Mendebaldeko euskararen lexikoaren tratamendua, enbor+hondarki banaketan oinarritua, hizketa dialektalaren ezagutze automatikorako

Igor Odriozola, Eva Navas, Jon Sanchez, Iker Luengo, Ibon Saratxaga, Iñaki Sainz, Inma Hernaez

Euskal Herriko Unibertsitatea.
igor, eva, ion, ikerl, ibon, inaki, inma @aholab.ehu.es

Abstract

In this paper a first approach based in the division of the dictionary elements into stems and endings is introduced to deal with Basque dialectal speech recognition. In this way, two objectives are achieved: on the one hand, the great dictionary decrease due to the treatment of the agglutinative grammatical cases of Basque as a finite group of endings; on the other hand, the treatment of the phonetic and phonological variants that show these grammatical cases in the different forms of the western dialect. In this paper, the procedure used in the experiments and the results obtained are shown.

Laburpena

Artikulu honetan, euskarazko hizketa dialektala automatikoki ezagutzeko lehen hurbilketa egin da, hiztegiko hitzak enbor+hondarkitan banatuta. Hurbilketa horrekin, bi abantaila lortzen dira: batetik, sistemaren hiztegia nabarmen txikiagotzea, euskararen kasu gramatikal eranskariak hondarki multzo mugatu gisa tratatzen baitira; bestetik, kasu gramatikalok mendebaldeko euskalkietako hizkera desberdinetan dituzten aldaera fonetiko eta fonologikoak tratatzea. Artikuluan, esperimentuetan jarraitutako prozedura eta lortutako emaitzak aurkeztu dira.

Keywords: Stems and endings, phonetic variations, dialectal ASR, western Basque

Gako hitzak: Enbor eta hondarkiak, aldaera fonetikoak, ASR dialektala, mendebaldeko euskara

1. Sarrera

Euskara hizkuntza aurre-indoeuroparra da, eranskaria eta oso malgukaria, eta, azken proposamen dialektologikoen arabera (Zuazo, 2003), sei euskalki nagusitan bana daiteke. Mendebaldeko euskalkia da, erdialdekoarekin batera, hiztun gehien dituen euskalkia: Bizkaiko eremu euskaldun guztietan, Arabako iparraldean eta Gipuzkoako mendebaldean mintzatzen da.

Euskara estandarizatzeko prozesua 1968an hasi zen; aski berria izaki, euskaldun zahar askok ez dute kode berria barneratu ahozko jardunerako; aitzitik, idatzizkoa erabat hedaturik dago arlo eta eremu guztietara. Horrenbestez, ez da harritzekoa euskalkiak agertzea ahozko ia komunikatze-prozesu guztietan. Gainera, dialektologiako teorikoen artean gero eta hedatuago dago (Zuazo, 2000) nahitaezkoa dela datozen urteetan euskalkien erabilera sustatzea, euskara batua behar bezala osatuko eta garatuko bada.

Euskara batua, zergati literarioak direla eta, erdialderengo euskalkietan oinarritu zen nagusiki, eta ertzetako euskalkiak, hein batean, bazterturik geratu ziren. Hala, mendebaldeko euskalkiak, esaterako, alde nabarmenak ditu batuarekin, batez ere maila morfofonologikoari dagokionez, eta, hortaz, euskara batua ezagutzeko diseinatutako sistema batek oztopo izugarriak izango lituzke euskalki hori (eta haren azpieuskalkiak) ezagutzeko.

Gaur egun, zenbait ekimenen emaitza gisa, euskara batua ezagutzeko diseinatutako zenbait datu-base daude, hala nola *SpeechDat_eu FDB1060* datu-basea (ELRAn[1] eskuragarri). Euskalkiei dagokienez, dauden datu-baseak xede linguistikoetarako diseinatuak dira batez ere, eta, beraz, ez dira egokiak ezagutze-sistema automatikoetan erabiltzeko.

Hori horrela, euskara baturako diseinatutako datu-base bat erabiliz mendebaldeko euskalkiak izango lituzkeen emaitzak aztertzeari ekin diogu. Horretarako, lexiko-sarrerak enbor+hondarkitan banatu dira, hizkuntzaren bi alderdi landu baitaitezke hala: batetik, euskararen beraren ezaugarri den morfema-eransketa handia; bestetik, mendebaldeko euskararen baitako aldaera dialektalak. Aipatu beharra dago mendebaldeko euskalkia ez dela inola ere euskalki homogeneoa eta hainbat azpieuskalkitan sailkatu daitekeela: 8 talde, ezaugarri morfofonologikoetan oinarrituta (Gaminde, 2000).

