Ortografia-erroreak eta konpetentzia-erroreak Webeko euskarazko testuetan

Izaskun Etxeberria, Iñaki Alegria
IXA taldea. Donostia (UPV/EHU)

Igor Leturia
Elhuyar Fundazioa. Usurbil

Laburpena: Lan honetan euskarazko ortografia-erroreen azterketa egin dugu webetik jasotako dokumentuekin osatutako hainbat corpusetan (testu-bildumetan), eta horrela corpus horien kalitatea estimatu dugu. Metodologia finkatzeko, ingeleserako eta alemanerako egin den antzeko lanean oinarritu gara (Ringlstetter et al., 2006), baina, euskararen ezaugarriak direla eta, ez dugu teknologia bera erabili erroreak identifikatzeko.

Euskarak morfologia aberatsa duenez, erroreak identifikatzeko berrerabili egin ditugu aurretik garatutako ortografia-zuzentzaileak. Bide horretatik, detekzioaren estaldura handiagoa da eta, gainera, prozesuaren garapena azkarragoa izan da berrerabilpena dela-eta. Horrekin batera, posible da ia automatikoki halako tresnak dituzten beste hizkuntzetan metodo bera erabiltzea. Analisiaren emaitzak balio dezake zuzentasunaren araberako testuen sailkapena egiteko, eta bide batez, aukera ematen du gutxieneko kalitate bat ez duten testuak baztertzeko.

Abstract: This work's aim is to analize the quality of corpora retrieved from the Basque Web. The main objective of the analysis is to detect documents containing errors over a given threshold. The second objective is to estimate how many errors are of each class. The metodology followed is similar to that used for English and Germany in Ringlstetter et al. (2006). The main difference lies in the fact we reuse spelling checkers for detecting errors. By this way, we obtain a higher error coverage and we can also develop the work faster because of reusing previous tools.

1. SARRERA, HELBURUAK ETA ERLAZIONATUTAKO LANAK

Egun maiz osatzen dira testu-bildumak Internetetik jasotako hainbat dokumenturekin (*Web as a corpus-WAC*), Internetek aukera paregabea eskaintzen baitu corpus handiak eta egungoak osatzeko.

Corpus deritzen testu-bilduma horiek terminologia edo hitz-hurrenkera bezalako hitzkuntza-ezaugarriak analizatzeko erabiltzen dira, baina

hala osatutako corpusen kalitatea bermatzen ez bada, emaitzak desegokiak suerta daitezke.

Testu-bilketa handia eta automatikoa egiten denean, beti dago haien artean eragozpenak sortzen dituztenak, hainbat arrazoi direla-eta: formatua, transkripzioak, dialektoen erabilera, kaleko erregistroa, beste hizkuntzetan dauden atalen tartekatzea, hizkuntzaren identifikazio okerra eta abar. Arazo horietako batzuk identifikatzen dira Kilgarriff eta Grefenstette (2009) lanean.

Hemen aurkezten den lanaren testuingurua, beraz, hauxe da: weba erabili datu-bilduma handia eta egungoa lortzeko, gero datu horiekin euskarazko tresnak analizatu eta garatu ahal izateko. Beraz, web dokumentuak iragazteko tresna lortzea da aurkezten dugun lanaren helburu nagusia; tresna horren bitartez dokumentuen gutxieneko kalitate bat bermatuko dugu.

Bide batez, askotariko dokumentuak analizatuko direnez, errore mota bakoitzaren presentzia estimatu nahi da dokumentuetan.

Webeko dokumentuen bilaketa euskaraz egitea ez da erraza. Hori delaeta, lankidetzara jo dugu eta *Web as corpus* gaiaren inguruan Leturia et al. (2008) lanetan lortutako emaitzen azpimultzo bat erabili dugu gure corpusak osatzeko. Dokumentu horien zuzentasuna aztertu nahi dugu haien artean sailkapen bat egiteko; gainera, finkatuko dugu horietatik zeintzuek sor ditzaketen eragozpenak beste ikerketetan erabiliz gero.

Lanaren metodologia finkatzeko unean, Ringlstetter et al. (2006) lana hartu dugu erreferentziazko lan gisa. Lan horretan, ortografia-erroreen azterketa sakona egin dute ingeleserako eta alemanerako. Hainbat corpus analizatu eta gero, ondorioztatzen dute, beste gauzen artean, zer erroretasa tarteak erabil daitezkeen dokumentuak lau multzotan sailkatzeko (oso onak, onak, txarrak, baztertzekoak). Alde batetik, gaiaren arabera antolatzen dituzte corpusak: gai orokorreko corpusak eta gai espezifikoetako corpusak, eta beste alde batetik, dokumentuen jatorrizko formatuaren arabera antolatzen dituzte: HTML formatuko dokumentuen corpusak eta PDF formatuko dokumentuen corpusak. Era horretara, ikusi ahal izango dugu ezaugarri horiek eragina ote duten erroreen banaketan.

