beren euskalkian mintzatzen dira herriko jendearekin eta euskara batuan euskaldun berriekin. Mendebaldean, hango euskalkiaren eta batuaren arteko desberdintasun nabariak direla eta, euskara-ikasleei iruditzen zaie ikasten ari diren hizkuntza oso urrun dagoela jendearen ahotan dabilen euskara horretatik. Bestalde, jada badaude euskara batua ama-hizkuntzatzat duten euskaldun batzuk, euskaldun berri askok beren umeei euskaraz egitea erabaki baitute, nahiz eta beren lehen hizkuntza gaztelania izan.

Hala eta guztiz ere, euskararen teorialariek [9] gero eta argiago ikusten dute, euskararen geroa bermatuko bada, beharrezkoa dela euskara batua ez ezik egungo euskalkiak ere sustatzea. Hortaz, euskalkiek izango dute nolabaiteko garrantzia euskaraz eskainiko diren HTko aplikazioetan.

Euskal hizkuntza-teknologiaren komunitatea eta ikerlariak ohartu dira zer-nolako garrantzia duten teknologiek hiztun gutxiko hizkuntza batek XXI. mendean aurrera egin dezan, eta sekulako ahaleginak egin dituzte euskara gehien erabiltzen diren hizkuntzen maila berean jartzeko teknologia aldetik. Eskarmentu zientifiko sendoa du euskarak, bai eta aldameneko zenbait hizkuntzek ere, hala nola katalanak eta galizierak; hori ez da Europan beste inon gertatu, ez eta eskualdeko hizkuntza batzuek hizkuntzarteko produktu eta zerbitzuak garatzea ere.

Argi dago zeinen garrantzitsua den euskararako hizkuntza-teknologiaren industria garatzea, eta horrexegatik sortu da *Langune* [10] elkartea ere. Langune hizkuntzen industriaren alorreko Euskal Herriko enpresen elkartea da. Elkartea 2010ean sortu zen, eta itzulpengintzaren, edukien, irakaskuntzaren eta hizkuntza-teknologiaren alorreko 30 enpresa baino gehiago biltzen ditu. Languneren helburu nagusia da hizkuntza-teknologiaren sektorea garatzea, eta erreferentzia-puntu bilakatzea Europako hizkuntzen industriarentzat, aha-

leginak biderkatu gabe eta sinergiak lortuz. Langune hasi besterik ez da egin, baina urrats ikaragarriak ari da egiten.

3.4 HIZKUNTZA-LANKETA

Euskararen ordezkari nagusia Euskaltzaindia da, euskal hizkuntzaren akademia ofiziala (1919). Hizkuntza ikertzen du, babesten saiatzen da eta erabilera-arauak ezartzen ditu. Onarpen ofizial osoa du Espainian (1976), eta onura publikorako kultura-elkartetzat onartzen da Frantzian (1995).

Euskal Autonomia Erkidegoan euskara hizkuntza ofizial deklaratu zenetik, Eusko Jaurlaritzak hainbat arau eta lege egin ditu euskararen erabilera babesteko eta bultzatzeko. Hainbat erakunde sortu dira harrezkero: Euskararen Aholku Batzordea (1982), EiTB (Euskal Irrati Telebista, 1982), HABE (Helduen Alfabetatze eta Berreuskalduntzerako Erakundea, 1983) eta beste hainbat. Euskara Biziberritzeko Plan Nagusia (EBPN) 1998an jarri zen abian, tresna estrategiko moduan eta hiru helburu nagusirekin: adostasun batera iritsi erakundeen xede eta ekintzen artean, eratze-programetarako lehentasunak ezarri eta euskararen alde lan egiten duten erakundeen, enpresen eta elkarteen jarduerak koordinatu. Plan estrategiko horren barruan, aldian behin egiten diren inkesta soziolinguistikoak baliagarriak dira beste helburu batzuk eta zuzenketa-ildoak ezartzeko.

Eusko Jaurlaritzak badu euskarari buruzko web-atari bat www.euskara.euskadi.net, eta, han, hizkuntzari, haren historiari eta gaur egungo egoerari buruzko informazioa ez ezik, hizkuntzarekin lotutako era guztietako zerbitzu, produktu eta aplikazioetarako estekak ere badaude – eratze-programa publikoetarakoak barne.

Frantziako aldean, "Euskararen Erakunde Publikoa" [11] 2004an sortu zen, interes publikoko elkarte moduan, lau erakunde publiko – herri- edo eskualde-erakundeak – eta estatua elkartuta, eta eskualdean hizkuntza-politika bateratua sortu eta ezartzeko asmoz.

3.5 HIZKUNTZA HEZKUNTZAN

Euskal Autonomia Erkidegoan, 1983an sartu zen euskara hezkuntza-sistema publikoan, Lehen eta Bigarren Hezkuntzan euskararen eta gaztelaniaren erabilera arautzen duen legearekin. Lehen eta Bigarren Hezkuntzarako, hiru eredu sortu ziren, eta ikastetxe bakoitzari aukeran eman zitzaion zein eredu eskaini.

A ereduan, komunikazio-hizkuntza gaztelania da, eta euskara "Euskal hizkuntza eta literatura" irakasgaian ematen da. D ereduan – euskaraz, c letra ez da erabiltzen, normalean –, euskara da komunikazio-hizkuntza, eta ikasgai bat ematen da gaztelaniaz, "Gaztelania eta literatura". B ereduan edo tarteko ereduan, ikasgai batzuk gaztelaniaz ematen dira (batez ere, irakurketa eta idazketa eta matematika) eta beste batzuk euskaraz (nagusiki, zientziak eta plastika).

Alabaina, A eredua gero eta ikasle gehiago galduz joan zen, eta B eredua gero eta ikasle gehiago hartuz, Haur Hezkuntzan eta Lehen Hezkuntzan batik bat – ikasleen erdiak baino gehiagok D ereduan ikasten du aldi horietan. Dena den, 15 urteko ikasleen % 85ek gaztelaniaz egin zituen PISA programaren [12] azterketak, eta % 15ek bakarrik egin zituen euskaraz; horrek argi uzten du gaztelania dela hizkuntza nagusia hezkuntzan.

Nafarroako Erkidegoan, euskarak hainbat mailatako estatus ofiziala duen tokian, laugarren eredu bat ere jarri zuten, euskara derrigorrezko ikasgai gisa ere eskaintzen ez zuena.

Iparraldeko probintziei dagokienez, euskarazko Lehen Hezkuntza eskola-sare pribatu batek ematen du, Seaskak, eta, gaur egun, 2.700 bat ikasle ditu 29 ikastetxetan – Bigarren Hezkuntzako ikastetxe bat eta lizeo bat bakarrik daude.

Azkenaldian, eredu berriak ari dira proposatzen eta probatzen, ingelesaren ikasketa goiztiarrari garrantzia ematen dioten ereduak. Eusko Jaurlaritzak eredu hirueleduna jarri du martxan duela gutxi, eta, Nafarroan, berriz, gaztelaniaz eta ingelesez ematen den hezkuntza elebi-

duna jarri dute, nahiz eta euskara aukeran eskaintzen duten. Hezkuntza-maila handiagoetan, gaztelania da nagusi, ezbairik gabe. Hiru unibertsitate daude, eta publikoa, bakarra: Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU). EHUk euskaraz ikasteko aukera ematen du, eta, euskaraz eta gaztelaniaz eskaintza bera egiteko sekulako ahaleginak egin diren arren, oso gradu gutxi ikas daitezke euskara hutsean. Nabarmentzekoa da Hizkuntzaren Azterketa eta Prozesamendua [13] master eta doktorego-programa euskara hutsean ematen dela 2001. urtetik. Mondragon unibertsitate pribatuak euskaraz eskaintzen ditu gradu gehienak, eta master batzuk ere bai. Hirugarren unibertsitateak, Deustuko Unibertsitateak, ikasketa batzuk baino ez ditu eskaintzen euskaraz.

3.6 NAZIOARTEAN

2009ko urtarriletik, Etxepare Euskal Institutua da euskal hizkuntza eta kultura mundu osoan zabaltzeaz arduratzen den euskal erakunde publikoa. Institutu horren asmoa da euskararen irakaskuntza, ikaskuntza eta erabilera bultzatzea mundu osoan, eta euskara hizkuntza komuntzat daukaten komunitate guztien ekarpenak biltzen ditu. Institutuaren xedea da, halaber, euskal kultura nazioarteko komunitatean barreiatzea, euskara hitz egiten duten talde horiei erreferentzia berezia eginez, euskal diaspora barne. Historian, euskaldun askok alde egin dute Euskal Herritik munduko beste txoko batzuetara, arrazoi ekonomiko eta politikoak direla eta; euskal diaspora izena jarri zitzaion aberritik kanpo bizi den euskal jatorriko jendeari. Gaur egun, euskal jatorriko jende dezente bizi da Txilen, Argentinan, Bolivian, Ekuadorren, Kolonbian, Kuban, Mexikon, Venezuelan, Kanadan eta Estatu Batuetan. Haietan guztietan, euskal kulturaguneak daude, Euskal Etxeak, helburu bera lortzeko sortutakoak: euskal kultura eta nortasunari eustea. 24 herrialdetako hiri handi gehientsuenetan daude Euskal Etxeak [14].

Euskararen jatorriak eta egitura bereziak euskal hizkuntza eta kultura ikasteko interesa piztu dute. Gaur egun, Amerikako eta Europako 13 herrialdetako 29 unibertsitatetan ikas daiteke.

Nazioarteko erakundeetan duen tokiari dagokionez, Espainiako gobernua saiatu da Europako erakundeen hizkuntza ofizialetan euskara sartzen, katalanarekin eta galegoarekin batera. Baina, gaur egun, ez dira hizkuntza ofizialtzat jotzen; erdiofizialak dira, eskoziera, gaelikoa eta galesarekin batera. Euskara oso leku gutxitan erabil daiteke: Eskualdeetako Lantaldearen eta Kontseiluaren saioetan hitz egin daiteke, baina Europar Legebiltzarraren osoko bilkuretan, ez. Herritarrek eskubidea dute, halaber, Europako erakundeei euskaraz idazteko eta erantzuna hizkuntza berean jasotzeko, baina Espainiako Gobernuaren bitartez egin behar dute beti, eta hark ordaindu behar ditu gastuak.

Euskara sartuta dago Europar Batasuneko Eskualdeetako Hizkuntzen eta Hizkuntza Txikien Zerrendan [15], eta, beraz, jasotzen du laguntza Europar Legebiltzarrak eskualde-hizkuntzetako eta hizkuntza txikietako ekintzak sustatzeko egindako ebazpenetatik.

> Mundu osoan zehar 29
> unibertsitatetan euskara irakasten da.

Hizkuntza-teknologiak beste ikuspegi batetik egin diezaioke aurre erronka horri, atzerriko hizkuntzako testurako itzulpen automatikoa edo hizkuntzarteko informazio-berreskurapena bezalako zerbitzuak eskainiz, eta, hala, lagundu egin dezake berezko ingeleshiztunak ez direnek dituzten desabantaila pertsonal eta ekonomikoak murrizten.

3.7 EUSKARA INTERNETEN

2010. urtearen lehen hiruhilabetekoan, Euskal Herriko etxeen % 61,4tan (513.000) ordenagailua zegoen.

460.000 familia baino gehixeago zeuden, eta horietatik % 54,9k Interneterako sarbidea zuten beren etxeetan. Horrenbestez, 15 urte edo gehiagoko milioi bat lagun baino gehiago Internet-erabiltzaileak ziren. Gehienek esan zuten egunero konektatzen zirela. % 22,9k bakarrik erabiltzen zuen euskara Interneten. Hala eta guztiz ere, euskaldunen artean Internet-erabiltzaileen talde sendo eta gogotsu bat dago. Euskarazko blogek, euskarazko Wikipediak nahiz lineako zerbitzuek eta doako softwarean oinarritutako sistema eragile eta tresnen kokapenak euskara eta euskal kultura Interneten nahiz IKTetan egotearen aldeko apustua egin dute, eta, hala, euskara zabaldu dute. Esaterako, Euskal Wikipediak 120.000 artikulu baino gehiago ditu; wikipedia guztien artean artikulu gehien dituen 36.a da. Eta ahalegin handiak egin dira software-programa [16, 17] eta baliabide arruntak [18, 19, 20, 21, 22] euskaraz eskaintzeko.

Interneteko 1.000 webgune garrantzitsuenen artean, % 0,5etan erabiltzen da euskara.

Lehen mailako domeinu berri bat erregistratu da, .eus, eta 2012. urtean jarriko da abian. Aurretiko izenemateak 193 dira jada. Proposaturiko .eus domeinua euskal hizkuntza eta kulturaren komunitatea Interneten ordezkatuko duen izena izango da. Ikur hori euskal kultura eta euskara sustatzeko tresna bihurtuko da, eta, alde horretatik, .eus domeinua mekanismo eraginkorra izango da euskara mundu osoan estandarizatzeko. .eus domeinuak, Interneteko toki birtualean, euskara modu eraginkorrean sustatzen dela ziurtatuko du, eta, era berean, nazioartearen onespena bermatuko du. Era berean, .eus domeinuak Interneten kulturaniztasuna indartu eta zabalduko du, hizkuntza- eta kultura-komunitateei beren domeinua izaten uzteak Interneten bihotz-bihotzean jartzen baitu kulturaniztasuna. Hizkuntzarekin eta kulturekin zerikusia duten domeinuak indargarri eta onuragarri dira hizkuntza- eta kultura-komunitate horientzat, baina Internetentzat berarentzat ere bai [23].

Euskal Wikipediak 123.787 artikulu ditu, eta 36. Wikipediarik handiena da, artikulu kopuruari dagokionez.

Hizkuntza-teknologiarentzat garrantzitsua da Internet gero eta indartsuagoa izatea, bi arrazoirengatik. Batetik, eskuragarri dauden hizkuntzari buruzko datu digitalizatuak iturri aberatsa dira hizkuntza naturalaren erabilera aztertzeko, informazio estatistikoa bilduz nagusiki. Bestetik, hizkuntza-teknologia erabiltzen duten era askotako aplikazio-eremuak eskaintzen ditu Internetek.

4

HIZKUNTZA-TEKNOLOGIA EUSKARARAKO

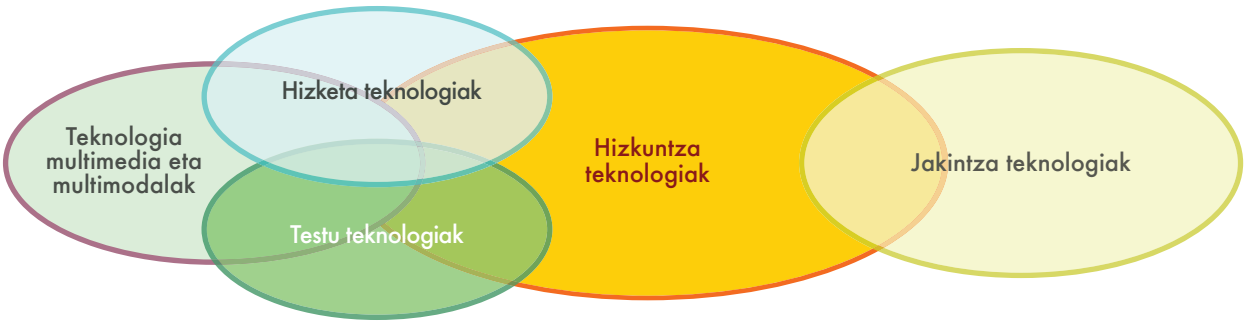
Hizkuntza-teknologiak giza hizkuntzarekin lan egiteko espezializatutako informazio-teknologiak dira. Horregatik, giza hizkuntzaren teknologia izenpean ere ezagutzen dira maiz teknologia hauek. Giza hizkuntza ahoz nahiz idatziz agertzen da. Hizkuntza-komunikazioko modurik zaharrena eta naturalena hizketa bada ere, informazio konplexua eta giza ezagutzaren zati handiena testu idatzien bidez gorde eta transmititzen da. Hizketa- eta testu-teknologiek bi modu horietan prozesatzen edo sortzen dute hizkuntza. Baina, hizkuntzak baditu hizketan nahiz testuetan agertzen diren alderdiak ere, hala nola hiztegiak, gramatikaren zati handi bat eta esaldien esanahia. Hortaz, hizkuntza-teknologiako atal asko ezin dira bietako batean sartu, hizketa-teknologian ala testu-teknologian. Horien artean daude hizkuntza ezagutzarekin lotzen duten teknologiak. 1. irudiak hizkuntza-teknologiaren egoera irudikatzen du. Elkarrekin komunikatzen garenean, beste komunikazio-modu batzuk eta beste informazio-bide batzuk erabiltzen ditugu hizkuntzarekin batera. Hizketarekin batera, imintzioak eta aurpegierak erabiltzen ditugu. Testu digitalak irudiekin eta hotsekin lagunduta joan ohi dira. Filmetan, hizkuntza ahoz eta idatziz ager daiteke. Beraz, hizketa- eta testu-teknologiak gainjarri egiten dira eta interakzioan daude askotariko komunikazioa eta multimedia dokumentuak errazago prozesatzeko aukera ematen duten beste teknologia askorekin.

4.1 HIZKUNTZA-TEKNOLOGIA APLIKATZEKO ARKITEKTURAK

Hizkuntza prozesatzeko software-aplikazio ohikoenetan, hizkuntzaren hainbat alderdi eta haien zereginak kopiatzen dituzten zenbait osagai egon ohi dira. 2. irudian, testuak prozesatzeko sistema batean topa dezakegun arkitektura bat ageri da, asko sinplifikatuta. Lehenengo hiru moduluek sarrerako testuaren egitura eta esanahia hartzen dute kontuan:

- Aurretratamendua: datuak garbitu, formatua kendu, sarrerako hizkuntza detektatu eta abar.
- Analisi gramatikala: aditza eta haren objektuak, modifikatzaileak eta abar aurkitu; esaldiaren egitura detektatu.
- Analisi semantikoa: desanbiguazioa ("hori" hitzaren zein adiera da egokia testuinguru jakin batean?), anaforak eta erreferentziako adierazpenak (adib. "bera" edo "autoa") ulertzea; esaldi baten esanahia ordenagailuak irakurtzeko moduan eman.

Zeregin espezifikoko moduluek era askotako eragiketak egiten dituzte, hala nola sarrerako testu baten laburpen automatikoa, datu-baseko bilaketak eta beste hainbat. Hemen behean, aplikazio-eremu komunak erakutsiko ditugu, eta eremu horietako modulu nagusiak nabarmendu. Hor ere, aplikazioen arkitekturak oso sinplifikatuta eta idealizatuta ageri dira, hizkuntza-teknologiako (HT) aplikazioak edonork ulertzeko mo-



1: Hizkuntza-teknologien ingurua

duan azaltzearren. Tresna eta baliabide garrantzitsuenak azpimarratuta daude testuan, eta kapitulu amaierako taulan ere ageri dira. Aplikazio-eremu komunei buruzko ataletan, dagokion euskarazko esparruan lanean ari diren industrien ikuspegi orokor bat ere ematen da.