[1] catalog.elra.info/product_info.php?products_id=5

2. Mendebaldeko euskararen ezaugarri fonologiko nagusiak

Atal honetan, euskararen mendebaldeko euskalkiaren ezaugarri morfofonologiko nabarmenenak aurkeztu dira. Aurreko atalean aipatu denez, mendebaldeko euskalkia da aldaera gehien dituen euskalkia eta, hortaz, konplexuena, gainerako euskalkietan agertzen ez diren arau fonologiko eta aldaera fonetiko asko baititu.

2.1. Euskararen egitura morfologiko orokorra

Euskara postposiziozko hizkuntza eranskaria da, flexio-sistema aberatsekoa, bai aditzei, bai kasu-sintagmei dagokienez. Euskal aditza oso-oso malgukaria eta, beraz, konplexua da; hain konplexua, ezen, aditz-laguntzaileak esaterako, forma zatiezin gisa tratatzen baitituzte hizkuntza naturalaren prozesamenduan diharduten zenbait taldek (Alegria et al., 1996). Dena dela, euskararen morfologiaren deskribapena oso gai zabala da, eta, beraz, kasu-sintagma besterik ez da azalduko hemen, horixe baita artikulu honetan baliagarri gertatuko zaiguna.

Kasu-sintagmak honako eskema honen arabera eraikitzen dira (parentesi arteko elementuek ez dute zertan agertu):

Lema + (artikulua) + (numeroa) + kasua(k)

17 kasu ditu euskarak, eta haietariko zenbait kateaturik ager daitezke. Hortaz, askoz handiagoa da lema bakoitzak izan dezakeen hondarki kopurua.

2.2. Artikulu singularraren egokitzapenaren egitura morfofonologikoa

Artikulu honetan landu den ezaugarririk garrantzitsuena da hori. «*a*» artikulu singularra da lemari eransten zaion lehendabiziko morfema. Hala, artikulua lemari egokitzeko moduaren arabera eransten dira ondorengo atzizkiak. Euskara batuan, artikuluaren eransketak emaitza bakarra du; euskalkietan, ostera, asimilazio- eta disimilazio-prozesuen ondorioz —eta hala sortzen diren *feeding* prozesuen ondorioz (Kiparsky, 1968)—, lemaren azken fonemaren araberakoa da. Egokitzapena, gainera, ez da homogeneoa mendebaldeko euskalkian, eta, arestian aipatu bezala, 8 azpieuskalkitan bana daiteke eremu hori, artikuluaren egokitzapena kontuan izanda.

1. taulan, "azken fonema + artikulua" multzo bakoitzaren emaitzak bildu dira, bai eremu azpidialektal bakoitzerako, bai euskara baturako (fonemak adierazteko erabili diren ikurrak SAMPAkoak dira —*Speech Assessment Methods Phonetic Alphabet*—; ikus http://aholab.ehu.es/sampa_basque.htm). Adibide gisa, «*-e+a*» multzoak hau adierazten du: «*e*» fonemaz amaitutako lemak eta «a» artikuluak osatutako multzoa («*C*» ikurrak kontsonantea adierazten du).

	-a+a	-e+a	-i+a	-o+a	-u+a	-C+a
A	[ea]	[ea]	[ie]	[oa]	[ue]	[Ca]/[Ce]
B	[ia]	[ia]	[ie]	[oa]	[ue]	[Ca]/[Ce]
C	[ie]	[ie]	[iZe]	[oa]	[ue]	[Ca]/[Ce]
D	[ia]	[ia]	[iZe]	[oa]	[ue]	[Ca]/[Ce]
E	[ie]	[ie]	[iZe]	[ue]	[ue]	[Ca]/[Ce]
F	[i]	[i]	[iZe]	[u]	[u]	[Ca]/[Ce]
G	[ia]	[ia]	[iZa]	[ua]	[ua]	[Ca]
H	[e]	[e]	[i]	[o]	[u]	[Ca]/[Ce]
Estandarra	[a]	[ea]	[ia]	[oa]	[ua]	[Ca]

1. taula: "Azken fonema + artikulua" multzo desberdinen ebakera fonetikoak, eremu bakoitzean (A-H) eta estandarrean

Zenbait lema+artikulu multzotan, ez da estandarrari dagokion emaitza (aldaerarik gabea) ageri; esate baterako, «-a + a» kasuak bost ebakera desberdin ematen ditu zortzi eremuetan, baina bat ere ez da estandarrekoa: «neska+a». Eremuaren arabera, «neskea», «neskia», «neskie», «neski» eta «neske» formak hartzen ditu; aitzitik, inoiz ere ez batuko «neska» (Oñederra, 2005). Aipatzekoa da, gainera, artikuluaren egokitzapenak, artikuluaren ebakerari ez ezik, lemaren azken fonemaren ebakerari ere eragiten diola (are gehiago, «-i + a» kasuan kontsonante bat txertatzen zaio), eta, horren ondorioz, alomorfoak sortzen dira, bai lemetarako, bai artikulurako (euskara estandarrean, -/a/ fonemaz amaitzen diren lemei artikulu plurala egokitzean baino ez da sortzen fenomeno hori). 2. eta 3. tauletan, lemen eta artikuluaren alomorfoak ageri dira, hurrenez hurren.