Erroreen tipologiari dagokionez, erreferentziazko lan horretan errore tipografikoak, ezagutzakoak, OCR erroreak zein kodeketa-erroreak analizatzen dituzte.

Bilaketa-sistema bat diseinatu ondoren, abian jartzen dute horrela hainbat testu mota lortzeko. Gero, testu horiek tokenizatu[1] egiten dituzte eta baita «garbitu» ere; izan ere eragozpenak sortzen dituzten hainbat hitz baztertzen

[1] Tokenizatu: tetu bat hitzetan banatzeari esaten zaio.

dituzte analisirako: letra larriz hasten diren hitzak, oso motzak direnak, zenbakiak eta abar. Aldi berean, errore-hiztegiak sortzen dituzte modu erdiautomatikoan: hiztegi bat errore mota bakoitzerako, eta azkenik, dokumentu bakoitzaren errore kopurua zenbatzen dute, beti ere errore mota kontuan izanik. Datu horietatik abiatuta estatistikak egiten dituzte eta hortik ondorioztatzen dituzte dokumentu sail bakoitzaren ezaugarriak.

Metodologia hori oinarritzat hartuta, euskarazko webarekin lan egin nahi genuen eta helburu berberak dituen sistema bat diseinatu dugu. Hala ere, argi utzi behar dira hainbat ezberdintasun:

— Lan honetan ez dugu egin webeko dokumentuak eskuratzeko bilaketa prozesurik, aurreko lanetan (Leturia et al., 2008) lortutako informazioa baliatu baizik. Lan horietan lortutako corpusetan zorizko laginketa egin da analizatu nahi diren gaien eta formatuen arabera (ikus 3 atala).
— Ingelesez eta alemanez «errore-hiztegiak» sortzen dituzte gero horiekin erroreak detektatzeko baina euskararen kasuan planteamendu hori ezinezkoa da euskararen ezaugarriak direla eta. Hizkuntza eranskaria denez, errore-zerrendak prestatzea ez da metodo eraginkorra, eta horren ordez, gure analisirako hizkuntza eranskariekin erabili ohi den teknologia erabili dugu: egoera finituko teknologiaren bitartez sortutako morfologia-prozesadoreak edo transduktoreak. Horietako batzuk berrerabiliak dira eta beste batzuk lan honetarako propio sortu ditugu (ikus 2 atala).
— Aztertu ditugun errore motak tipografia-erroreak, ezagutzakoak eta OCR erroreak izan dira. Egungo euskara idatzian azentu-markarik edo antzekorik erabiltzen ez denez, kodeketa-erroreak ez ditugu kontuan izan, ez baita arazorik antzeman alderdi horretatik.

Bestalde, euskarazko testuak analizatzeko unean, antzeman nahi genituzke euskara estandarra erabili beharrean, euskalkiak erabiliz idatziak izan diren testuak, edota testu diakronikoak direnak. Testu horiek guztiak estandarraren ikuspuntutik analizatuz gero, argi dago «errore» asko detektatuko direla. Dena den, une honetan lan honen esparrutik at geratzen da testu horiek identifikatzeko lana, eta etorkizuneko lanetarako planteatuta geratzen den erronka da hori.

Azkenik, ez dugu aipatu gabe utzi nahi beste gai interesgarri bat hemen planteatzen dugunarekin nolabait lotuta: webaren erabilera erroreak detektatzeko eta baita zuzentzeko ere (Whitelaw et al., 2009). Lotura izan arren, gure lana ez doa bide beretik, baina oso interesgarria da erabilera hori ere.

Ondoren, 2. atalean, egoera finituko teknologiaren sarrera bat egingo dugu, eta teknologia horri zuzentzaileak sortzeko ematen zaion erabilera-

ren berri ere emango dugu. Gero, 3. atalean esperimentuak eta lortutako emaitzak deskribatuko ditugu eta, azkenik, 4. atalean lanaren ondorioak eta etorkizuneko lanak azalduko ditugu.

2. EGOERA FINITUKO MORFOLOGIA ETA ZUZENTZAILEAK

Atal honetan, erroreak detektatzeko eta identifikatzeko erabili ditugun tresnen ezaugarriak azaldu nahi ditugu, hots, erabilitako morfologia-prozesadoreen (transduktoreen) ezaugarriak. Horretarako, horien oinarrizko teknologiaren azalpen laburra egingo dugu, eta gero transduktoreak sortzeko erabili dugun tresna aurkeztuko dugu.