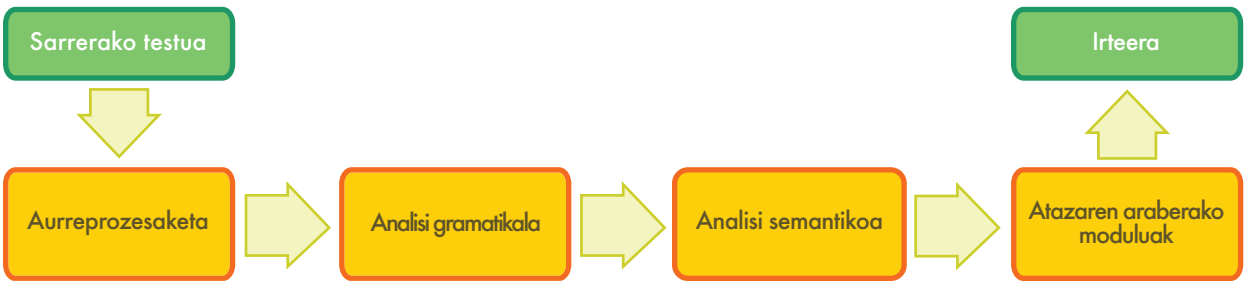
Aplikazio-eremu komunak aurkeztu ondoren, HTen ikerketa eta hezkuntzako egoeraren azalpen labur bat emango dugu, eta, bukatzeko, egin diren eta egiten ari diren ikerketa-programen berri emango dugu. Atal honen amaieran, adituaren ikuspuntutik HTen tresna eta baliabide komunen egoera nolakoa den azalduko dugu, hainbat alderditatik (eskuragarritasuna, heldutasuna eta kalitatea). Taula honek ikuspegi orokor on bat ematen du euskararako HTen egoeraren gainean.

4.2 APLIKAZIO-EREMU KOMUNAK

Atal honetan, HTen tresna eta baliabide garrantzitsuenak aztertu, eta euskarazko HTen jardueren ikuspegi bat emango dugu. Tresna eta baliabide garrantzitsuenak nabarmenduta daude testuan, eta kapitulu amaierako taulan ere ageri dira.

4.2.1 Hizkuntza-zuzentzailea

Testu-prozesadore bat erabili duen edonork egin du topo ortografia-zuzentzaile batekin – ortografia-akatsak markatu eta zuzenketak proposatzen dituen osagaia da. Ralph Gorin-ek ortografia zuzentzeko lehen programa asmatu zuenetik berrogei urte igaro ondoren, gaur egungo hizkuntza-zuzentzaileek ez dute, besterik



2: Testua prozesatzeko aplikazioen ohiko arkitektura

gabe, erauzitako hitzen zerrenda bat ortografia zuzeneko hitzen hiztegi batekin alderatzen; gero eta sofistikatuagoak dira.

Morfologiarako hizkuntzaren menpeko algoritmoez gainera (adibidez, plurala egiteko), batzuk gai dira orain sintaxiarekin lotutako akatsak identifikatzeko, hala nola aditz baten falta edo pertsona edo numeroan subjektuarekin komunztatzen ez duen aditza ("Haiek gutuna idazten ari **da*"). Alabaina, eskura dauden ortografia-zuzentzaile gehienek (Microsoft Word-ekoa barne) ez dute akatsik aurkituko Jerrold H. Zar-en poema bateko (1992) lehen bertso honetan [24]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Mota horretako akatsak aurkitzeko, testuingurua aztertu beharra dago maiz, hala nola euskaraz ergatiboa erabili behar den ala ez erabakitzeko orduan:

- *Liburua neskak dauka*
- *Irakurlea neska da.*

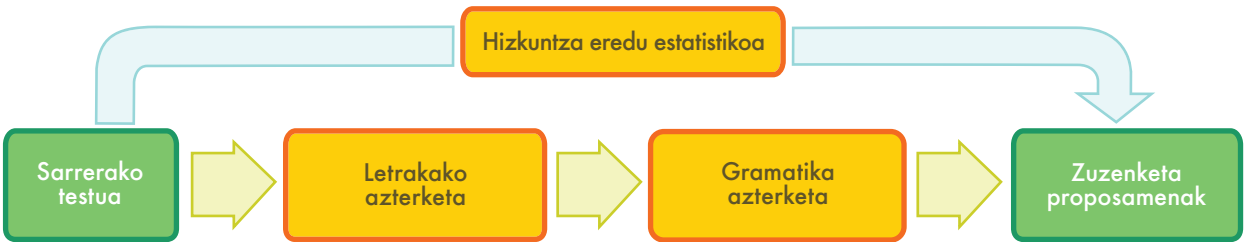
Hizkuntza zuzentzailea lortzeko (3), bietakoren bat egin behar da: hizkuntza espezifikoko **gramatika**-arauak formulatu – trebetasun eta eskulan handia eskatzen du horrek – edo hizkuntza-eredu estatistiko delakoa erabili. Ereduok hitz jakin bat inguru zehatz batean (hots, aurretik eta atzetik dituen hitzak) ateratzeko zer probabilitate dagoen kalkulatzen dute. Adibidez, *neskak dauka* hitz-sekuentzia agertzeko probabilitatea neska dauka agertzekoa baino askoz handiagoa da. Hizkuntza eredu estatistikoa automatikoki atera daiteke hizkuntza-datu zuzenen kantitate handietatik (hots, corpusetatik). Orain arte, ingelesezko hizkuntza-datuetan oinarrituta garatu eta ebaluatu dira metodo horiek. Horrek ez du esan nahi, ordea, euskarara zuzenean transferitu daitezkeenik, euskarak inflexio handiagoa eta morfologia eranskaria du eta. Egia esanda, zailtasun izugarriak daude euskararako hizkuntza-ereduak sortzeko, ezinezkoa baita balizko hitz-forma guztiak biltzea.

Hizkuntza-zuzentzailea ez da testu-prozesamenduko tresnetan bakarrik erabiltzen; testuak sortzen laguntzeko sistemetan ere erabiltzen da. Produktu teknikoek gora egin duten bezalaxe, dokumentazio teknikoa ere asko ugaritu da azken hamarkadetan. Enpresak, bezeroen aldetik erabilera txarragatiko kexak edo matxuragatiko erreklamazioak (gaizki idatziriko edo ulerturiko jarraibideengatik) jasotzeko beldurrez, dokumentazio teknikoan arreta gehiago jartzen hasi dira, eta nazioarteko merkatuan sartzen ere bai. Hizkuntza naturalaren prozesamenduan egin diren aurrerapenek testuak sortzen laguntzeko softwarea ekarri dute; programa horrek dokumentazio teknikoa jartzen du idazlearen eskura, arau jakin batzuk eta (enpresaren) terminologia-murrizketa batzuk dituzten hitzak eta esaldi-egiturak erabil ditzan.

Hizkuntza-zuzentzailea ez da testu-prozesamenduko tresnetan bakarrik erabiltzen; testuak sortzen laguntzeko sistemetan ere erabiltzen da.

Euskararako gehien erabiltzen den zuzentzaile ortografikoa Xuxen da [25], IXA unibertsitateko ikerkuntza-taldeak (http://ixa.si.ehu.es) garatu eta Eleka Ingeniaritza Linguistikoa enpresak eskaintzen duena. Zuzentzaile ortografiko hori ez da hiztegi bat erabiltzera mugatzen, ingelesean edo inflexio txikiagoko beste hizkuntza batzuetan egin ohi den moduan. Horren ordez, analisi morfologikoa egiten du. Zuzentzaile ortografiko honen bertsio berrienak gramatika eta estiloa ere zuzentzen ditu. Bertsio horretan, Hizkia [26] enpresak eta UZEI [27] erakundeak garatutako kodea ere badago.

Zuzentzaile ortografikoetan eta editatzen laguntzeko programetan ez ezik, ordenagailuz lagundutako



3: Hizkuntza azterketa (behean: arauetan oinarritua; goian: estatistikoa)

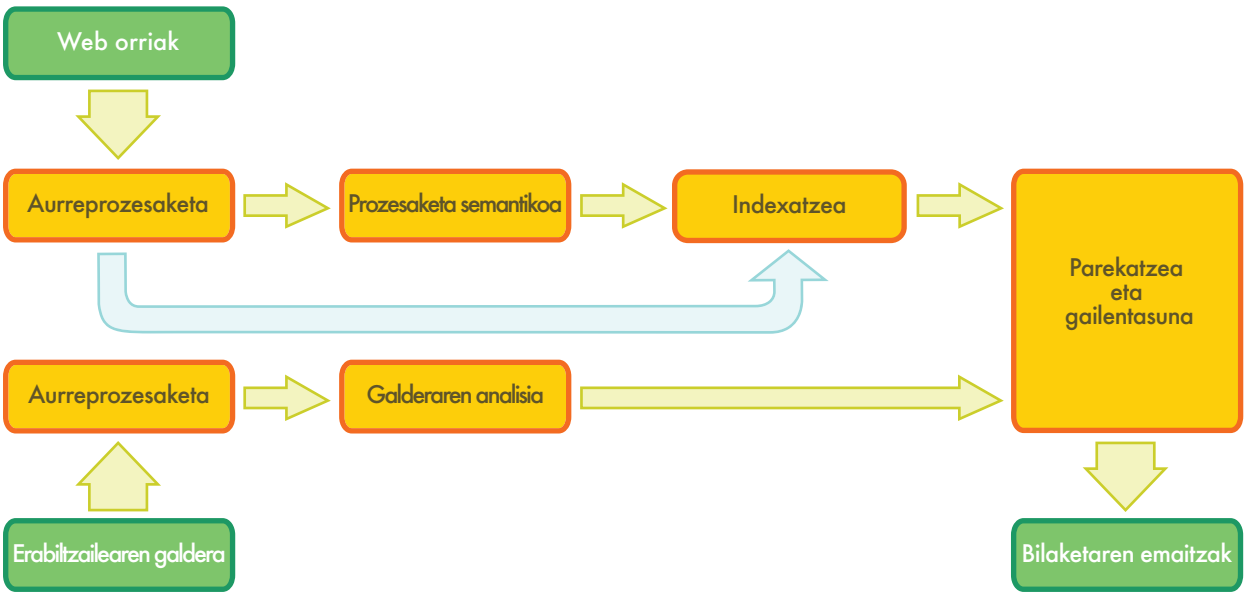
hizkuntza-ikaskuntzaren esparruan ere garrantzitsua da hizkuntza-zuzentzailea, eta web bilatzaileetara bidalitako dudak automatikoki zuzentzeko ere erabiltzen da; adib. Google-ren "esan nahi zenuen" iradokizunak.

4.2.2 Web bilaketak

Webeko, intranetetako edo liburutegi digitaletako, bilaketak dira gaur egun gehien erabiltzen den hizkuntza-teknologia, eta, hala ere, gutxi garatuta dago. Google bilatzailea 1998an sortu zen, eta mundu osoko bilaketa guztietatik % 80tan erabiltzen da gaur egun [28].
Lehenengo bertsio hartatik ez da aldaketa esanguratsurik egon, ez bilaketaren interfazean, ez berreskuratutako emaitzen aurkezpenean. Oraingo bertsioan, Google-k gaizki idatzitako hitzak ortografikoki zuzentzen ditu, eta, 2009an, oinarrizko bilaketa semantikorako ahalmenak gehitu zizkioten bere algoritmo-taldekatzeari [29]; hala, bilaketa zehatzagoa egin daiteke, sartutako terminoen esanahia testuinguruan aztertzen da eta. Google-ren arrakastak erakusten du, eskura datu asko eta horiek indexatzeko teknika eraginkorrak izanda, nagusiki estatistiketan oinarritzen den metodo batek emaitza onak eman ditzakeela.
Informazio-eskaera sofistikatuagoetarako, ezagutza linguistiko sakonagoa integratu beharra dago testu-interpretazioa errazteko. Makinek irakurtzeko moduko thesaurusak eta hizkuntza-baliabide ontologikoak -adibidez, Wordnet- eta antzeko **Baliabide lexikalak** erabili dituzten esperimentuek hobekuntzak izan dituzte orriak bilaketa-terminoen sinonimoen bidez bilatzeko aukerari esker. Aurrerapen horietarako ere hizkuntzaren baliabide espezifikoak behar dira. Euskal Herriko Unibertsitateko IXA ikerkuntza-taldeak EuskalWordNet (BasWN) garatu du, eta ELRAren bitartez eskuratu daiteke.

Bilatzaileen hurrengo belaunaldiak hizkuntza-teknologia sofistikatuagoa izan beharko du, batez ere galdera batean edo hitz-gakoz osatzen ez den esaldi batean oinarritutako kontsulta bati aurre egiteko. Esaterako, *Emadazu azken bost urteotan beste enpresa batek xurgatu dituen enpresa guztien zerrenda* kontsultarako, **analisi gramatikala** nahiz **analisi semantikoa** behar da. Erantzun egokia lortzeko, analisi sintaktikoa egin beharra dago, esaldiaren egitura gramatikala aztertzeko eta jakiteko erabiltzaileak bilatzen duena xurgatuak izan diren enpresak direla, eta ez beste enpresa batzuk xurgatu dituzten enpresak.

Azkenik, kontsulta prozesatua egituratu gabeko datu-kantitate handi batekin lotu beharra dago, erabiltzaileak lortu nahi duen informazioa aurkitze aldera. Informazio-berreskurapena esaten zaio horri, eta dokumentu egokiak bilatu eta mailakatzen ditu. Gainera, enpresa-zerrenda bat sortzeko, dokumentu bateko hitz kate bat enpresa-izen bat dela adierazten digun informazioa erauzi behar dugu. Era horretako informazioa entitateen izenen ezagutzaile izenekoen bitartez dago eskuragarri.



4: Web bilaketaren arkitektura

Are zailagoa da bilaketa bat beste hizkuntza batean idatzitako dokumentuekin lotzea. Hizkuntzarteko informazio-berreskuratzerako, balizko iturburu-hizkuntza guztietara itzuli behar dugu kontsulta automatikoki, eta berreskuratutako informazioa helburuko hizkuntzara eraman. Testu-formatuez bestelakoetan ematen diren datuen proportzioa handitu den heinean, gero eta gehiago eskatzen dira multimedia-informazioa berreskurapenerako zerbitzuak, hots, irudi-, audio- eta bideo-datuen bilaketak. Audio- eta bideo-fitxategietarako, **hizketa ezagutzeko** modulua izan behar da, hizketaren edukia testu edo irudikapen fonetiko bihurtzeko, eta, hala, erabiltzailearen kontsultak haiekin lotzeko.

Enpresa horien garapena interes bereziko atariei gehigarriak eta bilatzaile aurreratuak eskaintzean datza, gaiari dagokion semantika erabiliz. Oraindik ere prozesatze-indar handia eskatzen duela eta, bilatzaile horiek **testu-corpus** nahiko txikietan bakarrik erabil daitezke. Prozesatze-denbora bilatzaile estatistiko arrunt batena baino mila aldiz handiagoa da, gutxienez – adibidez, Google-k eskaintzen duen bilatzailearekin alderatuta. Bilatzaile horiek eskari handia dute gai espezifikoko domeinuen modelazioan, eta ezinezkoa da mekanismo horiek web-mailan erabiltzea.

> Bilatzaileen hurrengo belaunaldiak hizkuntza-teknologia sofistikatuagoa izan beharko du.

Euskal Autonomia Erkidegoan, Eleka Ingeniaritza Linguistikoa enpresa txikia buru-belarri aritu da lanean, euskararako aplikazioak eta webean oinarritutako zerbitzuak garatzen. HTren ikerketa-emaitzak eta baliabideak integratzen dituzte normalean, hala nola IXA taldearen eta Elhuyar Fundazioaren lematizatzaileak eta datu-base lexikalak. Elebila bilatzaile eleaniztunak kontuan hartzen ditu euskararen berezitasunak, eta hainbat hizkuntza-tresna eta -baliabide integratzen ditu, kalitatezko euskarazko emaitzak lortzeko. Beste adibide bat Miatu izeneko tresna da; liburutegi bat da, eta helburu bereziarekin indexatutako datu baseen gainean le-

matizatzaileak eta analisi morfologikorako beste tresna batzuk erabiliz bilaketak egiteko funtzioak eskaintzen ditu. www.zientzia.net zientziarekin lotutako ataria eta www.ikasbil.net eduki pedagogikoko ataria sortzeko erabili da.

4.2.3 Ahozko interakzioa

Erabiltzaile bati grafikoak, teklatua eta saguaren partez ahozko hizkera erabiliz makinekin interakzioan jartzeko aukera ematen dioten interfazeen oinarrian dagoen teknologia da Ahozko Interakzioa. Gaur egun, enpresek beren bezeroei, langileei edo lankideei telefonoz eskaintzen dizkieten automatizazio partzial edo osoko zerbitzuetan erabiltzen dira, normalean, ahotsaren bidezko erabiltzaile-interfazeak (AEI). Lan-arlo hauetan asko erabiltzen dira ahotsaren bidezko erabiltzaile-interfazeak: banketxeak, logistika, garraio publikoa eta telekomunikazioak. Ahozko interakzioaren teknologia gailu jakin batzuen interfazeetan ere egon ohi da – esaterako, autoan txertatuta doazen nabigazio-sistemetan –, eta erabiltzaile-interfaze grafikoen sarrera/irteera modalitateen ordez ahozko hizkera ezartzeko ere erabiltzen dira, Smartphone gailuetan esaterako.