Lemaren azken fonema	Alomorfoak
-/a/	-[e], -[i]
-/e/	-[e], -[i]
-/i/	-[i]
-/o/	-[o], -[u]
-/u/	-[u]
-/C/	-[C]

2. taula: Artikulua lemari egokitzean sortzen diren lema-alomorfoak

Eremua	Artikuluaren txandakatze fonetikoak
A	[a] ~ [e]
B	[a] ~ [e]
C	[a] ~ [e] ~ [Ze]
D	[a] ~ [e] ~ [Ze]
E	[a] ~ [e] ~ [Ze]
F	[a] ~ [e] ~ [Ze] ~ [Ø]
G	[a] ~ [Za]
H	[a] ~ [e] ~ [Ø]
Estandarra	[a]

3. taula: Artikulua lemari egokitzean sortzen diren artikulu-alomorfoak

Lemei dagokienez, aipatzekoa da -/a/, -/e/ eta -/o/ fonemez amaitzen diren lemek bina alomorfo dituztela; gainerakoek, bana (euskara batuan, alomorfo bakarra dute guztiek). Artikuluari dagokionez, berriz, F eremuan lau ebakera desberdin ageri dira; eremu bakoitzean, gutxienez, bi gauzatze fonetiko sortzen dira (euskara batuan, gauzatze bakarra).

2.3. Aditz-partizipioak

Euskaraz, aditz asko (partizipio-forman) ondoko hizkuntzetatik hartu dira, bereziki hizkuntza erromantzeetatik eta, gaur egun, gaztelaniatik eta frantsesetik. Erromantzeetatik mailegatzeko modua oso emankorra izan zen: «-ado»z amaitutako aditz-partizipioek, zenbait prozesu fonetikoren ondorioz, gauzatze fonetiko hauek dituzte egun: -[atu], -[au], -[eu], -[a]; «-ido» motako partizipioek, berriz, honako hauek: -[itu], -[idu], -[iu]. Ebakera bakoitza eremuaren araberako da.

Artikulu honetan, argigarri gisa, *Aditz berriak* izena jarri diegu ondoko hizkuntzetatik mailegatutakoei; *Aditz zaharrak*, berriz, mailegatuak ez direnei eta, hortaz, halako prozesurik jasan ez dutenei.

2.4. Beste zenbait aldaketa fonetiko

Mendebaldeko euskalkian, euskara estandarrean kontuan hartu ez diren hainbat aldaketa fonetiko gertatzen dira; esaterako, gaur egun, mendebaldeko euskalkian, desagertu egin da garai bateko /s'/ fonema igurzkari hobi-atzeko ahoskabea; fonema haren lekua /s/ fonema igurzkari hobikari ahoskabeak hartu du. Hala, bai /s'/ fonemak, bai /s/ fonemak [s] ebakera dute mendebaldeko hizkeretan.

Palatalizazioaren fenomenoa ere oso ugaria da mendebaldeko euskalkian: /i/ fonemaren ondoko kontsonante asko sabaikaritzen dira, eta gainerako euskalkietan agertzen ez diren ebakerak ere sortzen dira hala.

Aldaketa mota horiek, funtsean, fonemaz fonemako aldaketak dira.

3. Erabilitako datu-baseen deskripzioa

Sarreran azaldu denez, lan honen helburua da ezagutze-sistema entrenatzeko euskara estandarrerako diseinatutako datu-base bat erabiltzea eta testa hizketa dialektalaz egitea. Datu-base hauek hautatu dira horretarako: *SpeechDat_eu* (Hernáez et al. 2003) eta *Bizkaifon* (Castelruiz et al., 2004).

Lehen datu-basea euskara baturako diseinaturik dago, telefonia finkoan jarduteko, eta horixe erabili da erreferentziazko eredu akustikoak eraikitzeko. Bigarrena, berriz, erabat dialektala da, eta horixe baliatu da hurrengo atalean azalduko diren esperimentuak egiteko. Bataren eta bestearen artean alde nabarmenak daude, 4. taulan ikus daitekeenez.