2.1. Teknologia

Morfologia-prozesadoreak oinarrizko tresnak dira Hizkuntzaren Prozesamenduan (NLP, *Natural Language Process*). Euskara edo turkiera, finlandiera eta zuluera bezalako hizkuntza eranskarietan, morfologia ezin da deskribatu hitz zerrenda baten bitartez eta hitz horien analisiaren bitartez. Hizkuntza hauetan, egoera finituko teknologia (FST, *Finite State Tecnology*) erabili ohi da morfologia-prozesadoreak (morfologia-analizatzaile zein sortzaile) garatzeko. Horrelakoetan, hizkuntzaren morfologia bi fitxategien bitartez deskribatzen da (Beesley & Karttunen, 2003): (1) lexikoaren fitxategia, non morfemak eta haien ostean egon daitezkeen morfema multzoak edo paradigma multzoak deskribatzen diren (horri, morfemen morfotaktika deritzo); (2) erregela fonologikoen fitxategia, non morfemen kateatzearen ondorioz gerta daitezkeen aldaketak deskribatzen diren. Fitxategien konpilazioak transduktore bakar batean elkar daitezke; horri esker, posible da programa bat erabiliz hitzen analisia zein sorkuntza oso azkar egitea.

Erregela fonologikoak paraleloak zein sekuentzialak izan daitezke (Alegria et al., 2009). Erregela paraleloak erabiltzen badira, haien ordena ez da garrantzitsua eta gainera, testuko hitzen eta lexikoaren adierazpenaren artean ez dira tarteko lengoaiak definitu behar. Testuan ageri diren hitzek «azaleko maila» osatzen dute, horiek baitira irakurleak irakurtzen dituenak. Aldiz, hitz bakoitzaren analisi baten ondorioz lortzen den informazioari «sakoneko maila» deritzo, hor jasotzen baita morfologia-analisia.

Erregela paralelo horiek definitzeko unean, ordea, kontuan hartu behar da beste erregelek izan dezaketen eragina, eta hori, maiz, errore-iturri bihurtzen da. Sekuentziazko-erregelak erabiltzen badira, berriz, ordenak garrantzia du baina, oro har, erregela horien definizioa erosoagoa suertatzen da luzera begira.

Egoera finituko teknologia oinarritzat duten hainbat morfologia-prozesadore garatu dira finlandiera, turkiera, euskara, zuluera eta ingelesa bezalako hizkuntzetarako. Batez ere, morfologia aberatsa duten hizkuntzetan izan da teknologia hori arrakastatsua, aurretik erabiltzen ziren metodo bakunagoak ez baitziren egokiak.

Morfologia-prozesadore bat edukiz gero, erraza da ortografia-zuzentzaile bat lortzea; izan ere, hitz zuzenak lexikoa eta erregelen arabera analiza daitezkeenak dira (Kukich, 1992) (Alegria et al., 2002).

2.2. **Tresnak**

Xerox etxeko tresna-bildumak (Beesley & Karttunen, 2003) hainbat aukera eskaintzen ditu transduktoreak eraikitzeko: erregela paraleloak definitzeko (*twolc*), sekuentziazko erregelak definitzeko (*xfst*) eta lexikoa deskribatzeko (*lexc*). Haien esperientzien arabera, sekunetziazko erregelen erabilera erosoagoa da luzera begira, eta hori berretsi egin da ondoren egin diren beste inplemetazioetan.

Xeroxeko tresnak kalitate handikoak dira eta oso transduktore trinkoak lortzen dituzte, baina eragozpen handi bat daukate: lizentzia oso mugatua da eta, salbuespenak salbuespen, ezin dira erabili merkatuko aplikazioetarako. Hori dela-eta, software askean garatu dira tresnak, Xeroxekoen ordez erabiltzeko. Bi aipatu nahi ditugu hemen: *hunspell* eta *foma*.

Hunspell tresnak ez ditu transduktoreak erabiltzen eta bere aurrekariak (*ispell*, *aspell* eta *myspell*) baino hobea da; izan ere, paradigma gehiago defini daitezke eta atzizkien kateatze bikoitza egin daiteke. Lexikotik kanpo ezin dira erregela paraleloak edo sekuentzialak definitu, aldaketak lexikoaren barruan definitu behar baitira. Tresna hau garrantzitsua da Open Officeko eta Mozillako aplikazioek onartzen dutelako eta hala, edozein hizkuntzaren deskribapena *hunspell* bitartez eginez gero, automatikoki lortzen da tresna horiekin erabilgarri den ortografia-zuzentzaile bat.