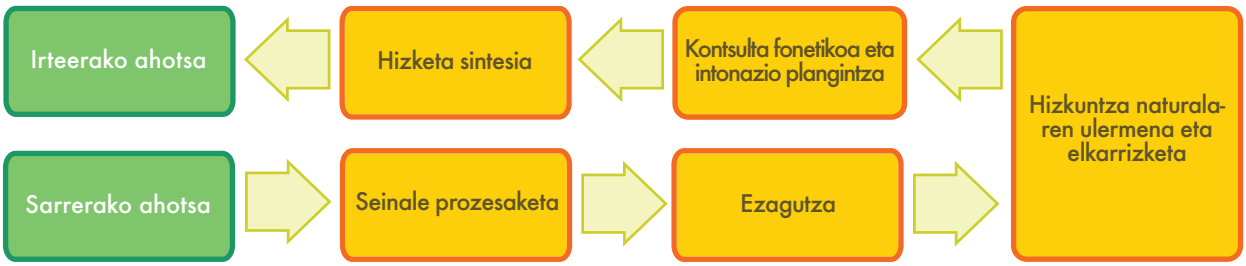
Ahozko interakzioaren oinarrian, lau teknologia hauek daude:

- **Hizketaren ezagutza** automatikoak: zer hitz esan diren ateratzen du, erabiltzaileak egindako hots-sekuentzia batetik.
- Analisi sintaktikoak eta interpretazio semantikoak: erabiltzaile batek esandakoaren egitura sintaktikoa aztertzen du eta interpretatu egiten du sistemaren helburuaren arabera.
- Elkarrizketa-kudeaketa: beharrezkoa da erabiltzailearen interakzioko sistema-zatian zer egin zehazteko, behin erabiltzailearen inputa emandakoan eta sistemaren funtzioak kontuan hartuta.
- **Hizketaren sintesirako** teknologia (TTS edo Text-to-Speech): esandako hitz horiek hots bihurtzeko eta erabiltzailearentzako irteera gisa emateko erabiltzen da.

> Erabiltzaile bati ahozko hizkera erabiliz makinekin interakzioan jartzeko aukera ematen dioten interfazeen oinarrian dagoen teknologia da Ahozko Interakzioa.

ASR sistemen erronkarik handienetarikoa da zehaztasun osoz ezagutzea erabiltzaile batek esandako hitzak. Horretarako, bietakoren bat egin behar da: erabiltzailearen balizko esaldiak gako-hitzen sorta mugatu batera murriztu, edo hizkuntza-ereduak sortu eskuz, hizkuntza naturalaren erabiltzailearen esaldi sorta handia hartzen dutenak. Ikasketa automatikoko sistemak erabiliz, automatikoki ere sor daitezke hizkuntza-ereduak **hizketa-corpusetatik**; alegia, hizketa duten audio-fitxategien eta haien testu-transkripzioen bilduma handietatik. Esaldien edukia mugatuz gero, ahots bidezko interfazeak era mugatuan erabiltzera behartzen da erabiltzailea, eta horrek eragina izan dezake haren erosotasunean; hala ere, hizkuntza-eredu aberatsak sortu, doitu eta zaintzeak asko igo ditzake gastuak. Hizkuntza-ereduak erabiltzen dituzten eta hasieran erabiltzaile bati zer nahi duen adierazteko malgutasuna ematen dioten –*Nola lagundu diezazuket?* edo antzeko galderen bidez– ahots bidezko erabiltzaile-interfazeek onarpen zabalagoa dute.

Ahots bidezko erabiltzaile-interfazeen emaitzetarako, enpresek erabiltzen dituzten esaldiak aldez aurretik grabatutakoak izan ohi dira eta hiztun profesionalek esandakoak – ahal izanez gero, enpresakoak bertakoak. Esaldi estatikoak diren kasuan, hitzen formulazioa erabilera-testuinguru jakin baten edo erabiltzaile horren datu pertsonalen menpekoa ez denean, erabiltzailearen esperientzia ona izango da. Aldiz, esaldi batek zenbat eta eduki dinamikoagoa hartu behar duen kontuan, are txarragoa izango da erabiltzailearen esperientzia, audio-fitxategi solteak elkarrekin lotzeagatik sortutako prosodia kaskarragoa izango baita. Gaur egungo



5: Ahots bidezko elkarrizketa sinple baten arkitektura

TTS sistemak, ostera, hobeak dira esaldi dinamikoen naturaltasun prosodikoari dagokionez, nahiz eta oraindik ere hobetu daitezkeen.

Ahozko interakzioaren teknologiaren merkatuan, azken hamarkadan, teknologia-osagaien arteko interfazeak asko estandarizatu ziren, eta aplikazio jakin baterako software-tresna jakin batzuk sortzeko irizpideak ere agertu ziren. Era berean, merkatua asko indartu da azken hamar urteotan, batez ere hizketaren ezagutza automatikoaren eta TTSen esparruetan. Esparru horietan, G20 herrialdeetako – populazio dezentea eta indar ekonomiko handia duten herrialdeak – merkatu nazionalak mundu osoko bost enpresa baino gutxiagoren eskuetan daude; Europan, Nuance eta Loquendo dira garrantzitsuenak. 2007. urtetik, Eusko Jaurlaritzak emandako babesari esker, Nuanceren produktu-katalogoan sartuta dago euskara. Alabaina, hizketaren ezagutza automatikoaren eskaintza tamaina txiki eta ertaineko hiztegi-aplikazioetara mugatzen da, eta ez da eskaintzen diktaketa-produkturik. TTSrako, emakumezko ahots bakarra eskaintzen da. Espainiako merkatuan, Verbio Speech Technologies [30] enpresa kataluniarrak bietarako eskaintzen du euskara, hizketaren ezagutza automatikorako nahiz TTSrako. Euskararako diktaketa-sistema komertzialik ez dago, ordea.

Elkarrizketa-kudeaketako teknologia eta ezagutzei dagokienez, enpresa nazionalak dira nagusi merkatuetan, ETEak normalean. TTSen Espainiako merkatuan, enpresa gehienak aplikazio sortzaileak dira. Espainiako merkatuko enpresa nagusiak hauek dira: Indsys [31] (Intelligent Dialogue Systems), Fonetic [32], Ydilo [33] eta NaturalVox [34]. Horietako zenbaitek badute eskaintza mugatu bat euskararentzat. Euskararako doaneko TTS softwarea ere badago, Euskal Herriko Unibertsitateko (UPV/EHU) Aholab [35] ikerkuntza-taldeak eskainia.

Gaur egungo teknologiatik harago begiratuta, aldaketa esanguratsuak egongo dira, Smartphone gailuak hedatu egingo baitira bezeroekiko harremanak kudeatzeko plataforma berri moduan – telefono, Internet eta posta elektronikoarekin batera. Joera horrek eragina izango du ahozko interakziorako teknologiaren erabileran ere. Alde batetik, epe luzera, behera egingo du telefonian oinarritutako ahotsaren bidezko erabiltzaile-interfazeen eskariak. Bestetik, gero eta gehiago erabiliko da ahozko hizkera Smartphonetarako sarrera-modalitate erabilerraz moduan. Joera hori erakusten dute hiztuna kontuan hartu gabeko hizketaren ezagutzaren zehaztasunean egin diren hobekuntza nabariek – Smartphone-erabiltzaileei zerbitzu zentralizatu moduan jada eskaintzen ari zaizkien ahozko diktaketa-zerbitzuetarako egindakoak. Ezagutzaren eginkizuna aplikazioen azpiegitura bideratzeko joera hori ikusita, uste da hizkuntza-teknologia komunen aplikazio espezifikoko erabilerak garrantzia hartuko duela.

4.2.4 Itzulpen automatikoa

Hizkuntza naturala itzultzeko ordenagailu digitalak erabiltzearen ideia A. D. Booth-ek izan zuen 1946an, eta esparru hori ikertzeko finantzaketa handia egin zen 1950eko hamarkadan, eta, berriro, 1980eko hamarkadan. Hala ere, **itzulpen automatikoak** (IA) ez dio oraindik behar bezala erantzun hasierako urteetan sortu zuen itxaropenari.

> Itzulpen Automatikoak, bere oinarrizko mailan, hizkuntza natural bateko hitzak kendu eta beste batekoak jarri besterik ez du egiten.

IAk, bere oinarrizko mailan, hizkuntza natural bateko hitzak kendu eta beste batekoak jarri besterik ez du egiten. Hori baliagarria izan daiteke esapide gutxiko hizkera oso mugatua darabilten esparruetan, hala nola eguraldi-iragarpenetan. Estandarizazio gutxiagoko testuak ondo itzultzeko, ordea, testu-unitate handiagoak (esapideak, esaldiak eta pasarte osoak ere bai) xedehizkuntzako baliokide aproposenekin lotu behar dira. Horko zailtasunik handiena giza hizkuntzaren anbiguotasuna da, erronkak ezartzen baititu hainbat mailatan; esaterako, adiera desanbiguazioa lexiko-mailan ("Jaguar" hitzak autoari edo animaliari egin diezaioke erreferentzia) edo beste maila batzuetan, adibidez:

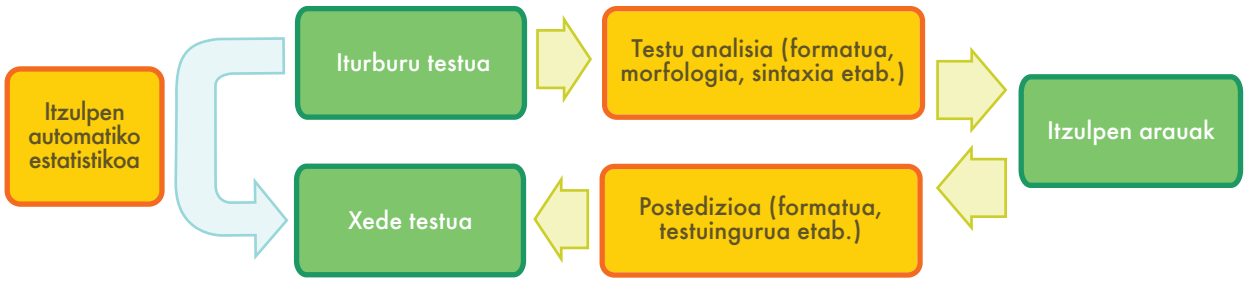
- *Egon garenetan ez dugu topatu*
  [Egon garen aldietan ez dugu topatu] edo
  [Egon garen tokietan ez dugu topatu]
- *Aitak semeari bere bizikleta eman dio*
  [Aitarena ala semearena?]

Halakoak konpontzeko modu bat hizkuntza-arauetan oinarritzen da. Familia bereko hizkuntzekin ari bagara lanean, beharbada zuzeneko itzulpena egin daiteke bigarren adibidearen gisako perpausetan. Baina, sarritan, arauan oinarritutako sistemek (edo ezagutzak gidaturikoek) sarrera-testua aztertu eta tarteko adierazpide sinboliko bat sortzen dute, eta hortik sortzen da xedehizkuntzako testua. Metodo horiek arrakasta izan dezaten, hiztegi handiak izan behar dituzte, informazio morfologiko, sintaktiko eta semantikoa biltzen dutenak, eta gramatika-arauen bilduma handia, hizkuntzalari aditu batek tentuz diseinaturikoa.

1980ko hamarkadaren amaieratik hasita, ordenagailua garrantzia hartuz eta merkatuz joan zen heinean, gero eta jakin-min handiagoa pizten zuten IArako eredu estatistikoek. Eredu estatistiko horien parametroak testu elebidunen corpusen analisitik atera dira; hor dugu, esaterako, Europarl **corpus paraleloa**, Europako Legebiltzarraren aktak 11 hizkuntza europarretan ematen dituena. Datu nahikoa izanez gero, IA estatistikoa nahiko baliagarria da atzerriko hizkuntzan idatzitako testu baten gutxi gorabeherako esanahia ateratzeko. Alabaina, ezagutzak gidaturiko sistemekin alderatuta IA estatistikoak (edo datuek gidaturikoak) duen desabantaila da emaitza agramatikalak sortzen dituela maiz. Bestalde, datuek gidaturiko IAk, gramatika idazteko giza ahalegin txikiagoa behar izateaz gainera, badu beste abantaila bat: ezagutzak gidaturiko sistemei ihes egiten dizkieten berezitasunak ondo trata ditzake, esapideak kasurako.

Ezagutzak gidaturiko IAren eta datuek gidaturiko IAren indarguneak eta ahulguneak elkarrekiko osagarriak direnez, bietako metodologiak nahasten dituzten metodo hibridoetara jotzen dute ikerlari guztiek gaur egun. Hori egiteko era bat baino gehiago daude. Batean, bi sistemak erabiltzen dira – ezagutzak gidaturikoa eta datuek gidaturikoa –, eta hautapen-modulu batek erabakitzen du zein den esaldi bakoitzaren irteerako esaldi onena. Esaldi luzeetarako, ordea, ez du topatzen emaitza egokirik. Konponbide hobea da irteera askotako esaldien zatirik onenak elkartzea; hori nahiko zaila izan daiteke, askotariko aukerei dagozkien atalak ez baitira beti agerikoak eta lerrokatu egin behar izaten baitira.

Euskararentzat, IA bereziki zaila da. Euskara hizkuntza eranskaria, morfologia aberatsekoa eta flexio maila han-



6: Itzulpen automatikoa (ezkerrean: estatistikoa, eskuinean: arau bidezkoa)

dikoa izanda, hiztegiaren analisia eta hiztegi-estaldura zailagoa da. Gainera, esaldiko osagaien hurrenkera dela bide, lan nekeza da **corpus paraleloak** kudeatzea.

Matxin transferentzian oinarritutako IA sistema bat da, gaztelaniatik euskararakoa, IXA taldeak Euskal Herriko Unibertsitatean garatutakoa. Irekia da, berrerabilgarria, eta euskarri elkarreragingarria eskaintzen du beste hizkuntza bikoteentzat ere (matxin.sourceforge.org). Kode irekiko beste programa batzuk erabiltzen ditu, hala nola Freeling, eta euskararen morfologia berrerabiltzen du morfologia sorkuntzarako. IXA taldeak Itzulpen Automatiko Estatistikoko sistema bat ere sortu du euskara eta gaztelaniarako, hitzen segmentazioa eta berrordenaketa egiten duena (EUSMT). IA sistema horiek garatzeko, lankidetza estua dago unibertsitateko ikerkuntza-taldearen, Eleka Ingeniaritza Linguistikoa enpresa txikiaren eta Elhuyar Fundazioaren artean – azken horrek hizkuntza-baliabide asko jartzen ditu. Eleka enpresak Batua-Bizkaiera bihurtzailea ere atera du. Alacanteko Unibertsitateko Transducens taldeak ere garatu du euskaratik gaztelaniara itzultzeko hasierako sistema bat, Apertium plataforma erabiltzen duena. Google Itzultzaileak alfa bertsio bat eskaintzen du euskararentzat.

Lucy Software enpresak – IAko aplikazioen sortzaile garrantzitsuenetakoa da nazioartean – filial bat dauka Espainian, Lucy Iberica [36], lehen Translendium zena. Eusko Jaurlaritzak enpresa hori hautatu zuen 2008an, gaztelaniatik euskararako itzulpen-sistema bat sortzeko, eta 2011an berriro hautatu zuen lan horrekin jarrai zezan.

Erabiltzailearekiko espezifikoa den terminologia eta lanprozesuen integrazioa behar bezala egokituz gero, oro har, uste da IAren erabilerak produktibitatea asko hobetu dezakeela. Halaber, IA sistemen kalitatea oraindik asko hobetu daitekeela uste da. Hainbat erronka daude oraindik; besteak beste, hizkuntza-baliabideak esparru edo erabiltzaile-talde jakin batera egokitzea eta lehendik dauden prozesuetan integratzea, terminoen datu-baseekin eta itzulpen-memoriekin batera. Gainera, hizkuntza-bikote asko falta dira oraindik.

Ebaluazio-kanpainek IA sistemen kalitatea, metodoak eta hizkuntza pare bakoitzerako sistemaren egoera alderatzeko balio dute. 7. irudia (p. 24) Euromatrix+ proiektuan prestatu zen, eta Europako 23 hizkuntzatatik 22tarako lortutako binakako emaitzak (irlandera ez dago ebaluatua) erakusten ditu. Emaitzak zerrendatzeko, BLEU score bat hartu zen kontuan, non score handiagoak itzulpen hobea adierazten baitu [37]. Giza itzultzaile batek 80 puntu inguru lortuko lituzke. Emaitzarik onenak (berdez eta urdinez) programa koordinatuetan ikerketa-inbertsio garrantzitsuak izan dituzten eta hainbat corpus paralelo dituzten hizkuntzek dituzte (ingelesak, frantsesak, nederlanderak, espainierak eta alemanak, esaterako). Emaitza txarragoak dituzten hizkuntzak gorriz ageri dira. Hizkuntza horien kasuan,

	Xede hizkuntza — Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

7: 22 hizkuntza europar arteko itzulpen automatikoa – Machine translation between 22 EU-languages [38]

edo ez dute behar bezalako garapenik izan, edo oso desberdinak dira egitura aldetik beste hizkuntzekiko (adibidez, hungariera, maltera eta finlandiera).

4.3 BESTE ERABILERA-EREMU BATZUK

Hizkuntza-teknologiako aplikazioak sortzeko, era askotako lanak egin behar dira, eta, batzuetan, lan horiek ez dira agertzen erabiltzailearekiko elkarreraginean, baina zerbitzu-funtzio garrantzitsuak betetzen dituzte sistemaren "erraietan". Horregatik, ikergai garrantzitsuak dira, eta esparru akademikoko hizkuntzalaritza konputazionalaren barruko banakako diziplinak dira orain.

Galderei erantzutea ikerketa-arlo bizia da orain; corpus etiketatuak egin dira horretarako eta lehiaketa zientifikoak antolatu dira. Ideia da gako-hitzetan oinarritutako bilaketa atzean uztea (bilatzaileak garrantzitsuak izan daitezkeen dokumentuen sorta handi bat ematen du), eta bilaketa zehatza egin ahal izatea (erabiltzaileak galdera zehatz bat egiten du, eta sistemak erantzun bakar bat ematen dio):

> *Galdera: Zer adin zuen Neil Amstrongek ilargira iritsi zenean?*
>
> *Erantzuna: 38.*

Argi dago hori lehen aipatu dugun web-bilaketaren esparru komunarekin dagoela lotuta, baina, galderei erantzutea, gaur egun termino zabala da eta hainbat zalantza sortzen ditu, hala nola "Zer galdera mota ezberdindu behar lirateke eta nola tratatu behar lirateke", "Nola azter eta erka daitezke erantzuna izan dezaketen dokumentuak" (bat ez datozen erantzunak eman ditzakete?) eta "Nola erauz daiteke informazio espezifikoa (erantzuna) testu batetik testuingurua gehiegi baztertu gabe?".

Horrek lotura du informazio-erauzketa (IE) delako lanarekin ere; garrantzi eta eragin handiko alorra izan zen hori 1990eko hamarkadaren hasieran, hizkuntzalaritza konputazionalaren barruan "aldaketa estatistikoa" gertatu zen garaian. IEren xedea da dokumentu mota espezifikoetan informazio espezifikoa aurkitzea; adibidez, egunkarietako pasarteetan kontatzen diren enpresa-xurgatzeetan partaide nagusiak zein diren aurkitzea. Landutako beste alor bat terroristen erasoei buruzko albisteena da; horietan, erasoaren egilea, helburua, ordua, tokia eta ondorioak jasotzen dituen txantiloi batera pasatzea da arazoa. Esparru espezifikoko txantiloiak betetzeko gaitasuna IEren bereizgarri nagusia da, eta, horregatik, "atzeko" teknologiaren beste adibide bat da; ondo zedarritutako ikerketa-eremuak ditu, baina, erabilgarria izan dadin, aplikazio-inguru egoki batean txertatu behar da.