	SpeechDat_eu	*Bizkaifon*
Hizketa mota	Euskara estandarra	Mendebaldeko euskalkia
Kanala	Telefonia finkoa	Mikrofonoa (zinta digitala eta magnetikoa)
Lagintze-maiztasuna	8 kHz	8, 16 eta 32 kHz
Audio-fitxategiak	5.200	11.868
Lexikoia	3.968 hitz estandar	8.199 forma dialektal (5.224 estandar)
Mailegu ez-onartuak	0 (% 0)	993 (% 12,38)
Hizlarien generoa	Giz.: % 45,28 Emak.: % 55,72	Giz.: % 4,70 Emak.: % 95,30

4. taula: SpeechDat_eu eta Bizkaifon datu-baseen arteko alderik nabarmenenak

SpeechDat_eu datu-baseak 1.060 hizlariren grabazioak ditu (480 gizonezko eta 580 emakumezkorenak), telefonia finkoko sarearen gainean grabatuak, eta *SpeechDat* proiektuaren zehaztapenak betetzen ditu (Winski, 1997). Datu-basean, nahiz oso kopuru txikian, badira hizketa dialektala ezagutzeko berariaz egindako grabazioak ere. Hego Euskal Herriko euskararen soinuen inbentario osoaren laginak izateko garatu zen atal hori; izan ere, [Z] alofonoa, igurzkari aurresabaikari ahostuna, aurreko atalean azaldutako arau morfofonologikoen emaitza gisa baino ez da sortzen (ikus 1. taula), eta ez dago euskara batuaren soinu-inbentarioaren barnean. *SpeechDat_eu* datu-basearen lexikoian, ahoskera alternatibo gisa jotzen da.

Bizkaifon datu-basean, hitz solteen 11.868 audio-fitxategi daude. 4. taulan ageri denez, hizlari gehien-gehienak emakumezkoak dira, eta grabazioak era askotara egin dira (hainbat motatako mikrofonoak, zinta magnetikoa, zinta digitala eta abar erabiliz). Gainera, lagintze-maiztasun desberdinez digitalizatu dira, eta, beraz, esperimentuei ekin aurretik, 8 kHz-era egokitu da fitxategi guztien lagintze-maiztasuna.

Bizkaifon datu-baseko fitxategi bakoitzak forma dialektalaren transkripzio ortografikoa eta estandarrari dagokion transkripzio ortografikoa ditu. Aurkitu dugun lehen arazoa izan da euskara estandarrean onartzen ez diren hainbat hitz dialektal daudela datu-basean, batez ere mailegu aski berriak direlako eta tokiko erabilera dutelako; esaterako, «abuelie» hitza, gaztelaniazko «abuela»tik mailegatua. Auzi horri beste azterketa sakonago bat dagokiolakoan eta aldaera fonologikoek sortzen duten arazoetan sakontzeko asmoz, mailegu ez-onartuak dituzten fitxategiak datu-basetik kentzea iritzi da, guztira 1.275 fitxategi. Hala, fitxategi kopurua 10.593ra jaitsi da. Fitxategi horietan 7.029 forma dialektal desberdin daude, estandarreko 4.721 hitzi dagozkienak. Gainera, 4.721 forma estandar horietarik

1.143 hitzek baino ez dute bat egiten forma dialektalen batekin.

Datu-basea morfologikoki etiketatu da, prozesu erdi-automatiko bati jarraituz: euskara estandarrerako analizatzaile morfologiko bat erabili da (Ezeiza, 1998), euskara batuko transkripzioekin batera, eta, halaxe lortu dira forma dialektaletarako etiketa morfologikoak. Horren ondoren, eskuz aztertu dira etiketak. *Bizkaifon* datu-basearen banaketa morfologikoa 5. taulan ageri da.

Kategoria gramatikala	Fitxategi kopurua	Forma dialektalen kopurua
Izenak + adj (art. sg.)	8.239	5.466
-a	1.874	1.508
-e	898	681
-i	1.290	826
-o	1.252	889
-u	746	474
-C	2.179	1.088
Izenak + adj (art. pl.)	345	253
Aditzak (partizipioak)	1.435	859
Aditz berriak	510	267
Aditz zaharrak	925	592
Bestelakoak	574	451
Guztira	10.593	7.029

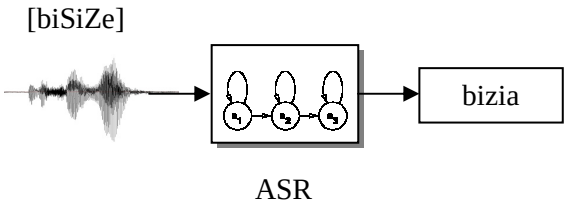
5. taula: *Bizkaifon* datu-basearen edukia, ikuspegi morfologikotik

Izenak eta adjektiboak talde berean sailkatu dira, hitz solteen testuinguruan jokaera berbera baitute. Talde hori da, kopuruari begiratuz gero, datu-basearen bihotza, eta, taulan ikus daitekeenez, talde horretako hitz gehienek artikulu singularra dute erantsirik. Artikulu plurala duten izen eta adjektiboen multzoa aparte jarri da, oso aldaketa fonetiko gutxi baitute. Bigarren multzorik handiena aditz-partizipioez osatutakoa da, bai aditz berriez, bai aditz zaharrez (ikus 2.3 atala).