Foma tresna (Hulden, 2009), berriz, Xeroxeko *xfst* eta *lexc* tresnen baliokidea da, baina software askean: *fomak* inplementazio independentea du eta GPL lizentzia (http://foma.sourceforge.net/). Beraz, egokia da morfologiaren deskribapena egiteko egoera finituko teknologiaren bitartez, sekuentziako erregelak erabiliz. Tresnaren egileak dioenez, Xeroxerako egin diren deskribapenek zuzenean funtzionatzen dute *foman* eta konpilazioa eraginkorragoa da. Abantaila garrantzitsua da Xeroxeko *lexc* eta *xfst* aplikazioetarako eginak diren deskribapenak zuzenean konpilatu ahal izatea *fomarekin* (salbuespenak salbuespen, eta datuak Unicode formatuan egonik): aplikazio horien komando berberak eta sintaxi berbera erabiltzen ditu *fomak*. Hala, lehen aipatutako liburua (Beesley & Karttunen, 2003) *fomako* eskuliburu gisa erabil daiteke hein handi batean.

2.3. **Ortografiaren egiaztatzailea eta errore detektatzaileak/ zuzentzaileak**

Hitz zuzenak identifikatzeko eta errore mota bakoitza (tipografikoak, ezagutzakoak eta OCR motakoak) detektatzeko honako baliabideak erabili ditugu:

1. Euskara estandarraren transduktorea. Hitz zuzenak identifikatzeko erabiltzen da.
2. Ezagutza-erroreak detektatzen dituen transduktorea.
3. OCR motako erroreak detektatzeko transduktorea.
4. *fomako med* funtzioa transduktore estandarrari aplikatuta. Horrela errore tipografiko gehienak detektatzen dira: gehitzeak, ezabapenak eta ordezkapenak.
5. Transposizioak detektatzeko transduktorea.

Ikus ditzagun banan-banan[2].

Euskara estandarraren transduktorea

Euskara estandarraren transduktorea aspalditik dago garatuta (Alegria et al., 2002), eta 2009 urtean zehar *foma* aplikaziora migratu da (Alegria et al., 2009).

Berrerabili dugun euskara estandarraren transduktorea sortzeko behar den lexikoaren fitxategia oso handia da: 90.000 lerro inguru ditu eta bertan biltzen dira sarrera lexiko guztiak (80.000 inguru), morfotaktika eta informazio morfologikoa. Aldaketa fonologikoak deskribatzeko 23 erregela paralelo konplexu erabili ziren (Alegria et al., 2002), gero 75 sekuentziazko erregela bakun bihurtu zirenak (Alegria et al., 2009). 2009ko artikuluan erregela paraleloak sekuentziazko erregelak bihurtzeko prozesua deskribatzen da.

Ezagutza-erroreen transduktorea

Aurrekoa bezala, ezagutza-erroreak edo ereduzko errore «tipikoak» detektatzen dituen transduktorea aspalditik dago garatuta. Ortografia-zuzentzailea garatzean landu zuten Eukal Herriko Unibertsitateko IXA taldean euskarazko ohiko erroreak detektatzeko transduktorea, funtsezkoa baita an-

[2] Transduktoreen eta tresnen inguruko xehetasun gehiago lortu nahi izanez gero, helbide honetara jo daiteke: http://ixa.si.ehu.es/Ixa/Argitalpenak/Hiztek%20proiektuak. Bertan HAP (Hizkuntzaren Azterketa eta Prozesamendua) masterrean aurkeztutako tesia dago: «Errore ortografikoen azterketa web-eko dokumentuetan», Izaskun Etxeberriak aurkeztua.

tzemate hori zuzentzaileak zuzenketa egokiak proposatzeko. Errore horiek detektatzeko, bai lexiko-sarrerak bai erregela berriak gehitzen zaizkio estandarrari (Alegria et al., 2002), eta horrela transduktore berri bat sortzen da.

Adibide bat jarrita hobeto ulertuko dugu nolakoa den estandarrari gehitzen zaion informazioa. Transduktore estandarrak *etxetik* hitza analizatzean, *etxe + Etik*[3] adierazpena lortzen du. Aldiz, *etxetikan** forma ematen bazaio analizatzeko, ez du analisi egokirik lortuko, forma hori ez baita estandarra. Gipuzkeraz, ordea, ohikoa da *etxetikan** forma erabiltzea, eta euskara estandarraren ikuspuntutik zuzenketa egokia proposatzeko, interesgarria da aldaera hori antzematea. Aldaera hori detektatzeko, lexikoan gehikuntza bat egiten da *+Etikan* atzizkia *+Etik* ordez ager daitekeela zehaztuz; horrela, transduktore berria gai da *etxetikan** forma analizatzeko, eta *etxetik* formarekin lotzeko.

Lexiko-sarreraz gain, erregela fonologiko berriak ere gehitzen dira. Adibidez:

$$h\ (\rightarrow)\quad 0\quad |\ |\ \_\ ;$$

Erregela horrek adierazten du lexikoko h letra gal daitekeela (galera hori aukerakoa dela adierazteko → eragilea parentesi artean jartzen da).