Hizkuntza-teknologiako aplikazioek zerbitzu-funtzio garrantzitsuak betetzen dituzte software-sistema handiagoen baitan.

Bi "mugako" eremu daude: testu-laburpenak eta **testu-sorkuntza**; batzuetan, aplikazio autonomoak dira, eta, beste batzuetan, azpiko aplikazio laguntzaileak. Testu-laburpenak, agerikoa den moduan, testu luze bat laburtzeko lanari egiten dio erreferentzia, eta MS Word-en barruko funtzio moduan eskaintzen da, esaterako. Batez ere estatistiketan oinarrituta egiten du lan; lehenik, testu bateko hitz "garrantzitsuak" aurkitzen ditu (adibidez, testuan oso maiz eta hizkuntzaren erabilera orokorrean askoz gutxiagotan agertzen diren hitzak), eta, gero, hitz garrantzitsu asko dituzten esaldiak detektatzen ditu. Esaldi horiek dokumentuan markatu edo handik erauzi egiten ditu, eta laburpena egiteko erabiltzen ditu. Zeregin horretan – eta horixe da bere zeregin nagusia –, testu-laburpena eta esaldi-erauzketa gauza bera dira: testua txikiagotu egiten da bere esaldien azpimultzo batera. Testu-laburpenerako tresna komertzial guztiek ideia horixe erabiltzen dute. Zertxobait ikertu den beste metodo bat laburpenean esaldi berriak sartzea da; hots, jatorrizko testuan forma horretan agertu beharrik ez duten esaldiez osaturiko laburpena egitea. Horretarako, testua sakonagotik ulertu behar da, eta, beraz, ez da horren metodo sendoa. Azken batean, testu-sortzaile bat normalean ez da izaten aplikazio autonomo bat, software handiagoko batean txertatutako aplikazioa baizik; adibidez, informazio klinikoaren sistemetan, pazienteen datuak bildu, gorde eta prozesatu egiten dira, eta txostena sortzea da sisteman txertatutako testu-sortzailearen funtzioetako bat.

Euskararako eta, halaber, hizkuntza gehienetarako, testu-teknologia gehienetako ikerketa ez dago ingeleserako bezain garatua.

Ikerkuntza-eremu horiek guztiak ez daude ingeleserako bezainbeste garatuta euskararako. Bada, ingelesean, hainbat eta hainbat lehiaketa ireki egin dira galderei erantzutearen, informazio-erauzketaren eta testu-laburpenen esparruetan, Estatu Batuetako DARPA eta NISTek antolatuta. Horrek aurrerapen handiak ekarri ditu esparruan, baina jomuga ingelesa izan da beti; lehiaketa batzuetan sartu izan dira hainbat hizkuntza, baina euskara ez da inoiz horietako bat izan. Hori dela eta, corpus etiketatu edo baliabide gutxi daude eskuragarri zeregin horietarako. Testu-laburpenetarako sistemak, metodo estatistikoak bakarrik erabiltzen dituztenean, hizkuntzaren mendekoak dira, neurri handi batean, eta, beraz, zenbait ikerketa-prototipo daude eskuragarri. Testu-sorkuntzarako, osagai berrerabilgarriak gainazala egiteko moduluetara ("sorkuntza-gramatiketara") mugatuta egon izan dira; hor ere, eskura dauden software gehienak ingeleserako dira.

4.4 HIZKUNTZA-TEKNOLOGIA HEZKUNTZAN

Hizkuntza-teknologia diziplinarteko alorra da, eta hainbat adituren lana hartzen du; besteak beste, hizkuntzalariak, informatikariak, matematikariak, filosofoak, psikolinguistak eta neurozientzialariak. Hortaz, gaur egun, Espainian, hizkuntzalari konputazional izateko oinarrizko prestakuntza filologia edo hizkuntzalaritzako gradu baten barruan emango da, beharbada – irakasgai komunen artean hizkuntzalaritza konputazionala ematen bada –, edo informatika-fakultatean, bestela. Lehen aukera eskaintzen duten unibertsitateak hauek dira: Universitat de Barcelona, Universitat Pompeu Fabra, Universitat Oberta de Catalunya eta Universidade de Vigo. Hizkuntzalaritza konputazionala irakasgaitzat ematen duten informatika-fakultate nagusiak, berriz, beste hauek dira: Universidad Politécnica de Madrid, Universidad Carlos III, Universidad Autónoma de Madrid, Universitat d'Alacant, Universidad Nacional de Educación a Distancia eta Euskal Herriko Unibertsitatea. Bi aukerak eskaintzen dituenik ere bada: Universidad Complutense.

Graduondoko ikastaroek helburu zehatzagoa duen prestakuntza profesionala eskaintzen dute. Doktoreikastaro batzuetan ematen dira hizkuntzaren eta hizketaren prozesamenduarekin lotutako masterrak eta irakasgaiak. Euskal Herriko Unibertsitateak hizkuntzaren prozesamenduari buruzko doktoretza-ikastaro oso bat eskaintzen du, euskara hutsean ere ikas daitekeena. Beste master edo doktoretza-ikastaro batzuetako ikasleei ere eskaintzen zaizkie hizkuntza-teknologiako moduluak, hizketaren prozesamendua batik bat (adib. EHUren Sare Mugikorretako Informazio- eta Komunikazio-teknologiak masterreko ikasleei).

Euskal Autonomia Erkidegoko hiru unibertsitateetan banatuta dauden zenbait ikerketa-talde ari dira gai hauek lantzen: hizketaren prozesamendua, hizketaren sintesia eta bihurketa, hizketaren eta hiztunaren ezagutza, hizkuntzaren ezagutza, hizkuntza naturalaren prozesamendua, testutik testurako itzulpena eta hizketatik hizketarako itzulpena. Guztiak dira Hizkuntza Naturalaren Prozesamendurako Espainiar Elkarteko kide (SEPLN, Sociedad Española para el Procesamiento del Lenguaje Natural). SEPLN irabazi asmorik gabeko erakundea da, esparru akademikoko nahiz industria-arloko 300 kide baino gehiago dituena, eta 1984an sortu zen, helburu honekin: irakaskuntza, ikerkuntza eta LNPren garpenarekin lotutako jarduerak sustatu eta zabaltzea, Espainian bertan nahiz nazioartean. SEPLN elkarteak mintegiak, sinposioak eta hitzaldiak antolatzen ditu, eta Espainiako nahiz nazioarteko erakundeekiko lankidetza sustatzen du.

SEPLNk urteroko kongresu bat antolatzen du, eta, urtetik urtera, LNP lantzen duten ikerlari gehiago erakartzen ditu, Espainiatik nahiz kanpotik datozenak. Elkarteak aldizkari bat ere kaleratzen du, eta web-zerbitzari bat du, hizkuntza naturalaren prozesamenduarekin lotutako gaiei buruzko informazioa eta kideentzako foro ireki bat eskaintzen dituena.

Espainiako Hizketa Teknologietako Sarea (RTTH) [39] foro komun bat da, eta han elkartzen diren hizketateknologien ikerlariek (25o baino gehiago dira gaur egun) zeregin osagarriak egin eta esperientziak elkarbanatzen dituzte, helburu hauekin:

- Hizketa-teknologien ikerkuntza sustatzea, alor horretara ikerlari gazte gehiago erakartzeko prestakuntza, ikasle-trukeak, bekak eta sariak eskainiz.
- Negozio-ikerkuntzarako inbertsioak erakartzea, beste negozio aukera batzuk eskaintzen dituzten aplikazio berriak aurkituz.
- Aurrerapenak egitea elkarteak sortzen, eta sareko kideak integratzea, Espainiak gaztelaniaren ikerkuntzan duen nagusitasunari eusteko eta hizkuntza koofizialei ere bultzada emateko (katalana, euskara eta galegoa).

RTTH elkarteak "Hizketaren teknologiari buruzko jardunaldiak" antolatzen ditu urtero, 2000. urtetik. Ikastaro horren xedea da elkargune bat izatea, iberiar penintsulako hizkuntzetan dauden hizketa- eta hizkuntza-teknologien gaineko ikerkuntzaren emaitzen berri emateko eta horiei buruzko eztabaida sustatzeko. Enpresen eta unibertsitatearen arteko elkarlana ere sustatzen du. Era askotako jarduerak antolatzen ditu: txosten teknikoen aurkezpenak, hitzaldi magistralak, proiektu-txostenen eta laborategiko jardueren aurkezpenak, erakustaldiak eta azken doktore-tesien aurkezpenak.

4.5 HIZKUNTZA-TEKNOLOGIAKO PROGRAMAK

Euskararako teknologia-programak Eusko Jaurlaritzak eta Espainiako Gobernuak bultzatu dituzte, batez ere.

Espainiako Hezkuntza Ministerioak eta Zientzia eta Berrikuntzako Ministerioak ikerketa-programa nazionalen bidez bultzatu dute informazio-teknologien alorreko ikerketa. Programa horietarako, ikerketa-proiektu asko egin behar izan dira, eta elkarlana egin nazioarteko ikerketa zentro eta enpresekin. Euskararen prozesamendu automatikorako aplikazio komertzialen eta aurrerapen teknologikoaren oinarria proiektu horien ondorioz sortu da, hein batean.

2000. urtetik hona, Espainiako Gobernuak, Ikerketarako eta Teknologia Berrikuntzarako Plan Nazionalaren barruan, proiektu batzuk bultzatu ditu hizketa-teknologia eleaniztunen alorrean: TEHAM, AVIVAVOZ eta BUCEADOR. Haien helburu nagusia zen hizketaren ezagutza, hizketaren itzulpena eta testutik hizketarako sintesia hobetzea Espainiako hizkuntza ofizial guztietan: euskara, galegoa, katalana eta gaztelania.

Industria Teknologia Garatzeko Zentroa (CDTI) Espainiako erakunde publiko bat da, Zientzia eta Berrikuntza Ministerioaren mendekoa, eta Espainiako enpresen maila teknologikoa areagotzen laguntzea du helburu. CDTIk I+G proiektuak ebaluatu eta finantzatzen ditu, CENIT (2010ean bukatutakoa) eta AVANZA bezalako programen bitartez.

Eusko Jaurlaritzak ikerketa eta berrikuntzaren alde egiten du "Zientzia, teknologia eta berrikuntzarako plana" ren bitartez (ZTBP). Plan horren barruan, erakunde eta ikerketa- eta berrikuntza-agentzia batzuk eratu dira azken urteotan: Zientzia, Teknologia eta Berrikuntzaren Euskal Kontseilua (ikerketa eta berrikuntza bultzatzeko eta garatzeko ekintzak egiten dituen erakunde politiko ahaltsuena), InnoBasque (berrikuntzaren euskal agentzia) eta IkerBasque (zientziaren euskal fundazioa, talentudun ikerlariak euskal zientzia- eta teknologia-sistemara erakartzeaz arduratzen dena). ZTBP planaren tresna garrantzitsuenak ikerketa- eta berrikuntza-proiektuetarako deialdiak dira: ETORTEK programa – Zientzia, Teknologia eta Berrikuntzarako Euskal Sareko erakundeei zuzendutakoa – eta ETORGAI programa – enpresa pribatuei zuzendutakoa.

Azken ZTBP planean (2010ekoan) identifikatutako alor estrategiko bat hizkuntza-teknologiena izan da, aurrekoetan bezalaxe. Hala, azken hamar urteotan, HIZKING21, ANHITZ eta BERBATEK [40] proiektuak gauzatu dira ETORTEK programaren barruan. Euskararako gaur egun dauden baliabide eta tresna gehienak proiektu horien bidez eskuratutakoak dira.

4.6 EUSKARARAKO TRESNA ETA BALIABIDEAK

Atal honetako 8. taulan, euskararako dauden Hizkuntza-teknologien gaur egungo egoeraren laburpena ageri da. Adostutako estimazioetan oinarrituta, dauden tresna eta baliabideak balioetsi dituzte zenbait adituk, zazpi irizpideei jarraiki (0tik 6ra):

Liburu Zurien bilduma honetan, Europako hainbat hizkuntzaren egoera orokorra balioesteko lehendabiziko ahalegina egin da, hizkuntza-teknologien egoerari dago-

	Kantitatea	Eskuragarritasuna	Kalitatea	Estaldura	Heldutasuna	Iraunkortasuna	Moldagarritasuna
Hizkuntza teknologiak (Tresna, Teknologiak eta Aplikazioak)							
Hizketa Ezagutza	2	1	1	1	4	3	2
Hizketa Sintesia	2	3	4	4	4	3	3
Analisi gramatikala	4	2.5	4	4	4	2.5	2.5
Analisi semantikoa	1	1.5	2	1	1	1	1
Testu-sorkuntza	1	0	0	0	0	0	0
Itzulpen automatikoa	3	5	2	3	3	2	2
Hizkuntza baliabideak (Baliabideak, Datuak eta Jakintza-Baseak)							
Testu-corpusak	2	4	3	2	3	4	2.5
Hizketa-corpusak	3	2	3	2	3	3	2
Corpus paraleloak	2	4	2	2	2	2	1
Baliabide lexikalak	4	4	4	5	5	4	3
Gramatikak	2	2	2	2	2	2	2

8: Hizkuntza-teknologien sustapenaren egoera euskararako.

kienez. Hutsuneak eta beharrak zehatz-mehatz konparatzeko eta identifikatzeko aukera emango du azterketa horrek.

Euskararako, teknologiei eta baliabideei erreparatuz ateratako ondorioak honako hauek dira:

- Gaur egun, hizketaren prozesamendurako tresnek argi adierazten dute hizketaren sintesia garatuago dagoela hizketa-ezagutza baino. Hala ere, oso zaila da euskararako eguneroko aplikazioak aurkitzea; hala nola, telefono mugikorretarako ahots bidezko interfazeak, auto-nabigazio sistemak edo ahots bidezko elkarrizketa-sistemak.
- Ikerketa-talde batzuek hizketaren eta hizkuntzaren prozesamenduan dihardute lanean. Hala ere, ikerketako ahaleginak eta norabideak ez daude koordinatuta, eta tokiko eta aldizkako finantzaketaren menpe daude.
- Euskararako HT ikerketak arrakasta lortu du kalitate handiko tresna jakin batzuk diseinatzerakoan, baina zaila da ebazpen jasangarriak eta estandarrak proposatzea. Era berean, baliabide asko ez dira estandarrak, hau da, existitzen badira ere jasangarritasuna ez dago bermatua; programa eta ekimen itunduak behar dira datuak eta truketarako formatuak estandarrak bihurtzeko.
- Semantika sintaxia baino zailagoa da prozesatzeko; testu-semantika hitz- eta perpaus-semantika baino zailagoa da prozesatzeko. Tresna batek gero eta semantika gehiago kontuan hartu, gero eta zailagoa da datu zuzenak aurkitzea; prozesaketa sakona sustatzeko ahalegin gehiago behar dira.

Horrenbestez, argi dago ahalegin handiagoak bideratu behar direla, bai euskararako baliabideak sortzeko, bai ikerketarako, berrikuntzarako eta garapenerako. Datu kopuru handien beharra eta hizkuntza-teknologietan oinarritutako sistemen konplexutasun handia direla medio, nahitaezkoa da harremanetarako eta lankidetzarako azpiegitura berriak garatzea.

4.7 HIZKUNTZARTEKO KONPARAKETA

HTen gaur egungo egoera oso desberdina da hizkuntza-komunitate batetik bestera. Hizkuntzen arteko egoerak alderatzeko, bi aplikazio-eremutan (itzulpen automatikoa eta hizketa-prozesaketa), oinarrizko teknologia batean (testu-analisia) eta, orobat, HTetan oinarritutako aplikazioak garatzeko behar diren funtsezko baliabideetan oinarritutako ebaluazio bat aurkeztu da atal honetan.Hizkuntzak multzokatzeko, bost puntuko honako eskala hau baliatu da:

1. HTen egoera bikaina
2. egoera ona
3. egoera ertaina
4. egoera osagabea
5. egoera apala

HTen egoera irizpide hauen bidez neurtu da:
Hizketa Prozesaketa: hizketa ezagutzeko dauden teknologien kalitatea, hizketa-sintesirako dauden teknologien kalitatea, landutako eremuen kopurua, dauden hizketazko corpusen kantitatea eta tamaina, hizketan oinarritutako aplikazio eskuragarrien kantitatea eta motak.
Itzulpen Automatikoa: Dauden MT teknologien kalitatea, landutako hizkuntza pareen kopurua, landutako fenomeno linguistikoen eta eremuen kopurua, dauden corpus paraleloen kalitatea eta tamaina, MT aplikazio eskuragarrien kantitatea eta motak.

Testu Analisia: Testua analizatzeko dauden teknologien kalitatea eta motak (morfologia, sintaxia, semantika), landutako fenomeno linguistikoen eta eremuen kopurua, eskuragarri dauden aplikazioen kantitatea eta motak, dauden testu-corpusen (etiketatuen) kalitatea eta tamaina, dauden baliabide lexikalen (adibidez, WordNet) eta gramatiken kalitatea eta motak.

Baliabideak: Dauden *testu-corpusen* kalitatea eta tamaina, *hizketa-corpusak* eta *corpus paraleloak*, dauden *baliabide lexikal* eta *gramatiken* kalitatea eta motak.

Goiko taulek erakusten dute ezen, azken hamarkadetan espainiar eta euskal gobernuek HTak diruz laguntzeko programei esker, euskarak Europako gainerako hizkuntza gehienak bezalako baliabideak dituela. Euskara eta hizlari kopuru handiagoa duten hizkuntzak parean daude, baina kontuan izan behar da beste hizkuntza horiek EBko hizkuntza ofizialak direla. Argi dago euskarazko HTetako baliabideak eta tresnak ez direla iristen gaztelaniazko maila bereko baliabide eta tresnen kalitatera eta estaldurara; izan ere, gaztelania ondo kokatuta dago ia HT eremu guztietan. Oraindik ere hizkuntza-baliabideetan hutsune asko dago euskararako, kalitate handiko aplikazioak sortzeari begira.