4. Esperimentuak

Atal honetan aurkeztu diren esperimentuen helburua da mendebaldeko euskarazko hitz solte edo isolatuak ezagutzeko sistema bat lortzea euskara estandarrerako entrenatutako eredu akustikoak erabiliz. Hortaz, sistemaren irteera euskara batuan izango da, sarrerako aldaera dialektala zeinahi delarik ere (ikus 1. irudia).

Lan honetan gauzatu diren ezagutze-prozesu guztiak *HTK Speech Recognition Toolkit* softwareaz egin dira (Young et al., 2000). Atal honetan, lehendabizi, euskara estandarra erabiliz egindako atariko esperimentu bat azaldu da; ondoren, transkripzio dialektalez osatutako hiztegi dialektalaz egindako esperimentu bat; 4.2 atalean, lexikoaren tratamendua azaldu da, enbor+hondarkitan banatzean oinarritua; azkenik, 4.3 atalean, emaitzen analisi bat egin da.



1. irudia: Ezagutze-sistemaren sarrera (dialektala) eta irteera (estandarra)

4.1. Atariko esperimentua

Mendebaldeko euskalkiko hitz isolatuak ezagutzeko sistema garatzeko abiapuntu gisa, euskara baturako diseinatutako *SpeechDat_eu* datu-baseaz egindako ebaluazio-esperimentuak hartu ditugu kontuan. (Hernáez et al., 2003)n deskribatutako esperimentuan, % 17,20ko WERa (*Word Error Rate*) eman zuen hitz fonetikoki aberatseko azpicorpusaren ebaluazioak. Sistemaren hiztegia 3.968 sarrera lexikalez osatua zen, eta hitz isolatuak baino ez ziren kontuan hartu. Eredu akustikoetan euskara estandarreko soinuak soilik zeuden modelatuta, eta Markoven eredu ezkutuetan oinarrituak ziren (3 egoerakoak eta 32 gaussiarrekoak). Erabili ziren parametro akustikoak hauek izan ziren: lehendabiziko 13 MFCC koefizienteak eta haien lehen eta bigarren deribatuak (guztira, 39 parametro, 10 ms-an behin eta 25 ms-ko trama-zabalerarako lortuak).

Estandarreko 4.721 lexemez (euskara batuko transkripzio ortografikoak) eta erregelatan oinarritutako G2P grafema-fonema transkibatzaile bat baliatuz, euskara estandarreko hiztegi bat sortu da, SAMPA alfabeto fonetikoa erabiliz. G2P bihurgailu hori Euskaltzaindiak ahoskeraren inguruan emandako arauetan oinarriturik dago (Euskaltzaindia, 1998).

*Bizkaifon*eko audio-fitxategiak euskara estandarrean egindako hiztegiaz eta *SpeechDat_eu*ko fitxategietatik sortutako eredu akustikoez testatuta, % 46,27ko WERa lortu da. Oso kaxkarra da emaitza, baina espero izatekoa, honako kontu hauengatik:

- Lexikoi dialektalak 7.029 forma desberdin ditu, eta 1.143 formak (% 16,26k) soilik egiten dute bat haren estandarreko baliokidearekin.
- Mendebaldeko euskalkian /Z/ soinua dago (igurzkari aurresabaikari ahostuna), eta *Bizkaifon*en maiztasun handia du. Soinu hori ez zen modelatu *SpeechDat_eu* datu-basetik eredu akustikoak sortzean; haren eredu akustikorik ez dago, beraz.
- Datu-base bien desberdintasun akustiko nabarmenak. Eragin esanguratsua du horrek, zalantzarik gabe, emaitzetan.

Faktore horiek esperimentuetan duten eragina hobeto ezagutzearren, ezagutza 'dialektaleko' esperimentu bat egin da: lehendabizi, hiztegi dialektal bat sortu da 7.029 forma dialektalak lexema gisa jarrita eta lexemaren transkripzio fonetikoak laborategiko G2P transkribatzaileaz lortuta. Gainera, /Z/ soinuaren eredu akustikoa sortu da *SpeechDat_eu*ko fitxategietatik abiatuta, nahiz eta 71 lagin soilik izan. Azkenik, *Bizkaifon*eko 10.593 audio-fitxategiak testatu dira.

Ezagutze dialektal horretan, % 23,14ko WERa lortu da. Hiztegia egiteko, datu-baseko forma dialektalak soilik erabili dira; beraz, emaitza horrek hurrengo esperimentuetarako goiko muga teorikoa adierazten du. Horrenbestez, datu-base bien arteko desberdintasun akustikoak nabarmenak diren arren, emaitza ez da batere etsigarria.