Aurreko bi gehikuntzak konbinatuta, posible da honako forma ezestandarrak ezagutza-errore gisa identifikatzea transduktore berriarekin: *zuhaitzetikan**, *zuaitzetik** eta *zuaitzetikan**.

Erregela berriak ere *fomara* migratu dira (Alegria et al., 2009) eta sekuentzialki konposatu dira erregela estandarrekin. Horrela ez zaie inolako aldaketarik egin behar estandarrei.

Transduktore berriarekin, ezagutza-erroreak detektatzeaz gain, posible da errore horiek haien forma estandarrarekin lotzea, eta nahi izanez gero, posible da lotura hori egiteko zein erregela aplikatu den jakitea (Alegria et al., 2009). Hori oso baliagarria izan daiteke konputagailuz lagunduriko irakaskuntza-sistemetan.

OCR transduktorea

OCR aplikazioek testua sortzen dute irudietatik abiatuta, baina emaitza ez bada errebisatzen erroreak gertatu ohi dira prozesu horretan. OCR erroreak identifikatzeko beste transduktore bat definitu dugu. Aurrekoaren ildotik, ideia sinplea da: transduktore estandarrari hautazko erregela batzuk

[3] E sinboloa morfofonema bat da eta atzizkia lotzean *e* karakterea sor daitekeela adierazteko erabiltzen da.

gehitu dizkiogu ohiko errore horiek detektatzeko. Erregelak idazteko, aurretik egindako esperimentu batzuk hartu ditugu kontuan eta baita erreferentziazko lanean ingeleserako eta alemanerako erabiltzen dituzten patroiak ere, OCR erroreak ez baitaude hizkuntzarekin berarekin lotuak, letren itxurarekin baizik: *1* eta *l* karaktereak nahastu, *o* eta *c* nahastu...

Hala, *c*/*e*, *n*/*ri*, *rn*/*m* eta antzeko aldaketak sortzen dituzten 24 erregela berri idatzi ditugu *fomarako*. Erregela horiek paraleloan eratu dira, horrela saihesten baita aldaketa bat beste batekin kateatzea. Adibidez, bi OCR aldaketa hauek posible dira: *d* letra *cl* bihurtzea eta *c* letra *e* bihurtzea; bi aldaketa horiek bata bestearen atzetik egiten badira, bukaeran ikusiko dugu *d* letra *el* bihurtu dela, eta hori ez da ohiko OCR errorea.

Azkenik, beraz, erregela horiek transduktore estandarrari gehitu dizkiogu transduktore berria lortzeko eta hala OCR erroreak detektatzeko edozein hitz estandarrean.

med *funtzioa*

Tipografia-erroreak identifikatu ahal izateko, aurreko prozesu bera egin daiteke: transduktore estandarrari erregela batzuk gehitu horien bitartez tipografia-erroreak detektatzeko. Ezabapenak, gehitzeak, ordezkapenak eta transposizioak dira tipografia-errore ohikoenak (Kukich, 1992). Ez da zaila errore horiek detektatzeko erregelak definitzea, baina erregela horiek izugarri zabaltzen dute automataren tamaina: edozein letra ezaba daiteke, edozein biren artean beste bat ager daiteke eta edozein letra ordez daiteke beste edozeinekin. *Foma* aplikazioak, ordea, beste aukera bat eskaintzen digu, askoz hobea, errore tipografiko gehienen identifikazioa egiteko: *med* komandoa (Hulden, 2009).

Komando horren bitartez posible da transduktoreari hitz bat ematea (zuzena ala ez) eta berak hitz horretatik gertuen dauden hitzak itzultzea (horiek bai zuzenak). Hitzen arteko distantzia kalkulatu ahal izateko, aldaketa posible bakoitzak zenbateko kostea duen adierazi behar zaio komandoari, eta guk adierazi dugu karaktere bat gehitzeak, ordezteak zein ezabatzeak 1 kostua duela. Hala, hitz batetik gertuen daudenak kalkulatzeko unean, *med*ek koste horiek hartuko ditu kontuan. Horrez gain, bilaketaren distantzia 1 balioan mugatu dugu, hau da, *med*i ematen diogun hitzaren eta berak itzultzen dituenen arteko distantzia, gehienez, 1 da. Adibidez, sarrerako hitza *zuhatzak** bada, komandoak *zuhaitzak* eta *zehatzak* itzuliko ditu, biak baitira 1eko distantziara dauden hitz zuzenak. Bestalde, hitz batean dagoen errorea tipografikoa dela identifikatzeko, nahikoa da *med* komandoak hitz zuzen bakar bat aurkitzea. Hori dela eta, mugatu egin dugu kopuru hori, komandoak bilaketa gelditzeko hitz zuzen bat aurkitzean.