Hizketa-prozesaketarako, gaur egungo teknologiek aski emaitza onak dituzte, hainbat aplikazio industrialetan integratzeko, hala nola IVR elkarrizketa-sistemetan, nahiz eta diktatu-sistemetan betetzeko hutsunea izan, baita eremu mugatuetan ere. Bestalde, Itzulpen Automatikoko sistemek ez dute emaitza onik oraindik; euskara oso hizkuntza desberdina da, izan ere, hizkuntza aurreindoeuroparrekin alderatuta. Sailkatzaile estatistiko sakonagoak behar dira, jatorri bera duten beste hizkuntza pare batzuekin (esaterako, katalana-gaztelania edo galiziera-gaztelania pareekin) konparatuta. Alderdi linguistiko gehiago kontuan hartzen dituzten eta sarrerako testuaren analisi semantiko sakonagoa egiteko aukera ematen duten baliabideen eta teknologien behar garbia dago. Oinarrizko baliabide eta teknologia ho-

rien kalitatea eta estaldura hobetuta, hainbat aplikazioeremu aurreratu (kalitate handiko itzulpen automatikoa barne) garatzeko aukera berriak sortzeko gai izango gara.

4.8 ONDORIOAK

Liburu Zurien bilduma honetan, lehendabiziko ahalegin garrantzitsu bat egin dugu, Europako 30 hizkuntzatako hizkuntza-teknologien egoera aztertu eta hizkuntza horien arteko goi-mailako konparaketa bat erakusteko. Hutsuneak, beharrak eta gabeziak identifikatuz, Europako hizkuntza-teknologien komunitatearentzat eta interesa duten parteentzat errazagoa izango da teknologian oinarritutako benetako Europa eleaniztun bat eraikitzea helburu duen eskala handiko ikerketa- eta garapen-programa bat diseinatzea.

Ikusi dugu alde handiak daudela Europako hizkuntza batetik bestera. Zenbait hizkuntzatarako eta aplikazioeremutarako, kalitate oneko softwarea eta baliabideak badaude ere, beste hizkuntza batzuek (eskuarki hizkuntza "txikiagoek") hutsune handiak dituzte. Hizkuntza askori testu-analisirako oinarrizko teknologiak eta teknologia horiek garatzeko oinarrizko baliabideak falta zaizkie. Beste batzuek, berriz, badituzte oinarrizko tresna eta baliabideak, baina oraindik ez dute prozesaketa semantikoan inbertitzen. Hortaz, eskala handiko ahalegina egin beharra dugu, Europako hizkuntza guztien arteko kalitate handiko itzulpen automatikoa garatzeko helburu handira iristeko.

Euskararen kasua, hizkuntza-teknologien egoerari dagokionez, baikor baina zuhur aztertu beharra dago. Badago HT ikerketa-komunitate bideragarri bat Euskal Herrian, espainiar eta euskal ikerketa-programen bidez bultzatzen dena. Hainbat baliabide eta punta-puntako teknologia ekoitzi eta banatu dira euskararako. Hala ere, garatu diren baliabideen irismena eta tresnen multzoa oso mugatuak dira oraindik ere, gaztelaniarako (eta, noski, ingeleserako) dauden baliabide eta tresnekin alderatuta; beraz, ez dira nahiko, ez kalitateari dagokionez, ez kantitateari dagokionez, benetako jakintzaren gizarte eleaniztun bat sustatzeko beharrezkoak diren teknologia motak garatzeko.

Gaur egun, hizkuntza-teknologien industria aski hedatuta dago, eta ETT askok lantzen dute eremu hori, batez ere hizkuntza idatzirako. Haien produktuak euskararen estandarizazio-prozesua eta erabilera bultzatzeko tresna eraginkorrak izan dira eta dira oraindik ere. Euskara ez da ageri enpresa handien katalogoetan, ekimen jakin zenbaitetan izan ezik, Eusko Jaurlaritzaren laguntzaz, eskuarki.

Hainbat ikerketa-talde ari dira hizketaren eta hizkuntzaren prozesaketan 1988tik. Euskara salbuespena da hizkuntzaren tamainaren eta HBen egoeraren arteko korrelazioarekiko, eta horren zergatia ikerketa-talde horien lan koordinatuan datza. Baliabide gutxiago dituzten hizkuntzetako ikerketa eta garapena bultzatzeko, goi-mailako estandarizazio-irizpideei jarraitu behar zaie eta, orobat, kode irekien aldeko apustuari eta dagoeneko eginda dauden lan, tresna eta aplikazioen berrerabilerari.

Gure azterketek agerian uzten dute euskarazko HT baliabideak sortzeko ahalegin handia egitea eta baliabideok aurrera begirako ikerketa, berrikuntza eta garapena bideratzeko erabiltzea dela bide bakarra. Datu kopuru handien beharrak eta hizkuntza-teknologietan oinarritutako sistemen konplexutasun handiak ezinbesteko egiten du azpiegitura berriak eta ikerketa-antolaketa koherenteagoa garatzea, harreman eta lankidetza hobeak suspertuko badira. Kode irekia eta 2.0 komunitateak tresna lagungarriak izan daitezke tresna eta baliabide jasangarriak azkar garatzeko baliabide gutxiago dituzten hizkuntzetarako.

Jarraitutasun-falta ere badago ikerketaren eta garapenaren finantzaketan. Txandakatu egin ohi dira epe laburreko programa koordinatuak eta laguntza urri edo batere gabeko aldiak. Gainera, koordinazio-falta oroko-

rra dago EBeko beste herrialde batzuetako eta Europako Batzorde mailako programekin ere.

Beraz, ondoriozta dezakegu behar-beharrezkoa dela ekimen handi eta koordinatu bat, Europako hizkuntzen artean hizkuntza-teknologien desberdintasunak orekatzea xede duena.

META-NETen epe luzeko helburua da hizkuntza guztietarako kalitate handiko hizkuntza-teknologiak garatzea, aniztasun kulturalaren bidez bateratze politikoa eta ekonomikoa lortzeko. Dauden oztopoak eraisten eta Europako hizkuntzen artean zubiak eraikitzen lagunduko du teknologiak. Horretarako, baina, behar-beharrezkoa da interesdun guztiek -politikariek, ikertzaileek, enpresek eta gizarteak- indarrak batzea etorkizunerako.

Bikaina egoera	Ona egoera	Ertaina egoera	Osagabea egoera	Apala/Ez egoera
	Ingelesa	Alemana Espainiera Finlandiera Frantsesa Nederlandera Italiera Portugalera Txekiera	**Euskara** Bulgariera Katalana Daniera Eslovakiera Esloveniera Estoniera Galiziera Grekoa Hungariera Irlandera Norvegiera Poloniera Serbiera Suediera	Islandiera Kroaziera Letoniera Lituaniera Maltera Errumaniera

9: Hizketa-prozesaketarako hizkuntza-multzoak

Bikaina egoera	Ona egoera	Ertaina egoera	Osagabea egoera	Apala/Ez egoera
	Ingelesa	Frantsesa Espainiera	Alemana Katalana Nederlandera Hungariera Italiera Poloniera Errumaniera	**Euskara** Bulgariera Kroaziera Daniera Eslovakiera Esloveniera Estoniera Finlandiera Galiziera Grekoa Irlandera Islandiera Letoniera Lituaniera Maltera Norvegiera Portugalera Serbiera Suediera Txekiera

10: Itzulpen automatikorako hizkuntza-multzoak

Bikaina egoera	Ona egoera	Ertaina egoera	Osagabea egoera	Apala/Ez egoera
	Ingelesa	Alemana Espainiera Frantsesa Nederlandera Italiera	**Euskara** Bulgariera Katalana Daniera Eslovakiera Esloveniera Finlandiera Galiziera Grekoa Hungariera Norvegiera Poloniera Portugalera Errumaniera Suediera Txekiera	Kroaziera Estoniera Irlandera Islandiera Letoniera Lituaniera Maltera Serbiera

11: Testu-analisirako hizkuntza-multzoak

Bikaina egoera	Ona egoera	Ertaina egoera	Osagabea egoera	Apala/Ez egoera
	Ingelesa	Alemana Espainiera Frantsesa Nederlandera Hungariera Italiera Poloniera Suediera Txekiera	**Euskara** Bulgariera Katalana Kroaziera Daniera Eslovakiera Esloveniera Estoniera Finlandiera Galiziera Grekoa Norvegiera Portugalera Errumaniera Serbiera	Irlandera Islandiera Letoniera Lituaniera Maltera

12: Baliabideetarako hizkuntza-multzoak

META-NETI BURUZ

Europako Batzordeak sortutako bikaintasunezko sarea da META NET. Sareak Europako 33 herrialdetako 54 kide ditu, gaur egun. META-NETek META, Europa Eleaniztunaren Teknologia Aliantza, babesten du, hizkuntza-teknologiako aditu eta erakundeen talde europar gero eta handiagoa. META-NETek oinarri teknologikoak eman nahi ditu informazio-gizarte zinez eleaniztuna sortzeko Europan, eta hari eusteko. Horrelako gizarte bat lortu nahi da:

- Hizkuntzen arteko komunikazio eta lankidetzarako aukera ematen duena.
- Hizkuntza guztietan aukera berdinak ematen dituena informazioa eta ezagutzak eskuratzeko.
- Europarrei informazio-teknologia aurreratua eskaintzen diena sarean eta modu onean.

Merkatu digital eta informazio-esparru bakar batek osatutako Europa bateratu bat bermatu nahi du META-NETek eta, horretarako, Europako hizkuntza guztietarako bultzatzen eta sustatzen ditu teknologia eleaniztunak. Teknologia horiei esker, era askotako aplikazio eta esparruetan erabil daitezke itzulpen automatikoa, eduki-sorkuntza, informazio-prozesamendua eta ezagutza-kudeaketa. Halaber, hizkuntzan oinarritutako interfaze intuitiboak garatzeko aukerak ere sor daitezke hainbat teknologiatan, hala nola etxetresna elektronikoetan, makineria eta ibilgailuetan, nahiz ordenagailu eta robotetan. META-NET 2010eko otsailaren 1ean jarri zen abian, eta dagoeneko zenbait jarduera garatu ditu bere hiru ekintza-lerroetan: META-VISION, META-SHARE eta META-RESEARCH.

META-VISIONen helburua da akziodunen komunitate bizia eta eragin handikoa sortzea ikuspegi partekatu baten eta ikerketa-programa estrategiko (IPE) komun baten inguruan. Proiektu horren lan-ildo nagusia da Europan HTen komunitate koherente eta kohesiboa eratzea, akziodunen talde zatitu eta anitzetako ordezkariak elkartuz. Liburu Zuri hau beste 29 hizkuntzatako aleekin batera prestatu da. Teknologia bateratuaren ikuspegia alor banatan lantzeko hiru Hausnarketa Taldetan garatu zen. META Teknologia Kontseilua osatu zen hausnarketak egin eta IPEa prestatzeko, HTen komunitate osoarekin elkarlan zuzenaz eraikitako ikuspegian oinarrituz.

META-SHARE proiektuak baliabideak trukatzeko eta partekatzeko bitarteko ireki eta partekatua eskaintzen du. Biltegiekin osatutako parekoen sareak hizkuntza-datuak, tresnak eta web-zerbitzuak izango ditu, kalitatezko metadatuekin dokumentatuak eta kategoria estandarizatuetan antolatuak. Baliabideak erraz eskura daitezke, eta bilaketa uniformea da. Baliabideok kode irekikoak izan daitezke – doakoak, beraz – edo ordainduta eskuratu beharreko salgai mugatuak.

META-RESEARCH proiektuak zubiak eraikitzen ditu hurbileko teknologia-esparruetara iristeko. Xedea beste esparruetan aurrerapenak eragitea da, eta hizkuntza-teknologiaren onerako izan daitekeen ikerketa berritzailea aprobetxatzea. Ekintza-ildoaren oinarria da itzulpen automatikoan muturreko ikerketa garatzea, datuak biltzea, datu-bildumak prestatzea eta baliabide linguistikoak ebaluazio-lanetarako antolatzea, tresna eta metodoen inbentarioak sortzea eta komunitateko kideentzat ikastaroak eta prestakuntza-saioak

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

Language is the primary means of communication between humans. It allows us to express ideas and feelings, helps us to learn and teach, is essential for living, is the primary vehicle of transmission of culture, and is a symbol of identity.

> Language is the primary means of communication between humans.

In our current level of globalization, we have many ways to easily communicate with people from all over the world. For example, the new information and communications technologies have enabled the development of social networks that have encouraged and enhanced interaction between people from virtually all countries and cultures. Also, in recent years, we have seen large movements of foreign people between our countries, i. e., tourism or immigration, that creates the necessity for communication among different languages. This cross-lingual communication problem is often solved thought the use of a lingua franca.

The countries of Europe provide is a clear example of linguistic and cultural diversity despite the fact that, during the last 60 years, Europe has increasingly become a distinct political and economic entity. This means that from Basque to Polish and from Italian to Icelandic, language challenges are inevitably confronted by people in everyday life as well as in the spheres of business, politics and sciences. The European Union's institutions spend about a billion euros a year on maintaining their policy of multilingualism, i. e., translating texts and interpreting spoken communication. In parallel, English is becoming a lingua franca in the communication between European citizens.

In Spain, as a case in point, we find the same scenario. Spain has an official language, Spanish, also known as Castilian, and three co-official languages: Catalan, Galician and Basque. Preserving multilingualism in Spain has not been an easy task. It is the result of a complex process to intentionally preserve cultural and linguistic identity within and among the various regions and people of Spain. Similar to the use of English in the European case, direct communication between citizens of different language areas of Spain, often need the use of Castilian as a lingua franca.

> Multilingualism is a cultural heritage to be preserved.

At both, the European and the Spanish levels, multilingualism is a cultural heritage to be preserved. Globalization should not become a mechanism that promotes the abandonment of our rich linguistic and cultural heritage as it invites us to abandon the use of our own language in favor of a lingua franca. In a global communication environment, we should find ways to communicate broadly with the world while preserving our own language and, with it, our cultural identity.

Modern language technology and linguistic research can make a significant contribution to bridging these linguistic borders. When combined with intelligent devices and applications, language technology will in the

future be able to help citizens talk easily to each other and do business with each other even if they do not speak a common language. Language technology solutions will eventually serve as a unique bridge between different languages. However, the language technologies and speech processing tools currently available on the market (ranging from question answering systems to natural language interfaces, and including translation systems and summarization tools, among many others), still fall short of this ambitious goal.

> Language technology solutions will eventually serve as a unique bridge between different languages.

As early as the late 1970s, the EU realised the profound relevance of language technology as a driver of European unity, and began funding its first research projects. At the same time, national and autonomic projects were set up that generated valuable results but never led to concerted European action. The dominant actors in the field are primarily privately owned for-profit enterprises based in Northern America. The predominant language technologies today rely on imprecise statistical approaches that do not make use of deeper linguistic methods and knowledge. For example, sentences are automatically translated by comparing a new sentence against thousands of sentences previously translated by humans. The quality of the output largely depends on the amount and quality of the available sample corpus. While the automatic translation of simple sentences in languages with sufficient amounts of available text material can achieve useful results, such shallow statistical methods are doomed to fail in the case of languages with a much smaller body of sample material or in the case of sentences with complex structures. Analysing the deeper structural properties of languages is the only way forward if we want to build applications that perform well across a wide range of languages.

The solution to the cross-language communication problem is therefore to build key enabling technologies. To achieve this goal and preserve Europe's cultural and linguistic diversity, it is necessary to first carry out a systematic analysis of the linguistic particularities of all European languages, and the current state of language technology to support them. This is the purpose of the present book in what concerns the Basque language. This volume shows a detailed analysis of the language technologies, applications and solutions for Basque.

In the field of language technology, the Basque language shows a number of products, technologies and resources. There are application tools for speech synthesis, speech recognition, spelling correction, and grammar checking. There are also some applications for automatic translation, mainly between Spanish and Basque.

> Basque is one of the EU languages that still needs further research before truly effective language technology solutions are ready for everyday use.

As this series of white papers demonstrate, there is a dramatic difference between Europe's member states in terms of both the maturity of the research and in the state of readiness with respect to language solutions. One of the major conclusions is that Basque is one of the EU languages that still needs further research before truly effective language technology solutions are ready for everyday use. At the same time, there are good prospects for achieving an outstanding position in this important technology area. This development of high-quality language technology for Basque is urgent and of utmost importance for the preservation for a minority language as Basque.

2

RISK FOR OUR LANGUAGES AND A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digital information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many of the processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail send and receive documents faster than a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter, and Google+ facilitate communication, collaboration, and information sharing.

Although such tools and applications are helpful, they are not yet capable of supporting a sustainable, multilingual European society for all where information and goods can flow freely.

2.1 LANGUAGE BORDERS HINDER THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. But there is a strong likelihood that the revolution in communication technology is bringing people speaking different languages together in new ways. This is putting pressure on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In a global economic and information space, more languages, speakers and content interact more quickly with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

The global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages (English is the most common foreign language followed by French, German and Spanish.). 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the Web [2]. A few years ago, English might have been the lingua franca of the Web – the vast majority of content on the Web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital divide due to language borders has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our languages?

The variety of languages in Europe is one of its richest and most important cultural assets.

Europe's approximately 80 languages are one of its richest and most important cultural assets, and a vital part of its unique social model [41]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [3].

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investment efforts in language preservation focused on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [4]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport, energy and disability needs among others.

Digital language technology (targeting all forms of written text and spoken discourse) helps people collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us:

- find information with an Internet search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- hear the verbal instructions of a car navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core technologies are for each European language.

Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology adapted to all European languages that is robust, affordable and tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text (a page) using a suitably powered printing press. Human beings had to do the hard work of looking up, reading, translating, and summarizing knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies. Digital language technology can now automate the very processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive languagespeech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for highly specialised domains, and often exhibit limited performance. But there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, cultural heritage sites, edutain-

ment packages, libraries, simulation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. But citizens need to communicate across these language borders criss-crossing the European Common Market, and language technology can help overcome this final barrier while supporting the free and open use of individual languages. Looking even further forward, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to enable their own multilingual communities. Language technology can be seen as a form of 'assistive' technology that helps overcome the 'disability' of linguistic diversity and make language communities more accessible to each other.

Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow.

The current pace of technological progress is too slow.

Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technology challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

Technological progress needs to be accelerated.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to use it, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between its parents, siblings and other family members. From the age of about two, children produce their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples. Learning a foreign language gets harder with age.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

The two main types of language technology systems 'acquire' language capabilities in a similar manner. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach can require millions of sentences and performance quality increases with the amount of text analysed. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate all rely on statistical approaches. The great advantage of statistics is that the machine learns fast in continuous series of training cycles, even though quality can vary arbitrarily.

The second approach to language technology and machine translation in particular is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. But due to the high cost of this work, rule-based language technology has so far only been developed for major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focuses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology. Due to its multilingual community, this is particularly true of Europe's economic and information space. Although language technology has made considerable progress in the last few years, there is still huge potential in improving the quality of language technology systems. In the following, we will describe the role of Basque in European information society and assess the current state of language technology for the Basque language.