Akatsen analisia osatzeko asmoz, beste esperimentu bat burutu da, *SpeechDat_eu* datu-baseko eredu akustikoak *Bizkaifon*eko audio-fitxategiez egokituta. Trifonema-ereduak bost aldiz entrenatu dira, % 2,66ko WER egonkorra lortu den arte. Emaitza horrek egiaztatzen du datu-base bien arteko desberdintasun akustikoak oso esanguratsuak direla. Hala eta guztiz ere, hurrengo esperimenturako, jatorrizko eredu akustikoak baino ez dira erabili.

4.2. Lexikoa enbor+hondarkitan banatuz egindako esperimentua

Hitz solte edo isolatuak ezagutzeko testuinguruan, zilegi da aldaera fonetiko eta morfofonologiko dialektal guztiak kontuan hartzen dituen hiztegi bat erabiltzea, baina, hizketa jarraitua ezagutzeko sistema bat garatuko bada, behar-beharrezkoa da hitza baino unitate txikiagoak erabiltzea, batez ere hizkuntza eranskarien kasuan.

Euskararen izaera eranskariak kasu-sintagmak bitan zatitzeko aukera ematen du (baita aditzak multzo txikiagotan zatitzeko aukera ere, baina, lan honetan, partizipioak baino ez dira kontuan hartu); hala, enbor+hondarki gisa tratatu dira hitz gehienak. Euskaraz, 17 kasu gramatikal daude 4 testuinguru numerikotan (singularra, plurala, mugagabea eta plural hurbila), 68 hondarki, guztira. Gainera, bigarren mailan kateatuz gero, 275 hondarki sortzen dira, eta aski ohikoa da, gainera, hori (Alegria et al., 1996). Kontuan izanda mendebaldeko hizkeran artikuluak 4,17 alomorfo dituela izen-lemako, batez beste 1.146,75 hondarki fonetikoki desberdin izango genituzke mendebaldeko euskararako. Horrexegatik iruditzen zaigu hain garrantzitsua hitza baino unitate txikiagoak erabiltzea euskararako eta, bereziki, mendebaldeko euskalkirako.

Lema flexionatzen denean inoiz aldatzen ez den zatia enborra dela kontsideratuz, -/a/, -/e/ eta -/o/ fonemez amaitzen diren lemetan enborrak ez luke lemaren azken fonema edukiko (ikus 2. taula); gainerako kasuetan, bat dira lema eta enborra. 6. taulan, lan honetarako baliatu diren enbor eta hondarkiak ageri dira, artikulu singularra duten izen eta adjektiboen multzorako; adibidez, «neska» lema -/a/ fonemaz amaitzen da, eta, hortaz, enborra «nesk» da, eta hondarkiak, berriz: -«ea», -«ia», -«ie», -«i», -«e». Hala, 2.2 atalean azaldu diren aldaera guzti-guztiak lortzen dira.

Lemaren azken fonema	enborra	hondarkia	
		Mend. euskara	Bat.
-/a/	lema - /a/	-[ea],-[ia],-[ie],-[i],-[e]	-[a]
-/e/	lema - /e/	-[ea],-[ia],-[ie],-[i],-[e]	-[ea]
-/i/	lema	-[Ø],-[Za],-[Ze],-[e]	-[a]
-/o/	lema - /o/	-[oa],-[ua],-[ue],-[u],-[o]	-[oa]
-/u/	lema	-[Ø],-[a],-[e]	-[a]
-/C/	lema	-[a],-[e]	-[a]

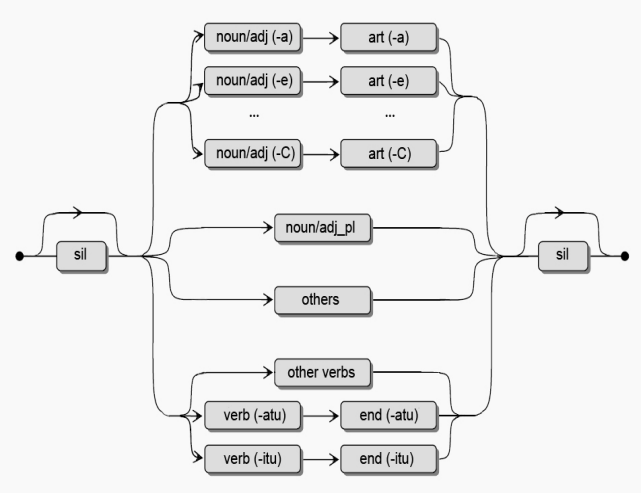
6. taula: Izen eta adjektiboetarako enborrak eta hondarkiak

Bizkaifon datu-basea 7.029 forma dialektal desberdinez osaturik dago, eta haietatik 5.733 banatu dira enbor+hondarkitan. 7. taulan, datu-basearen eduki morfologikoa ageri da, xehetasun handiagoz. Halaber, zein hitz zatitu den eta zein ez ere adierazi da.