Beraz, *med* komandoarekin detektatu dezakegu karaktere bat gehitu dela, kendu dela edo ordeztu dela hitz batean. Errore tipografiko guztiak detektatzeko transposizioak detektatzea falta zaigu.

Transposizioen transduktorea

Zoritxarrez, *med* komandoarekin ezin ditugu transposizioak modu egokian identifikatu, horien distantzia 2koa baita. Izan ere, bi letren arteko transposizioa suertatzen da hitz batean, bi letren ordena aldatuta dagoenean: adibidez *kal**ba**aza**, *kal**ab**aza* hitzaren ordez.

Transposizioak tratatzeko erregela multzo berri bat definitu dugu *foman* eta transduktore estandarrari gehitu diogu, horrela hitz estandarretan gertatu ahal diren transposizioak detektatzeko.

Adibidez, *b* letraz hasten den bikote batean transposizioa gerta daitekeela adierazteko bi erregela hauek erabiliko ditugu:

«b» Alf (→) «1» ... «b» ;
«1» Alf → 0 ;

Lehendabiziko erregelak transposizioa markatzen du *1* sinboloaren bitartez eta *b* letra gehitzen du bikotearen eskuinean. Gero, bigarren erregelak *1* sinboloa eta atzetik datorren letra ezabatzen ditu. Bi erregela horiek elkartuta, posible da *ba* bikotea *1bab* bihurtzea lehen erregelarekin, eta gero *1bab* katea *ab* bihurtzea bigarren erregelarekin. Beraz, posible da *kalbaaza** errorea *kalabaza* hitzaren transposizioa dela identifikatzea.

Transposizioa alfabetoko edozein bi letraren artean gerta daitekeenez, aurrekoen antzeko 26 erregela definitu ditugu, letra bakoitzeko bat, eta horrela sortu dugu transposizioak detektatzen dituen transduktorea.

3. **ESPERIMENTUAK ETA EMAITZAK**

Sarrerako atalean esan dugunez, webetik lortutako euskarazko dokumentuetan erroreak aztertzea dugu helburu: zenbat errore eta zein motakoak ageri diren. Analisi horren bidez ikusi nahi dugu posible ote den webetik osatutako corpusen kalitatea bermatzeko iragazki bat diseinatzea.

3.1. **Esperimentuen diseinua**

Corpusa

Corpusen kalitatearen lagin adierazgarri bat izateko, WAC alorrean euskaraz egin diren lanetan oinarritu gara eta, erreferentziatzat hartu dugun lanean bezala (Ringlstetter et al., 2006), bi corpus mota erabili ditugu gaiari dagokionez (orokorrak eta espezifikoak) eta bi dokumentu mota (HTML eta PDF). Horrela, 5 corpus osatu ditugu.

Corpus orokorra sortzeko euskarazko maiztasun handieneko 500 hitzeko zerrenda osatu zen eta zoriz lortutako hirukoteak eman zitzaizkien bilatzaileei, emaitzari gaiaren inguruko iragazkirik ezarri gabe; hala egiten du Sharoff-ek (2006). Horrela 71.500 dokumentuz (81 milioi hitz) osatutako corpusa lortu zen. Corpus horretatik zoriz aukeratu ziren HTML eta PDF formatuko hainbat dokumentu. Dokumentu kopurua finkatzeko, Ringlstetter et al. (2006) lanean erabilitako bera hartu dugu. 1. taulan ageri dira ezaugarri zehatzak.

1. taula. Corpus orokorreko laginaren tamaina

Mota	Dokumentuak	Hitzak
HTML	2.000	3,2 M
PDF	1.000	2,5 M

Corpus espezifikoei dagokienez, esperimentuetan erabili ditugunak osatzeko lagin bat aukeratu da zoriz, hainbat dokumenturen artean. Leturia et al. (2008) lanean azaltzen den metodologiari jarraituz, hiru arloetako dokumentuak lortu zituzten: turismokoak, informatikakoak eta bioteknologiakoak. 2. taulan ageri dira haien tamainak.

2. taula. Corpus espezifikoen tamaina lagina hartu baino lehen

Corpusa	Dokumentuak	Hitzak
Turismoa	1.238	1,5 M
Informatika	1.672	2,5 M
Bioteknologia	302	0,4 M

3. taula. Corpus espezifikoen laginen tamaina

Corpusa	Dokumentuak	Hitzak
Turismoa	500	0,6 M
Informatika	500	0,7 M
Bioteknologia	302	0,4 M

Mota bakoitzeko dokumentuen artean zoriz aukeratu da lagin bat esperimentuetan erabiltzeko.