3

BASQUE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Basque – or *euskara*, in Basque –, known as 'Lingua Navarrorum' in Latin because it was the popular language in the Kingdom of Navarre, is the only surviving pre-Indo-European language in western Europe. It is considered an isolated language, with no known connections with other languages other than ancient Aquitanian. Both the origin of the language and its relationship with other languages continue to be controversial and of interest for many researchers.

Basque is presently spoken in a small region located at the west of the Pyrenees, on both sides of the border between Spain and France, in the region called *Euskal Herria* (Basque Country, in Basque) by the Basque community. The language has been loosing territory for centuries mainly on the south side. More recently, during the years of Franco's dictatorship when the use of Basque was forbidden, the language suffered an irreparable loss. Enormous efforts of revitalisation of the language were overtaken particularly from the 60s, where a network of schools was created introducing Basque into the educational system, clandestinely during its first years of existence. However, it is only from the 80's, with the linguistic political competences given to the Basque Government after the creation of the Autonomies, that Basque language started a recovery process.

In spite of the tremendous efforts made, in 2009 Basque appeared in the Unesco Map of the World's Languages in Danger [5] as a "vulnerable" language. Nowadays, Basque is estimated to be spoken by about 26% of the population of the Basque Country [6], either on the Spanish administration side or on the French administration side, but its status is not at all homogeneous. On one hand, the Spanish area of the Basque Country is divided into two political regions: in the Basque Autonomous Community, Basque is legally co-official along with Spanish, but with certain inequalities in favour of Spanish; in the Navarresse Community there are three different areas depending on the legal status of Basque: Basque-speaking, non-Basque-speaking, and mixed. The support for the language and the linguistic rights of the citizens vary depending on which of the three areas they are in. On the other hand, on the French side, Basque is spoken in the western half of the Département of Pyrénées-Atlantiques, but it has never had any legal status of any kind, and it is not official in any institution. However some years ago (2004), a public Agency was created to promote Basque language in French Basque country.

Basque has around 800,000 native speakers.

Spoken Basque shows a very high degree of dialectal dispersion. It is now commonly accepted that it is comprised of six dialects which have great differences among them. Standard or Unified Basque was not officially established until 1968 when the Academy of the Basque language *Euskaltzaindia* [7] made the first standardi-

sation proposal. These dialects have great differences between them in many aspects: lexical, phonetic, morphophonological and also prosodical, in accent and intonation. The dialects are not homogeneous entities; instead, they change continuously from one to another, and in several cases the limit between two or more of them is not so clear.

3.2 PARTICULARITIES OF THE BASQUE LANGUAGE

Basque is an agglutinative and high-inflective language whose major characteristic is that it is an ergative-absolutive language. That means that the subject of an intransitive verb is in the absolutive case (which is unmarked), and the same case is used for the direct object of a transitive verb; the subject of the transitive verb is marked differently, with the ergative case: the suffix -k.

Basque uses six different vowel sounds and thirty five consonant sounds.

Basque is postpositional; so, case and postpositional phrases are formed by attaching a suffix or concatenating more than one to the end of a phrase, according to the following scheme:

root + (article) + (number) + (case(s))

For example, «mutilarengana» (*towards the boy*) is formed by: «mutil+a+Ø+r+en+gan+a», – in which «mutil» is the lemma, or noun root; «a» is the article; «» the mark of singular; «r» an epenthetic particle; «en» the possessive genitive; «gan» the animate-being marker and «a» the allative.

This is an important characteristic to be taken into account in natural language and speech processing, since each noun-phrase can be inflected in 17 different ways, multiplied by 4 ways for its definiteness and number. These first 68 forms are further modified based on other parts of sentence, which in turn are inflected for the noun again. It has been estimated that, with two levels of recursion, a Basque noun may have 275 inflected forms, which is, on the other hand, very common [8]. This implies that it is necessary to find a way of dealing with all these ending variations starting from a basic lexicon.

The verbs are another example of the agglutinative character of Basque. The auxiliary verb, which accompanies most main verbs, agrees not only with the subject, but with any direct object and the indirect object present. Among European languages, this poly-personal agreement is only found in Basque, some languages of the Caucasus, and Hungarian (all non-Indo-European). Verbs in Basque follow the next scheme:

[verb_radical+aspect_suffix] [aux_verb]

For example, in Standard Basque «esaten zenizkidaten» (*you – 2nd person plural – used to tell me some things*) is formed by «esan» (*tell*, verb radical) + «ten» (frequentative aspect) and the auxiliary verb «zen+i+zki+da+Ø+te+n», in which «zen» marks the ergative second person; «i» is the auxiliary verb radical; «zki» the absolutive third person plural; «da» the dative first person singular; «Ø» is the indicative marker; «te» the ergative plural marker; and «n» the marker for the past tense. Due to this complexity, it is usual in Natural Language Processing research to opt for treating each of the auxiliary verbs as a whole, instead of dividing them into morphemes.

As far as the word order of the sentence is concerned, the basic syntactic construction is Subject-Objects-Verb (unlike Spanish, French or English where Subject-Verb-Objects construction is more common). The order of the phrases within a sentence can be changed with thematic purposes, whereas the order of the words within a phrase is usually rigid. As a matter of fact, Basque

phrase order is topic-focus, meaning that in neutral sentences (such as sentences to inform someone of a fact or event) the topic is stated first, then the focus. In such sentences, the verb phrase comes at the end. In brief, the focus directly precedes the verb phrase. This rule is also applied in questions, for instance, What is this? can be translated as «Zer da hau?» or «Hau zer da?», but in both cases the question tag «zer» immediately precedes the verb «da». This rule is so important that, even in grammatical descriptions of Basque written in other languages, the Basque word *galdegai* (focus) is used.

Basque orthography is almost phonemic: each grapheme corresponds to one phoneme, and so, the pronunciation of a word can be easily figured out from its written form. Nevertheless, there are a few exceptions: <l> and <n> are usually palatalised when they are preceded by <i> and followed by a vowel: mutila → <mutiLa> (*the boy*). Another example is that the consonant phoneme at the end of the negative particle "ez" (*no*) converts the contiguous next phoneme in a voiceless phoneme: ez dira → <eztira> (*they are not*).

3.3 RECENT DEVELOPMENTS

A standardised form of the Basque language, called *Euskara Batua*, was developed by *Euskaltzaindia*, the Academy of the Basque Language in the late 1960s. Euskara Batua was created so that Basque language could be used – and easily understood by all Basque speakers – in formal situations (education, mass media, literature...), and this is its main use nowadays. For classic literary reasons, Standard Basque is based mainly on the Central and Navarrese-Labourdin dialects. The extreme dialects, differ noticeably from it, despite that the Western dialect is one of the most spoken dialects of the language together with the Central dialect.

Standard Basque has solid foundations and it is developing forward aspects as syntax and naturalness. At present, almost all the people that study Basque learn the Euskara Batua. This fact has created a phenomenon all around the Basque country in which Basque people speak their own local dialect with locals, and standard Basque with the 'new Basque speakers' (*euskaldun berri*). In the Western area, due to the great differences between the western dialect and the standard, it has led to a situation where people studying Basque feel that the language they are studying is pretty far from what Basque people speak. On the other hand, it is now already a fact that there are standard Basque speakers whose mother tongue is precisely standard Basque, because many new Basque speakers opt to speak to their children in Basque, even that their own primary language was Spanish.

However, the idea that the future of Basque is related not only to the development of Standard Basque but also to the promotion of the current dialects is more and more accepted by the theoreticians of the Basque language [9]. So, dialects will be somehow important in the future applications of LT for Basque.

The Basque LT community and researchers, conscious of the importance of technologies for languages spoken by small communities to evolve in the 21st century, have made a great effort to place Basque at the same technological level as the most used languages. There is a solid scientific experience along with other neighbouring languages, such as Catalan and Galician, that is virtually unique in Europe, such as the development of cross-lingual products and services between regional languages.

The importance of the development of a LT industry for Basque is evident taking into account the creation of *Langune* [10]. Langune is an association of Basque Country companies belonging to the Language Industry sector. This association was set up in 2010 and brings together over 30 companies in the spheres of translation, content, teaching and language technologies. Its main objective is to develop the sector of LT, which will be

a benchmark in the language industry in Europe, while avoiding the duplication of efforts and achieving synergies. Langune has just started but is taking giant steps.

3.4 LANGUAGE CULTIVATION IN BASQUE

The Basque language is mainly represented by '*Euskaltzaindia*', the Royal Academy of the Basque Language (1919). It carries out research in the language, seeks to protect it and establishes standards of use. It enjoys full official recognition as a royal academy in Spain (1976) and as a cultural association of public benefit within the territory of France (1995).

Since the declaration of Basque as the official language in the Autonomous Basque Community, the Basque Government has developed numerous norms and laws in order to protect and favour the use of the language. Various organisms and institutions have since been created: Basque Advisory Board (1982), Basque Radio-Television EiTB (1982), the Institute for Adults Literacy-HABE (1983) and many others.

The 'General Plan for the Promotion of the Use of Basque' was first introduced in 1998 as a strategic instrument with three main objectives: reach consensus in goals and actions of the different institutions, establish priorities for the founding programmes and coordinate the activities of institutions, companies and associations dealing with Basque. Within this strategic Plan, periodical sociolinguistic surveys serve as guide for establishing new goals and correction directions. The Basque Government has a web-portal www.euskara.euskadi.net dedicated to the Basque language, offering information not only about the language and its history and present situation, but also links to every kind of service, product or application related with the language, including public funding programmes. In the French area, the "Office Public de la Langue Basque" [11] was created in 2004, as a public Agency bringing together four local or regional public institutions and the state, with the goal of defining and applying a common linguistic policy in the region to promote Basque language.

3.5 LANGUAGE IN EDUCATION

In the Basque Autonomous Community, Basque was officially introduced in the public education system in 1983 with the law that regulates the use of Basque and Spanish in the Primary and Secondary School. For the Primary and Secondary School three models were created, giving the possibility to each institution to choose the model to offer. In model A the vehicular language is Spanish, and Basque is taught in the subject "Basque Language and Literature". In model D – the letter C is not normally used in Basque – Basque is the vehicular language and there is one subject "Spanish Language and Literature" taught in Spanish. Model B is an intermediate model, where some of the subjects are taught in Spanish (mainly Reading and Writing and Mathematics) and another part in Basque (mainly science and plastic). However, the Model A has been losing students progressively, in favour of Model B, mainly in pre- and primary school, where more than half of the students learn in Model D. Yet, 85% of the 15 years old students made the examinations for the PISA Study in Spanish whilst only 15% did them in Basque [12], clearly showing that Spanish is the dominant language in Education.

In the Navarresse Community, where Basque has different grades of official status depending on the area, a forth model was also available with no mandatory subject of Basque. As for the Northern provinces in France, primary education in Basque is offered by the private network of schools 'Seaska', which is managing presently almost 2700 students in 29 establishments that include one centre for secondary education and one 'lizeo'.

Very recently, new models are being proposed and tested, which consider the importance of early learning

of English. The Basque Government in Spain has recently introduced a trilingual model, while in Navarre bilingual education in Spanish and English has been introduced, although Basque is offered optionally.

At higher levels of education, the offer is clearly dominated by Spanish. From the three existing universities, the only public university, Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU), offers the possibility of learning in Basque, and although enormous efforts have been made to make equal offer in Basque as in Spanish, only very few degrees can be taken fully in Basque. Remarkably, a Master and Doctorate Program 'Analysis and Processing of Language' [13] totally offered in Basque exists since the year 2001. The private University Mondragon Unibertsitatea offers most of their degrees in Basque and some of their Master studies in Basque. The third University, Universidad de Deusto, offers only some of the courses in Basque.

3.6 INTERNATIONAL ASPECTS

Since January 2009, the Etxepare Basque Institute is the Basque public institution responsible for spreading the Basque language and culture all over the world. This institution is aiming to promote the teaching, study and use of Basque throughout the world and to include the contributions of all the communities that share Basque as a common language. The Institute also aims to disseminate Basque culture in the international community with very special reference to those groups that speak Basque, including the Basque Diaspora. Along the history, many Basques have left the Basque Country for other parts of the globe for economic and political reasons; Basque Diaspora is the name given to describe people of Basque origin living outside their traditional homeland. Currently there are substantial Basque origin populations in Chile, Argentina, Bolivia, Ecuador, Colombia, Cuba, Mexico, Venezuela, Canada and the United States. All of them have several Basque cultural centres (*Euskal Etxeak*) that were established to pursue the same objective: the perpetuation of Basque culture and identity. There are Basque cultural centres in most large cities of 24 different countries [14].

The origins and singular structure of Basque have raised the interest in the study of Basque language and culture. Currently it can be learned in 29 universities belonging to 13 different American and European countries.

Regarding the use of Basque in international institutions, the Spanish government has made efforts in favour of including it, together with Catalan and Galician among the official languages of the European institutions. But currently they do not enjoy the status of official languages; they are considered semi-official, together with Scottish, Gaelic and Welsh. Basque can only be used in very limited situations: it can be spoken at the work sessions of the Region Committee and the Council, but not in the plenary meetings of the European Parliament. Citizens can also write to the European institutions using Basque and have right to be answered in the same language, but always through the Spanish Government and this government must pay the derived fees.

29 universities from 13 different American and European countries offer Basque studies.

Basque is included in the list Regional and minority languages of the European Union [15] and as such it benefits from the resolutions adopted by the European Parliament to promote action on regional and minority languages.

Language technology can address this challenge from a different perspective by offering services like machine translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

3.7 BASQUE ON THE INTERNET

In the first quarter of 2010, 61.4% of the households (513,000) in the Basque Country had a computer. There were slightly over 460,000 families, of which 54.9%, had access to the Internet from their homes. This means that over a million people aged 15 and over were Internet users. Most of them stated to be online every day. Only 22.9% of them used Basque language on the Internet [42].

Basque is used by 0.5% of all the websites that rank in the top 1,000.

Nevertheless there is a strong and willing community of Internet users among Basque speaking people. The blogosphere in Euskara, the Wikipedia and online services in Euskara, as well as the location of tools and operating systems based on free software, have fostered the presence of Euskara and Basque culture, both on the Internet and ICT, encouraging, in this way, the expansion of its use. For instance, the Basque Wikipedia has more than 120,000 articles occupying the 36th place in number of articles among all the Wikipedia. And a big effort has been made in order to provide different common software programs [16, 17] and resources in Basque [18, 19, 20, 21, 22].

A new top level domain .eus has been registered and will be launched in mid 2012. It already counts with 193 pre-registrations. The proposed top-level domain .eus is the name that will represent the Community of the Basque Language and Culture on the Internet. This symbol will become a tool for the promotion of Basque culture and Euskara, and, in this sense, the .eus domain will be an effective mechanism for linguistic standardisation of Euskara worldwide. The .eus domain, through the virtual space of the Internet, will assure an efficient promotion of Euskara, guaranteeing simultaneously its international recognition. Similarly, the .eus domain will reinforce and extend the multicultural nature of the Internet, since allowing linguistic and cultural communities to have their own domain puts multiculturalism at the very heart of the Internet. Domains related to language and cultures strengthen and benefit not only those linguistic and cultural communities but also the Internet itself [23].

The Basque Wikipedia, with 123,787 articles, is the 36th largest Wikipedia in terms of the number of articles.

For language technology, the growing importance of the Internet is important in two ways. On one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas involving language technology.

4

LANGUAGE TECHNOLOGY SUPPORT FOR BASQUE

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction
- authoring support
- computer-assisted language learning
- information retrieval
- information extraction
- text summarisation
- question answering
- speech recognition
- speech synthesis

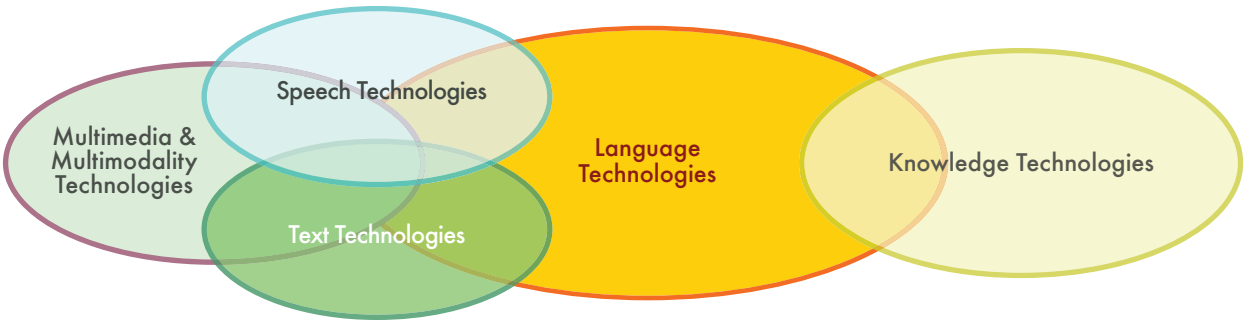
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [43, 44, 45, 46, 47].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input languages, and so on.
2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.



1: Language technologies

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns in the sentence); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

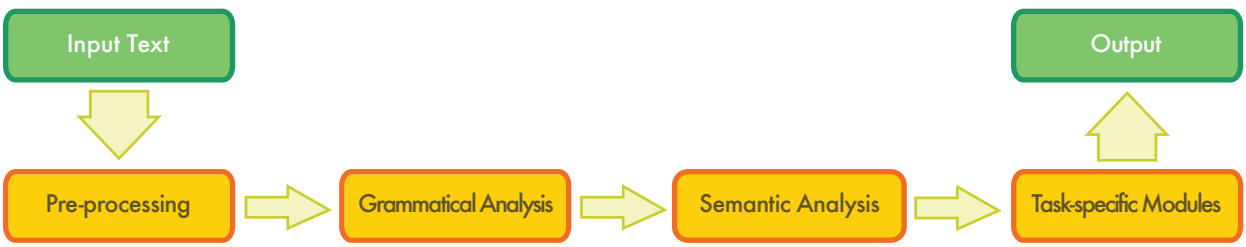
In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Basque in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Basque language is summarised in figure 7 (p. 60) at the end of this chapter. This table lists all tools and resources that are boldfaced in the text. LT support for Basque is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

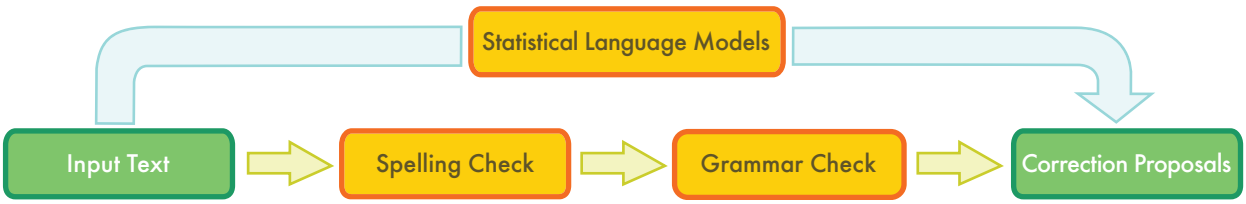
In this section, we focus on the most important LT tools and resources, and give an overview of LT activities in Basque. Tools and resources that are set in bold in the text can also be found in the table at the end of this chapter.