Kategoria gramatikala	Kopurua	Numeroa	Kopurua	Zatitua?
izenak	5.098	singularra	4.854	B
		plurala	137	E
		mugagabea	107	E
adjektiboak	621	singularra	612	B
		plurala	5	E
		mugagabea	4	E
aditzak	859	partizipio berria	267	B
		partizipio zaharra	592	E
adberbioak	281			E
postposizio-sintagmak	124			E
bestelakoak	46			E

7. taula: *Bizkaifon* datu-basearen eduki morfologikoa, xehe-xehe

Enbor bakoitzaren ondoren hari dagokion hondarki alomorfikoen multzoa soilik kateatu ahal izan dadin, 2. irudian adierazi den gramatika sortu da. Gainera, hitz-barneko fonema-testuinguruak ere izan dira kontuan, enborretik hondarkirako trantsizioak ahalik eta ondoen modelatzeko (Rotovnik et al., 2007).



2. irudia: Enbor+hondarki multzo egokiak sortzeko erabilitako gramatikaren eskema

Bestalde, sistemaren hiztegian, 2.4 atalean azaldutako aldaera fonetikoak ere txertatu dira alternatiba gisa, fonemaz fonemako aldaketa horien eragina ahalik eta txikiena izan dadin.

4.3. Emaitzak

Lortu diren emaitzak 8. taulan ageri dira, hitz barneko trantsizioko testuingurua kontuan izanda eta kontuan izan gabe. Azken lerroan azaltzen denez, testuingurua erabiltzeak nabarmen hobetzen ditu emaitzak; hala, lortzen den WERa % 28,82 da. Argi dago, beraz, hasierako % 46,27 hartatik asko hobetu dela sistema eta nabarmen hurbildu dela 4.1 atalean ezarri dugun % 23,14ko muga teorikora. Emaitza horren eta muga teorikoaren arteko desberdintasuna arrazoi hauek azaldu lezakete: batetik, hainbat aldaera fonetiko ez dira modelatu, garrantzitsuenak besterik ez; bestetik, hiztegia handitzeak sistemaren nahasmena handiagotzea dakar, sistemak enborraren aukera guzti-guztiak aztertzen baititu, ez bakarrik datu-basean dauden konbinazioak. Horrek, neurri batean, kaskartu egiten du sistemaren eraginkortasuna.

	Testuingurua kontuan izan gabe (%)	Testuingurua kontuan izanda (%)
Enborren ERa	41,64	25,77
Hondarkien ERa	19,20	9,43
Enbor zuzenetako hondarkien ERa	1,37	1,65
Zatitu ez diren lexemen ERa	27,22	37,15
WER osoa	39,77	28,82

8. taula: Lexema enbor+hondarkitan zatituz egindako esperimentuaren emaitzak

Taularen lehen eta bigarren lerroetan, hurrenez hurren, enborren eta hondarkien errore-tasa (ER, *Error Rate*) ageri da.

Hirugarren lerroan, ondo ezagututako enborra duten baina gaizki ezagutu diren hondarkien errore-tasa jaso da. Oso zifra txikia da; izan ere, lema multzo bakoitzari estandarreko hondarki bakarra dagokio irteeran (izenak eta aditzak enbor bera dutenean izan ezik).

Taulako laugarren lerroan, zatitu ez diren hitzen errore-tasa ageri da. Nabaria da hitz barneko testuingurua erabiltzeak nabarmen jaisten duela enborren nahiz hondarkien errore-tasa, baina emaitza txarragoak ematen ditu zatitu ez diren hitzetarako; izan ere, testuingurua kontuan hartzeak nahasmena eragiten du zatitu gabeko hitzak ezagutzeko unean. Gainera, erroreak analizatuz, ikusi dugu zatitu gabeko hitz gehienak aditz zaharretan gertatzen direla, oso hitz laburrak baitira (4 edo 5 fonemakoak).

5. Emaitzen azken ondorioak

Lan honetan, euskara estandarreko datu-base akustiko bat erabili da mendebaldeko euskalkiko hizketa ezagutzeko. Atariko esperimentuak argi erakusten duenez, euskalki horrek berariazko tratamendua behar du, batetik euskararen izaera eranskaria eta, bestetik, mendebaldeko euskalkiko aldaera fonetiko nabarmenak kontuan izateko.