Ezaugarri zehatzak 3. taulan ageri dira eta aurrekoan bezala, dokumentu kopurua finkatzeko irizpidea berriro izan da Ringlstetter et al. (2006) lanekoa: 500 dokumentu corpuseko. Bioteknologiako corpusa salbuespena da, abiapuntuko corpusa txikia baitzen eta kasu horretan zegoen guztia erabili da.

Prozesamendua

Benetako erroreak ez diren hitz asko eta asko erroretzat ez hartzeko, tokenizazio prozesua egin ondoren, iragazki bat pasa diogu dokumentu bakoitzari hainbat hitz kontuan ez hartzeko: hala nola luzera minimoa ez duten hitzak eta letra larriak edo karaktere bereziak dituztenak. Irizpide horiek, berriro, Ringlstetter et al., (2006) lanean finkatutakoak dira.

Iragazkia erabili ondoren, transduktore estandarra erabiltzen da dokumentuko hitz bakoitzarekin eta ontzat ematen dira analisia duten hitz guztiak. Analisia ez dutenekin, ondorengo urratsak ematen dira:

— Ezagutza-erroreak identifikatzen dituen transduktorearekin analizatzen dira, eta horrela lortzen da errore mota hori izan ahal dutenen hitz zerrenda.
— Transduktore estandarraren eta *med* funtzioaren bitartez, aldaketa bateko distantziara (aldaketa bakar bat onartzen da) dauden hiru errore tipografiko identifikatzen dira: ezabapenak, gehitzeak eta ordezkapenak. Horrez gain, transposizioen transduktorearen bitartez, transposizio erroreak identifikatzen dira. Errore horiek guztiak zerrenda bakar batean biltzen dira; izan ere gerta daiteke errore bat transposizio moduan eta baita ordezkapen moduan ere identifikatzea, eta ez dugu errore hori bi aldiz kontatzerik nahi.
— OCR erroreak identifikatzen dituen transduktorearekin analizatzen dira, errore mota hori duten hitz zerrenda lortzeko.

Aurreko prozesua corpus batean egin eta gero, lehen emaitzak analizatu genituen prozesuaren funtzionamendu egokia egiaztatzeko. Esan dugunez, hiru errore motak identifikatzeko unean automatei eman diegun hasierako informazioa bera izan da kasu guztietan: jatorrizko dokumentuan token ezestandarrak direnen zerrenda, hau da, transduktore estandarrak errore gisa markatu dituenak. Bestalde, hiru errore mota ezberdinak identifikatzen ditugunean (ezagutzakoak, tipografikoak eta OCR erroreak), emaitzak ez dira zertan disjuntuak izan eta egon daitezke hitz ezestandarrak automata bat baino gehiagorekin identifikatzen direnak. Adibidez, *zuaitza** hitza identifika daiteke ezagutza-errore gisa, *h* galtzea hala kontsideratuta dagoelako, baina posible da ere bai, errore tipografiko gisa identifikatzea, *h* letra (beste edozein bezala) ezabatua izan delako. Efektu hori agerian geratzen zen lehendabiziko estatistiketan dokumentu batzuen zerrenden kopuruak

aztertu genituenean: errore tipografikoak beti ziren gehiengoa, ezberdintasun nabarmena zegoela; izan ere, askotan, erroreen % 90 inguru errore tipografiko gisa identifika zitekeen. Efektu hori saihesteko, emaitzetako ebakidura horiek tratatu ditugu eta zerrenden artean banaketak egin ditugu; horrela argi ikusi dugu zenbat errore identifikatzen diren mota bakoitzeko. Hauxe izan da irizpidea hori egiteko: lehentasuna eman ezagutza-erroreen bitartez identifikatutakoari, hau da, hitz bera bi eratara identifikatzen bada, ezagutza-errore eta errore tipografiko gisa, edo ezagutza- eta OCR errore gisa, ezagutzakoa dela kontsideratuko dugu, hots, beste zerrendetatik kendu egingo dugu. Beste bi zerrenden artean, tipografikoen eta OCRkoen artean, ez dugu tratamendurik egin, datuetan ikusi genuelako OCR bitartez identifikatutakoa oso zerrenda motza zela beti.

Esan beharrekoa da errore gisa identifikatu diren hitz batzuk ez direla erroreak, eta hori gertatzen da batez ere errore tipografikoen zerrendan. Ez da bitxia lexikoan ez dagoen hitz zuzen baten eta bertan dagoen beste baten artean bateko distantzia egotea. Adibidez, *blogean* hitza ez da hitz zuzentzat jotzen, *blog* hitza ez dagoelako lexikoan, eta errore moduan identifikatzen saiatzean, *blokean* hitzaren errore tipografiko gisa identifikatzen da. Ringlstetter et al. (2006) lanean ere aipatzen da arazo hori. Egiaztatu dugun beste arazo bat euskaraz idatzita ez dauden hitzekin uztartuta dago. Testuetan zehar ageri dira beste hizkuntzetan zuzenak diren hitzak: batzuetan gaizki filtratutako HTML etiketak izan daitezke, baina beste batzuetan testuaren barnean dauden hitzak dira. Edozein kasutan, testuingurua analizatu gabe ezin da jakin hitz horiek beste hizkuntzetako hitzei dagozkien, edo euskaraz gaizki idatziak dauden hitzak diren.