4.2.1 Language Checking

Anyone using a word processing tool such as Microsoft Word has come across a spell checking component that indicates spelling mistakes and proposes corrections.



2: A typical text processing architecture



3: Language checking (top: statistical; bottom: rule-based)

Forty years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. Using language-dependent algorithms for **grammatical analysis**, they detect errors related to morphology (e. g., plural formation) as well as syntax-related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [24]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e. g., in Basque, for deciding if the ergative marker has to be used, as in:

- *Liburua neskak dauka*
  [The girl has the book]
- *Irakurlea neska da.*
  [The reader is a girl.]

Language checking(see figure 3) either requires the formulation of language-specific **grammars**, i. e., a high degree of expertise and manual labour, or the use of a statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i. e., the preceding and following words). For example, *neskak dauka* is a much more probable word sequence than *neska dauka*. A statistical language model can be automatically derived using a large amount of (correct) language data (i. e., a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Basque with its richer inflection and agglutinative morphology. In fact, language modelling for Basque poses enormous difficulties due to the impossibility of collecting all possible word-forms.

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and damage claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, and at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions.

Language checking is not limited to word processors but also applies to authoring systems.

The most used Spell Checker for Basque is Xuxen [25], which was developed by the university research group IXA (http://ixa.si.ehu.es) and is supplied by the SME Eleka Ingenieritza Linguistikoa. This Spell Checker is not limited to the use of a lexicon as it is common practice for English or other less-inflected languages. On the contrary, morphological analysis is performed. The newest version of this spell checker also performs grammar and style corrections. This version also includes code developed by the company Hizkia [26] and the institution UZEI [27].

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to Web Search engines, e. g., Google's 'Did you mean ...' suggestions.

4.2.2 Web Search

Search on the web, in intranets or in digital libraries, is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide [28].

Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, they incorporated basic semantic search capabilities into their algorithmic mix [29], which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

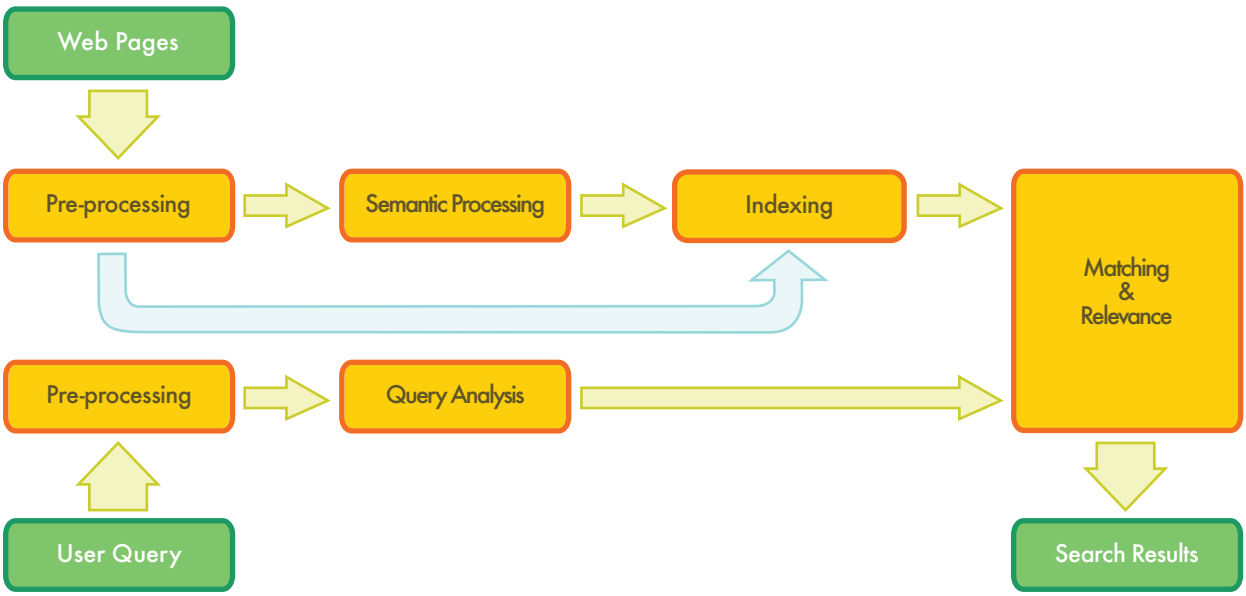
For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge to facilitate text interpretation. Experiments using **lexical resources** such as machine-readable thesauri or ontological language resources like WordNet have demonstrated improvements in finding pages using synonyms of the search terms. Again, these developments require language-specific resources. A Basque WordNet 'BasWN' has been developed by the research group IXA at the University of the Basque Country and is commercially available through ELRA.

The next generation of search engines will have to include much more sophisticated language technology.

The next generation of search engines will have to include much more sophisticated language technology, especially to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, a syntactic as well as **semantic analysis** is required. The system also needs to provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, rather than companies that have acquired other companies.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognisers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically



4: Web search architecture

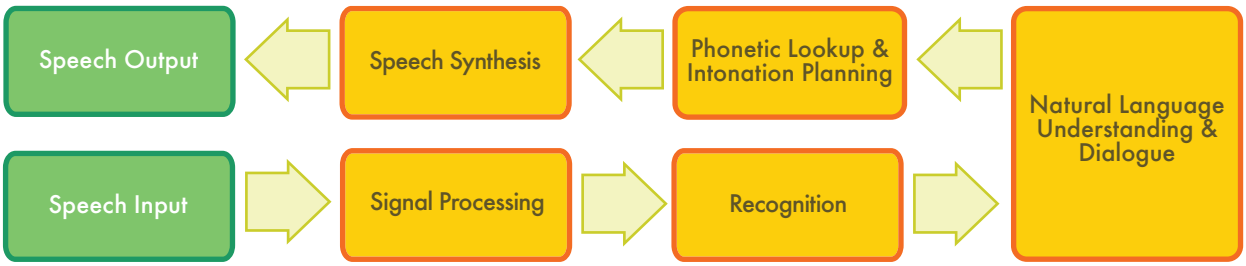
translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling multimedia information retrieval, i. e., information search on images, audio, and video data. For audio and video files, this involves a **speech recognition** module to convert speech content into text or a phonetic representation, to which user queries can be matched.

Focus on development for these companies lies on providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small **text corpora**. Processing time easily exceeds that of a common statistical search engine as, e. g., provided by Google by a magnitude of thousands. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.

In the Basque Autonomous Community, the small company Eleka Ingeniaritza Linguistikoa has been very active in the development of applications and web based services for Basque. They usually integrate LT research results and resources such as lemmatisers and lexical databases of the IXA group and Elhuyar Foundation. The multilingual search engine *elebila* considers the Basque language specifics and integrates various linguistic tools and resources to offer high quality search results for Basque. Another example is the tool called Miatu ('Examine' in Basque), a library offering functionality to search in special purpose indexed databases using lemmatisers and other morphology analysis tools. It has been used to develop the science related web portal www.zientzia.net and the educational content portal www.ikasbil.net.

4.2.3 Speech Interaction

Speech Interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e. g., a graph-



5: Speech-based dialogue system

ical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, and telecommunications. Other usages of Speech Interaction technology are interfaces to particular devices, e. g., navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e. g., in smartphones.

At its core, Speech Interaction comprises the following four different technologies:

- Automatic **speech recognition** (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpretting the latter according to the purpose of the respective system.
- Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- **Speech synthesis** (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be output to the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from **speech corpora**, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly – prompted by a *How may I help you?* greeting – are better accepted by users.

Speech interaction is the basis for interfaces that allow a user to interact with spoken language.

For the output part of a VUI, companies tend to use utterances pre-recorded by professional – ideally corporate – speakers a lot. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the more dynamic content an utterance needs to consider, the more the user experience may suffer from a poor prosody re-

sulting from concatenating different parts of audio files. In contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i. e., economically strong countries with a considerable population – are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe. Since 2007, thanks to the support given by the Basque Government, Basque language is included in the catalogue of products of Nuance. However, the offer in ASR is limited to small to medium size vocabulary applications and no dictation product is available. For TTS, just one female voice is available. On the Spanish market, the Catalan SME Verbio Speech Technologies [30] also offers Basque both for ASR and TTS, with more than one voice. Still, no commercial dictation system exists for Basque.

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Most of the companies on the Spanish TTS market are essentially application developers. Key players in the Spanish market are: Indsys [31] (Intelligent Dialogue Systems), Fonetic [32], Ydilo [33] and NaturalVox [34]. Some of them have a limited offer in Basque. Free TTS software for the Basque language is also offered by the research group Aholab [35] of the University of the Basque Country (UPV/EHU).

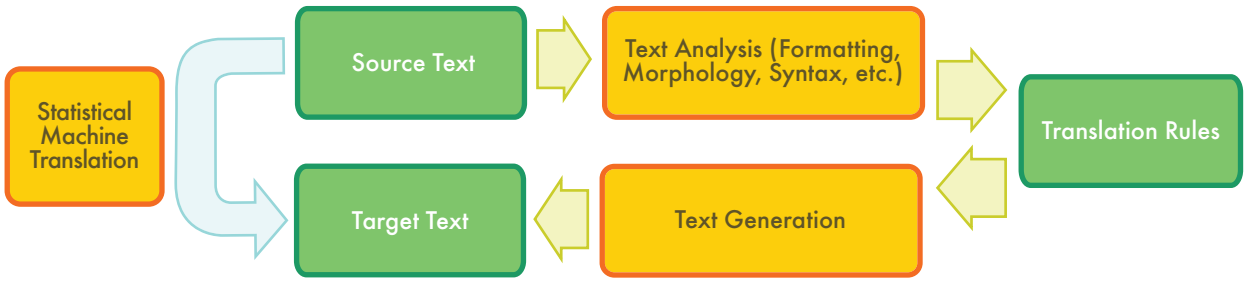
Looking beyond today's state of technology, there will be significant changes due to the spread of smart phones as a new platform for managing customer relationships – in addition to the telephone, Internet, and email channels. This tendency will also affect the employment of technology for speech interaction. On one hand, demand for telephony-based VUIs will decrease, in the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smart phones will gain significant importance. This tendency is supported by the observable improvement of speaker independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smart phone users. Given this 'outsourcing' of the recognition task to the infrastructure of applications, the application-specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

4.2.4 Machine Translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, **Machine Translation** (MT) still fails to fulfil the high expectations it gave rise to in its early years.

> At its basic level, Machine Translation simply substitutes words in one natural language with words in another language.

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e. g., weather reports. However, for a good translation of less standardised texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty here lies in the fact that hu-



6: Machine translation (left: statistical; right: rule-based)

man language is ambiguous, which yields challenges on multiple levels, e. g., word sense disambiguation at the lexical level ('Jaguar' can mean a car or an animal) or on other levels as in:

- *Egon garenetan ez dugu topatu*
  [Each time we were there we have not seen him/her]
  or [In every place we were we have not seen him/her]
- *Aitak semeari bere bizikleta eman dio*
  [The father has given his bicycle to his son]

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like in the second example above. But often rule-based (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by a skilled linguist.
Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl **parallel corpus**, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven MT are complementary, researchers nowadays unanimously target hybrid approaches combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For Basque, MT is particularly challenging. The rich morphology, the high degree of inflection and the agglutinative character of the language makes dictionary analysis and dictionary coverage difficult. Additionally, due to the order of the sentence components, parallel corpora are difficult to manage.

Matxin is a Transfer-based MT system from Spanish into Basque developed by IXA Group at the University of the Basque Country (UPV/EHU). It is an open, reusable and interoperable framework useful even for other language-pairs (matxin.sourceforge.org). It uses other open source codes such as Freeling, and reuses Basque morphology for morphological generation. IXA Group has also created an improved Statistical Machine Translation system for Basque Spanish that deals with morphological segmentation and word reordering (EUSMT. http://ixa2.si.ehu.es/openmt-demo/). For the development of these MT systems, there is strong collaboration between the university research group, the local SME *Eleka Ingeniaritza Linguistikoa* and the *Elhuyar Foundation*, which provides considerable amounts of linguistic resources. This SME has also developped the translator Standard Basque *batua* – Western dialect *bizkaiera*. Also, a Basque to Spanish initial system has been developed by the Transducens Group at Universitat d'Alacant, using the platform Apertium. Google's Translator offers an alpha version for Basque.

Leading international MT developer Lucy Software has an important subsidiary in Spain, Lucy Iberica [36], former Translendium. This company was selected in 2008 by the Basque Government to develop a Spanish-Basque translation system and again in 2011 to continue the work.

Provided good adaptation in terms of user-specific terminology and workflow integration, there is a wide consensus that the use of MT can increase productivity significantly. The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, many language pairs are still missing.

Evaluation campaigns help to compare the quality of MT systems, their approaches and the status of the systems for different language pairs. Figure 7 (p. 24), which was prepared during the Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [37]. A human translator would normally achieve around 80 points. The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese, Finnish).

4.3 OTHER APPLICATION AREAS

Building Language Technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities 'under the hood' of the system. Therefore, they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled, how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?), and how can specific information – the answer – be reliably extracted from a document, without unduly ignoring the context.

Language technology applications often provide significant service functionalities behind the scenes of larger software systems.

This is in turn related to the information extraction (IE) task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. IE aims at identifying specific pieces of information in specific classes of documents; this could e. g., be the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas, which sometimes play the role of stand-alone application and sometimes that of supportive, 'under the hood' component are text summarization and **text generation**. Summarisation, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying 'important' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarisation equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarisers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesise *new sentences*, i. e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

For Basque and for most languages, research in most text technologies is much less developed than for English.

For Basque, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarisation have since the 1990s been the subject of numerous open competitions, primarily those organised by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Basque was never a targeted language. Accordingly, there are hardly available annotated corpora or other resources for these tasks. Summarisation systems, when using purely statistical methods, are

often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realisation modules (the "generation grammars"); again, most available software is for English.

4.4 LANGUAGE TECHNOLOGY IN EDUCATION

Language Technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. Consequently, the current basic training of a computational linguist may be performed in Spain within the framework of a degree in Philology or Linguistics, which includes Computational Linguistics as a core subject, or by Computational Science faculties. Among the Universities that offer the first option: Universitat de Barcelona, Universitat Pompeu Fabra, Universitat Oberta de Catalunya and Universidade de Vigo. On the other hand, main computational science faculties offering Computational Linguistic as subject are: Universidad Politécnica de Madrid, Universidad Carlos III, Universidad Autónoma de Madrid, Universitat d'Alacant, Universidad Nacional de Educación a Distancia and Universidad del País Vasco / Euskal Herriko Unibertsitatea. Other cases, such as the Universidad Complutense combine both. Graduate courses offer a more targeted professional training. There are several doctoral programs which offer masters or subjects related to language and speech processing. A complete doctoral program on Language Processing is offered by Universidad del País Vasco / Euskal Herriko Unibertsitatea, also totally offered in Basque. Modules in Language Technology are also offered to students of other master or PhD courses, particularly in Speech Processing (e. g., Master TICRM of the UPV/EHU).

There are several research groups spread across the 3 universities of the Basque Autonomous Community, working on speech processing, speech synthesis and conversion, speech and speaker recognition, language recognition, natural language processing, text-to-text translation and speech-to-speech translation. All of them are members of the Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN, Spanish Society for Natural Language Processing), a non-profit organisation with over 300 members, both from academia and industry, which was created in 1984 with the purpose to promote and spread activities related to teaching, research and development of NLP, on both national and international level. SEPLN organises seminaries, symposiums and conferences and promotes collaboration with national and international institutions.

SEPLN organises an annual conference, which is attended yearly by an increasing number of researchers working on NLP, both from Spain and abroad. The association also edits a periodical journal and maintains a web server with information about issues related to the natural language processing and an open forum for members. The Spanish Network on Speech Technology (RTTH) [39] is a common forum where researchers (presently more than 250 researchers) in Speech Technology gather to combine efforts and share experiences in order to:

- Promote research in speech technology to attract new young researchers in this field through training, student exchanges, scholarships and awards.
- Attract investments for business research by finding new applications that offer new business opportunities.
- Progress in building partnerships and integration of network members to maintain Spain's leadership in the investigation of Spanish, and also enhance co-official languages such as Catalan, Euskara and Galician.

RTTH has been promoting every other year the "Jornadas en Tecnología del Habla" since 2000. This workshop pursues the aims of being a meeting point to present and discuss the results of the research on speech and language technologies on Iberian languages. They also aim at promoting industry/university collaboration. A wide variety of activities: technical papers presentations, keynote lectures, presentation of project reports and laboratories activities, demos, and recent PhD thesis presentations are defined.

4.5 LANGUAGE TECHNOLOGY PROGRAMS

Technology programs for the Basque language have been supported mainly by the Basque and the Spanish Government. The Spanish Ministries of Education and Science and Innovation have supported research in the field of information technologies through national research programs. These programs have impelled numerous research projects and collaboration with international research centres and companies. The basis of technology development and commercial applications for automated processing of the Basque language has been partly created as a result of these projects.

Since 2000 up till today, the Spanish Government supported within the National Plan of Research and Technology several projects in the area of Multilingual Speech Technologies: TEHAM, AVIVAVOZ, and BUCEADOR. Their main purpose was to improve the quality of Speech Recognition, Speech Translation and Text to Speech Synthesis in all the official languages spoken in Spain: Basque, Galician, Catalan and Spanish.

The Centre for the Development of Industrial Technology (CDTI) is a Spanish public organisation, under the Ministry of Science and Innovation, whose objective is to help Spanish companies to increase their technological profile. CDTI evaluates and finances R&D projects through programmes such as CENIT (finalised in 2010) and AVANZA.

The Basque Government supports research and innovation through the "Plan de Ciencia y Tecnología" (PCTI). Within this plan, several bodies and research and innovation agencies have been created in the last years: *The Basque Council for Science, Technology and Innovation* (the highest political body leading actions to promote and develop research and innovation), *InnoBasque* (The Basque Agency for Innovation) and *IkerBasque* (Basque Foundation for Science), whose main instrument is the attraction of talented researchers to the Basque Science and Technology system. Important instruments of the PCTI plan are the calls for research and innovation projects: the program *ETORTEK*, addressed to the agents of *Basque Network for Science, Technology and Innovation*, and the program *ETORGAI*, addressed to private companies.