Sistemaren hiztegiko lexemak enbor+hondarkitan banatuta, lortzen da hiztegia hondarki kopuruaren neurri berean ez haztea (gogoan izan euskara batuan 275 hondarki daudela bigarren kateatze-mailan) eta, orobat, datuak askoz azkarrago prozesatzea. Bestalde, prozedura horrek hondarkien alomorfoak modelatzeko aukera ematen du (artikulu singularrerako, 4,17 alomorfo daude izen-lemako), eta, hala, % 17,45eko hobekuntza lortzen da eta WERaren muga teorikora nabarmen hurbiltzen (% 23,14, ikus 4.1 atala). Teorikoki, % 5,66 dago oraindik ere hobetzeko, baina bi faktore izan behar dira kontuan: batetik, nahasmena: hitzak zatitzearen ondorioz, sistemak enbor bakoitzaren aukera guzti-guztiak hartzen ditu kontuan, eta, hortaz, jaitsi egiten da, pittin bat, sistemaren eraginkortasuna; bestetik, esperimentu honetan kontuan hartu ez diren hainbat aldaera fonetiko daude. Lehen faktorearen ondorioz, praktikan beti izango da errore-tarte bat muga teorikora iristea eragotziko duena.

Esperimentuetan eragin handia izan duen beste elementu garrantzitsu bat datu-baseen arteko desberdintasun akustikoak izan dira. Horixe da, atariko esperimentuan egiaztatu denez, muga teorikoa % 23,14an izatearen zergatia.

Lan honen hurrengo pausua da hemen aurkeztutako prozedura hizketa jarraitura hedatzea. Euskara batuak ez ditu arau asko ebakerari dagokionez, eta, hein handi batean, oso egoera formaletan baino ez da mintzatzen; oso ohikoa da bai irratian, bai telebistan, hizketa dialektala entzutea. Euskalkiak zein euskara batua

kontuan hartuko lukeen sistema bat garatuz gero, nabarmen hobetuko litzateke ezagutze-sistemen eraginkortasuna.

Ildo horretatik jarraitzeko arazorik handiena datu-base dialektalik eza da; hortaz, oso interesgarria deritzogu lan honetan proposatu dugun sistema egokitzea, irteera dialektal bat izateko. Hala, transkribatzaile erdi-automatiko bat izango genuke (eskuzko zuzenketez lagundurik, betiere) eta datu-base berriak sortu ahal izango lirateke, hala lortutako testu dialektalez modu estatistikoan ezagutza erauzteko teknikak ere aztertu ahal izateko.

6. Esker onean

Lan honek, zati batean bada ere, Eusko Jaularitzaren dirulaguntza jaso du, BERBATEK kodepean eta Zientzia eta Berrikuntza Ministerioarena, BUCEADOR (TEC2009-14094-C04-02) kodepean.

7. Aipamenak

Alegria, I.; Artola, X.; Sarasola, K.; Urkia, M. (1996). Automatic morphological analysis of Basque. *Literary & Linguistic Computing* 11(4):193-203.

Castelruiz, A.; Sánchez, J.; Zalbide, X.; Navas, E.; Gaminde, I. (2004). Description and Design of a WEB Accesible Multimedia Archive. *Proceedings of 12th IEEE Mediterranean Electrotechnical Conference, MELECON*, 681-684 orr., Dubrovnik.

Euskaltzaindia (1998). Euskara batuaren ahoskera zaindua. *Euskaltzaindiaren Arauak*, 805-808 orr., Bilbao.

Ezeiza, N.; Aduriz, I.; Alegria, I.; Arriola, J.M.; Urizar, R. (1998). Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *Proceedings of COLING-ACL'98*, 379-384 orr., Montreal.

Gaminde, I. (2002). Bizkaiko euskararen ezaugarri fonologiko batzuen inguruan. *Euskalingua* 1:4-14.

Hernáez, I.; Luengo, I.; Navas, E.; Zubizarreta, M.; Gaminde, I.; Sánchez, J. (2003). The Basque Speech_Dat (II) Database: A Description and First Test Recognition Results. *Proceedings of EUROSPEECH*, 1549-1552 orr., Geneva.

Kiparsky, P. (1968). Linguistic Universals and Linguistic Change. *Universals in Linguistic Theory*, 170-202 orr., eds. Bach and Harms.

Oñederra, M. L. (2005). Fonologiaren mugak: alabea eta birjinak elexan. *Nerekin yaio nun. Txillardegiri omenaldia*, 379-397 orr., IKER 17, Bilbao: Euskaltzaindia.

Rotovnik, T.; Sepesy Maucec, M.; Kacic, Z. (2007). Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Communication* 49(6):437-452.

Winski, R. (1997). Definition of corpus, scripts and standards for Fixed Networks. LE2-4001-SD1.1.1, Txosten teknikoa.

Young, S.; Kershaw, D.; Odell, J.; Ollason, D.; Valtchev, V.; Woodland, P. (2000). *The HTK Book*, Cambridge University, Cambridge.

Zuazo, K. (2003). *Euskalkiak, Herriaren lekukoak*. Donostia: Elkar argitaletxea.

Zuazo, K. (2000). *Euskararen sendabelarrak*. Irun: Alberdania argitaletxea.