In the last *PCTI2010*, as had already been in previous plans, Language Tecnologies have been identified as one strategic field. As such, during the last 10 years, the projects *HIZKING21, ANHITZ*, and presently *BERBATEK* [40] have been carried out under the *ETORTEK* program. Most of the existing resources and tools for Basque have been obtained through these projects.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Table 7 provides an overview of the current situation of Language Technology support for Basque. Several leading experts rated the existing tools and resources based on educated estimations using seven criteria (each ranging from 0 to 6). In this white paper series, a first effort has been made to assess the overall situation of many European languages with respect to language technology support in a way that allows for high level comparison

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	2	1	1	1	4	3	2
Speech Synthesis	2	3	4	4	4	3	3
Grammatical analysis	4	2.5	4	4	4	2.5	2.5
Semantic analysis	1	1.5	2	1	1	1	1
Text generation	1	0	0	0	0	0	0
Machine translation	3	5	2	3	3	2	2
Language Resources (Resources, Data and Knowledge Bases)							
Text corpora	2	4	3	2	3	4	2.5
Speech corpora	3	2	3	2	3	3	2
Parallel corpora	2	4	2	2	2	2	1
Lexical resources	4	4	4	5	5	4	3
Grammars	2	2	2	2	2	2	2

7: State of language technology support for Basque

and identification of gaps and needs. For Basque, key results include the following:

- Speech processing developments currently show a more mature situation for speech synthesis than for speech recognition. More efforts have to be done in the development of language models that account for the special morphology of Basque.
- Everyday applications that integrate speech technology such as voice-based interfaces to mobile phones, car navigation systems or spoken dialog systems are rarely available in Basque.
- The spelling checker is one of the most powerful tools in the ongoing standardisation of Basque and the most representative of the effective LT tools created to promote the use of Basque.
- Standard resources for Basque have adopted TEI and XML standards as a basis for linguistic annotation at the different levels of processing, and also to the definition of a general methodology for written corpus annotation. However, several resources lack standardisation, i. e., even if they exist, sustainability is not always given; concerted programs and initiatives are needed to standardise data and interchange formats.
- Text semantics is more difficult to process than word and sentence semantics. There is a Wordnet for Basque, and promising algorithms to examine similarity between words and to extract facts from text have been developed.

From this, it is clear that more efforts need to be directed into the creation of resources for Basque and into re-

search, innovation, and development. The need for large amounts of data and the high complexity of language technology systems make it also mandatory to develop new infrastructures for sharing and cooperation.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing) and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were clustered using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 7 to 10 show that, thanks to LT funding programs from the Spanish and Basque governments in recent decades, the Basque language is equipped as most of other European languages. It compares well with languages spoken by a bigger number of speakers despite these are official languages of EU countries. This is mainly due to coordinated efforts of research groups and small developers of LT tools. But LT resources and tools for Basque clearly do not yet reach the quality, size and coverage of comparable resources and tools for the Spanish language, which is in a good position in almost all LT areas. There are still some gaps in Basque language resources and tools with regard to high quality applications.

For speech processing, current technologies perform well enough to be successfully integrated into a limited number of industrial applications such as IVR spoken dialogue systems, although there is still a gap to fill for dictation systems, even in a constrained domain. Machine Translation systems do not get a good performance yet, due to the fact that Basque is very different from the Indo-European languages. Deeper statistical classifiers are needed compared to other language pairs with similar origin, such as Catalan-Spanish or Galician-Spanish. There is a clear need for resources and technologies to cover a wider range of linguistic aspects and to allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced

application areas, including high-quality machine translation and continuous speech recognition.

4.8 CONCLUSIONS

In this series of white papers, we have made an important initial effort to assess language technology support for 30 European languages, and provide a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled Europe.

We have seen that there are huge differences between Europe's languages. While there are good quality software and resources available for some languages and application areas, others (usually "smaller" languages) have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources for developing these technologies. Others have basic tools and resources but are as yet unable to invest in semantic processing. We therefore still need to make a large-scale effort to attain the ambitious goal of providing high-quality machine translation between all European languages.

The situation of Basque concerning language technology support gives rise to cautious optimism. There is a viable LT research community in the Basque Country, which has been mainly supported by Spanish and Basque research programmes. A number of resources and state-of-the-art technologies have been produced and distributed for Basque. However, the scope of the resources and the range of tools are still very limited when compared to the resources and tools for the Spanish language (and obviously for the English language) and they are simply not sufficient in quality and quantity to develop the kind of technologies required to support a truly multilingual knowledge society.

The Basque language technology industry is well established and a significant number of SME are active in this sector, although mostly for written technologies. Their products have been and still are effective tools supporting the standardisation process and promoting the use of Basque. Basque has not been included in the catalogue of large companies, except for a few specific actions, and usually supported by the Basque Government.

There are several research groups working in speech and language processing since 1988. If Basque is now an exception to the correlation between language size and LR scarcity is due to the coordinated efforts of those research groups. Research and development for less resourced languages should be faced following high standardisation criteria, open-source coding and reusing language foundations, tools and applications.

Our findings show that the only alternative is to make a substantial effort to create LT resources for Basque, and use them to drive forward research, innovation and development. The need for large amounts of data and the extreme complexity of language technology systems makes it vital to develop a new infrastructure and a more coherent research organisation to spur greater sharing and cooperation. Open source initiatives and the 2.0 communities can be important instruments for a rapid and sustainable development of tools and resources for less resourced languages.

There is also a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level. We can therefore conclude that there is a desperate need for a large, coordinated initiative focused on overcoming the differences in language technology readiness for European languages as a whole.

META-NET's long-term goal is to introduce high-quality language technology for all languages in order to achieve political and economic unity through cultural diversity. The technology will help tear down existing barriers and build bridges between Europe's languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts for the future.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	**Basque** Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

8: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	**Basque** Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

9: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	**Basque** Bulgarian Catalan Czech Danish Finnish Galician Greek Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian

10: Text analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	**Basque** Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

11: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence funded by the European Commission [48]. The network currently consists of 54 members from 33 European countries. META-NET forges META, the Multilingual Europe Technology Alliance, a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- grants all Europeans equal access to information and knowledge regardless of their language;
- builds upon and advances functionalities of networked information technology.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technolo¬gies support automatic translation, content production, information processing and knowledge management for a wide variety of subject domains and applications. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots.

Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META¬-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vi¬sion and a common strategic research agenda (SRA). The main focus of this activity is to build a coher¬ent and cohesive LT community in Europe by bring¬ing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised cate¬gories. The resources can be readily accessed and uni¬formly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technol¬ogy fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language re-sources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

AIPAMENAK REFERENCES

[1] Aljoscha Burchardt, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] European Commission Directorate-General Information Society and Media. *User language preferences online*. Flash Eurobarometer 313, 2011.

[3] UNESCO Director General. *Intersectoral mid-term strategy on languages and multilingualism*. Paris, 2007.

[4] European Commission Directorate-General for Translation. *Size of the language industry in the EU*. Kingston Upon Thames, 2009.

[5] UNESCO – Languages and Multilingualism . http://www.unesco.org/en/languages-and-multilingualism.

[6] Euskal Estatistika Erakundea (Basque Statistics Institute) . http://en.eustat.es.

[7] Euskaltzaindia (Royal Academy of the Basque Language) . http://www.euskaltzaindia.net/index.php?option=com_content&Itemid=1&id=18&lang=en&layout=blog&view=section.

[8] IXA Group. Automatic morphological analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203, 1996.

[9] Koldo Zuazo. *Euskararen sendabelarrak (The medicinal herbes of Basque)*. Alberdania, 2000.

[10] LANGUNE Hizkuntz industrien elkartea Euskal Herrian (The Basque Association of Language Industries). http://www.langune.com/home?set_language=en.

[11] Euskararen erakunde publikoa (Public Office of the Basque Language). http://www.mintzaira.fr.

[12] Amaia Arregi, Alicia Sainz, and José Ramón Ugarriza. PISA 2009 Euskadi. Informe de evaluación (PISA 2009 Euskadi. Evaluation report). http://www.isei-ivei.net/cast/pub/pisa2009/PISA2009-EUSKADI-1INFORME.pdf, 2009.

[13] Hizkuntzaren azterketa eta prozesamendua (Analysis and Processing of Language). https://ixa.si.ehu.es/master/en.

[14] Munduko euskal etxeen ataria (The web of and for the Basque clubs). http://www.euskaletxeak.net/i.

[15] EU policy – to protect and promote regional and minority languages. http://ec.europa.eu/education/languages/languages-of-europe/doc139_en.htm.

[16] Eusko Jaurlaritza (Basque Government). Euskarazko softwarea deskargatzea (Basque software download). http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_softwarea/eu_9567/aurkib.html.

[17] Softkat: Euskarazko software katalogoa (Softkat: Basque Software Catalog). http://softkat.ueu.org.

[18] Language Resources for Basque. http://aclweb.org/aclwiki/index.php?title=Resources_for_Basque.

[19] Hiztegia (Dictionary). http://www.hiztegia.net.

[20] Frantses-Euskara Hiztegi Elektronikoa (French-Basque Electronic Dictionary). http://www.nolaerran.org.

[21] Euskalbar (Basque translator for Firefox). http://euskalbar.eu.

[22] Euskara Institutuaren ataria (Basque Institute's website). http://www.ei.ehu.es.

[23] PuntuEus Association. http://www.puntueus.org/en/.

[24] Jerrold H. Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13, 1994.

[25] XuxenWeb (Spell Checker for Basque). http://www.xuxen.com.

[26] Hizkia, Informatique. http://hizkia.pagesperso-orange.fr.

[27] Terminologia eta Lexikografia Zentroa (Centre for Terminology and Lexicography). http://www.uzei.com.

[28] Google zieht weiter davon (Google moves further away). http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[29] Google rolls out semantic search capabilities. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[30] Verbio speech technologies. http://www.verbio.com.

[31] Indisys: Intelligent dialogue systems. http://www.indisys.es/default.aspx.

[32] Fonetic solutions. http://www.fonetic.es.

[33] Ydilo. http://www.ydilo.com/esp/index.php.

[34] Natural vox. http://www.naturalvox.com.

[35] Aholab. AhoTTS. http://aholab.ehu.es/tts.

[36] Lucy software. http://www.lucysoftware.com.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[38] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

[39] Red Temática en Tecnologías del Habla (Thematic Network on Speech Technologies). http://www.rthabla.es.

[40] Berbatek. http://www.berbatek.com.

[41] European Commission. *Multilingualism: an asset for Europe and a shared commitment*. Brussels, 2008.

[42] Statistics on the Information Society. http://en.eustat.es/estadisticas/opt_0/id_118/ti_Information_Society/subarbol.html#axzz1LTNljBpS.

[43] Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer, and Ralf Klabunde, editors. *Computerlinguistik und Sprachtechnologie: Eine Einführung (Computational Linguistics and Language Technology: An Introduction)*. Spektrum Akademischer Verlag, 2009.

[44] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2009.

[45] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[46] Language Technology World (LT World). http://www.lt-world.org.

[47] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1998.

[48] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

B

META-NETEKO KIDEAK META-NET MEMBERS

Alemania	Germany	Language Technology Lab, DFKI: Hans Uszkoreit, Georg Rehm
		Human Language Technology and Pattern Recognition, RWTH Aachen University: Hermann Ney
		Department of Computational Linguistics, Saarland University: Manfred Pinkal
Austria	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgika	Belgium	Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp: Walter Daelemans
		Centre for Processing Speech and Images, University of Leuven: Dirk van Compernolle
Bulgaria	Bulgaria	Institute for Bulgarian Language, Bulgarian Academy of Sciences: Svetla Koeva
Danimarka	Denmark	Centre for Language Technology, University of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Erresuma Batua	UK	School of Computer Science, University of Manchester: Sophia Ananiadou
		Institute for Language, Cognition and Computation, Center for Speech Technology Research, University of Edinburgh: Steve Renals
		Research Institute of Informatics and Language Processing, University of Wolverhampton: Ruslan Mitkov
Errumania	Romania	Research Institute for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufiș
		Faculty of Computer Science, University Alexandru Ioan Cuza of Iași: Dan Cristea
Eslovakia	Slovakia	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences: Radovan Garabík
Eslovenia	Slovenia	Jozef Stefan Institute: Marko Grobelnik
Espainia	Spain	Barcelona Media: Toni Badia, Maite Melero
		Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: Núria Bel
		Aholab Signal Processing Laboratory, University of the Basque Country: Inma Hernaez Rioja
		Center for Language and Speech Technologies and Applications, Universitat Politècnica de Catalunya: Asunción Moreno
		Department of Signal Processing and Communications, University of Vigo: Carmen García Mateo
Estonia	Estonia	Institute of Computer Science, University of Tartu: Tiit Roosmaa, Kadri Vider

Finlandia	Finland	Computational Cognitive Systems Research Group, Aalto University: Timo Honkela
		Department of Modern Languages, University of Helsinki: Kimmo Koskenniemi, Krister Lindén
Frantzia	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur: Joseph Mariani
		Evaluations and Language Resources Distribution Agency: Khalid Choukri
Grezia	Greece	R.C. "Athena", Institute for Language and Speech Processing: Stelios Piperidis
Herbehereak	Netherlands	Utrecht Institute of Linguistics, Utrecht University: Jan Odijk
		Computational Linguistics, University of Groningen: Gertjan van Noord
Hungaria	Hungary	Research Institute for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics: Géza Németh and Gábor Olaszy
Irlanda	Ireland	School of Computing, Dublin City University: Josef van Genabith
Islandia	Iceland	School of Humanities, University of Iceland: Eiríkur Rögnvaldsson
Italia	Italy	Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Language Technology Research Unit, Fondazione Bruno Kessler: Bernardo Magnini
Kroazia	Croatia	Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb: Marko Tadić
Letonia	Latvia	Tilde: Andrejs Vasiļjevs
		Institute of Mathematics and Computer Science, University of Latvia: Inguna Skadiņa
Lituania	Lithuania	Institute of the Lithuanian Language: Jolanta Zabarskaitė
Luxenburgo	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Malta	Malta	Department Intelligent Computer Systems, University of Malta: Mike Rosner
Norvegia	Norway	Department of Linguistic, Literary and Aesthetic Studies, University of Bergen: Koenraad De Smedt
		Department of Informatics, Language Technology Group, University of Oslo: Stephan Oepen
Polonia	Poland	Institute of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk
		University of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Department of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz University: Zygmunt Vetulani
Portugal	Portugal	University of Lisbon: António Branco, Amália Mendes

		Spoken Language Systems Laboratory, Institute for Systems Engineering and Computers: Isabel Trancoso
Serbia	Serbia	University of Belgrade, Faculty of Mathematics: Duško Vitas, Cvetana Krstev, Ivan Obradović
		Pupin Institute: Sanja Vranes
Suedia	Sweden	Department of Swedish, University of Gothenburg: Lars Borin
Suitza	Switzerland	Idiap Research Institute: Hervé Bourlard
Txekiar Errep.	Czech Republic	Institute of Formal and Applied Linguistics, Charles University in Prague: Jan Hajič
Zipre	Cyprus	Language Centre, School of Humanities: Jack Burston

Hizkuntza-teknologietako 100 bat adituk –META-NETen aurkezten diren herrialde eta hizkuntzetako ordezkariak– Liburu Zurien bildumaren ondorio eta mezurik garrantzitsuenak aztertu eta finkatu zituzten, Berlinen, Alemanian, izandako bilera batean, 2011ko urriaren 21 eta 22an. – About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

META-NETEN LIBURU ZURIEN BILDUMA

THE META-NET WHITE PAPER SERIES

Alemana	German	Deutsch
Bulgariera	Bulgarian	български
Daniera	Danish	dansk
Errumaniera	Romanian	română
Eslovakiera	Slovak	slovenčina
Esloveniera	Slovene	slovenščina
Espainiera	Spanish	español
Estoniera	Estonian	eesti
Euskara	Basque	euskara
Finlandiera	Finnish	suomi
Frantsesa	French	français
Galiziera	Galician	galego
Grekoa	Greek	ελληνικά
Hungariera	Hungarian	magyar
Ingelesa	English	English
Irlandera	Irish	Gaeilge
Islandiera	Icelandic	íslenska
Italiera	Italian	italiano
Katalana	Catalan	català
Kroaziera	Croatian	hrvatski
Letoniera	Latvian	latviešu valoda
Lituaniera	Lithuanian	lietuvių kalba
Maltera	Maltese	Malti
Nederlandera	Dutch	Nederlands
Norvegiera Bokmål	Norwegian Bokmål	bokmål
Norvegiera Nynorsk	Norwegian Nynorsk	nynorsk
Poloniera	Polish	polski
Portugalera	Portuguese	português
Serbiera	Serbian	српски
Suediera	Swedish	svenska
Txekiera	Czech	čeština

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the Basque language. It is part of a series that analyses the available language resources and technologies for 30 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

Europako hiritarrak, enpresak nahiz politikariak eragozpen linguistikoak gainditu beharrean izaten dira egunero-egunero. Hizkuntza-teknologiek aukera ematen dute eragozpen horiek gainditzeko eta, halaber, hainbat teknologia eta ezagupide erabiltzeko interfaze berritzaileak sortzeko. Liburu zuri honek euskararako hizkuntza-teknologien egoera aurkezten du, eta Europako 30 hizkuntzatarako eskuragarri dauden baliabide linguistikoak eta teknologiak aztertzen dituen bilduma baten lehenengo atala da. Europako batzordeak sortutako META-NET Bikaintasun Sareak bultzatu du azterketa hori, eta enpresamunduko, administrazio publikoko, ikerketa-alorreko, alor pribatuko, komunitate linguistikoko eta unibertsitate europarretako parte hartzaileekin lanean diharduten 33 herrialdetako 54 ikerketa-zentroz osatuta dago.

"The Language White Paper Series is an excellent initiative of META-NET, in keeping with our motto 'Give and spread knowledge'. We hope that it will further foster investment in Language Technology solutions for less resourced languages like Basque." – Iñaki Goirizelaia (Rector of the Universidad del País Vasco)

"Europa eleanitzaren testuinguruan, Informazioaren eta Komunikazioaren Teknologiak (IKT) arlo estrategikoa dira hizkuntza guztientzat baina, bereziki, hizkuntza minoritarioentzat. Egun, teknologia horien kontsumitzaileek, Interneti esker, muga geografikoak eta linguistikoak gainditua, aukera paregabea dute IKT produktuak nahi duten hizkuntzan eskuratzeko. Baina horretarako, gure hizkuntza txikiek, ezinbestez, merkatu horretan sartu behar dute. META-NET plataforma egokia da helburu hori erdiesteko."
– Blanca Urgell (Eusko Jaurlaritzako Kultura Sailburua)

www.meta-net.eu