Arazo hori tratatzeko, pentsa daiteke hasierako testuari ezartzen zaion iragazkia gehiago zehaztea, baina, bestalde, euskaraz kalitate minimoa duten corpusak osatzea da helburu, eta beraz badu bere logika kasu horiek erroretzat hartzeak.

Aurrekoarekin lotuta, garbi utzi behar da ez dela kontuan hartu ez *real-word errors* deritzen arazoaren tratamendua (Kukich, 1992), ez eta errore tipografiko bat baino gehiago dituzten kasuak ere. Hizkuntzaren barruan egon badauden baina testuinguru jakin batean ezegokiak diren hitzak dira *real-word errors* horiek (*nahiz* ↔ *naiz*, *esker* ↔ *ezker*...).

3.2. **Emaitzak**

Corpus motaren eta dokumentu motaren arabera lortu ditugun emaitzak aurkeztuko ditugu atal honetan. Grafiko asko egin ditugu baina horietako batzuk besterik ez dugu aurkeztuko hemen gehiegi ez luzatzearren.

Corpus orokorrei dagokienez, bakoitzaren batez besteko errore-tasa kalkulatu dugu (analizatutako 100 hitzetatik zenbat diren erroreak), errore

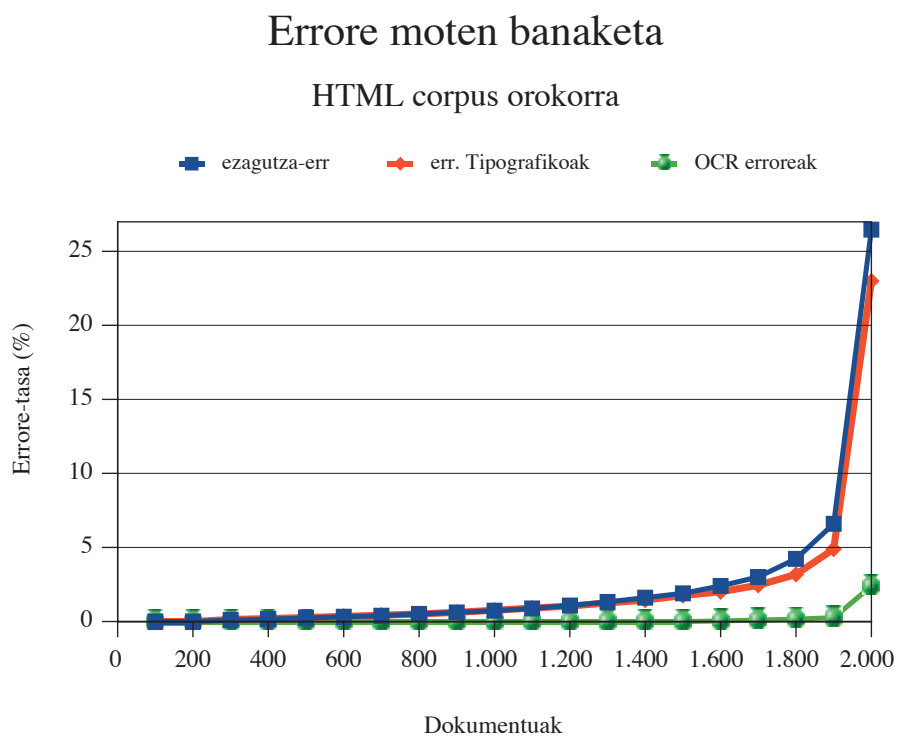
mota kontuan hartu gabe. Hauek dira lortutako zenbakiak: HTML corpusean, errore-tasa 3,09 da batez beste eta PDF corpusean, berriz, 2,62 da. Dena den, batez besteko hori ez da oso adierazgarria, erroreak ez baitira uniformeki banatzen dokumentuetan; gainera, badira errore asko dituzten dokumentuak eta ia batere errorerik ez dutenak. Hori agerian geratzen da 1. eta 2. irudietan non bi corpus orokorren hiru errore moten banaketa egin den (ezagutzakoak, tipografikoak eta OCR erroreak).

Oro har, bi corpus horien emaitzak aztertuta, hauek dira erdietsi ditugun ondorioak:

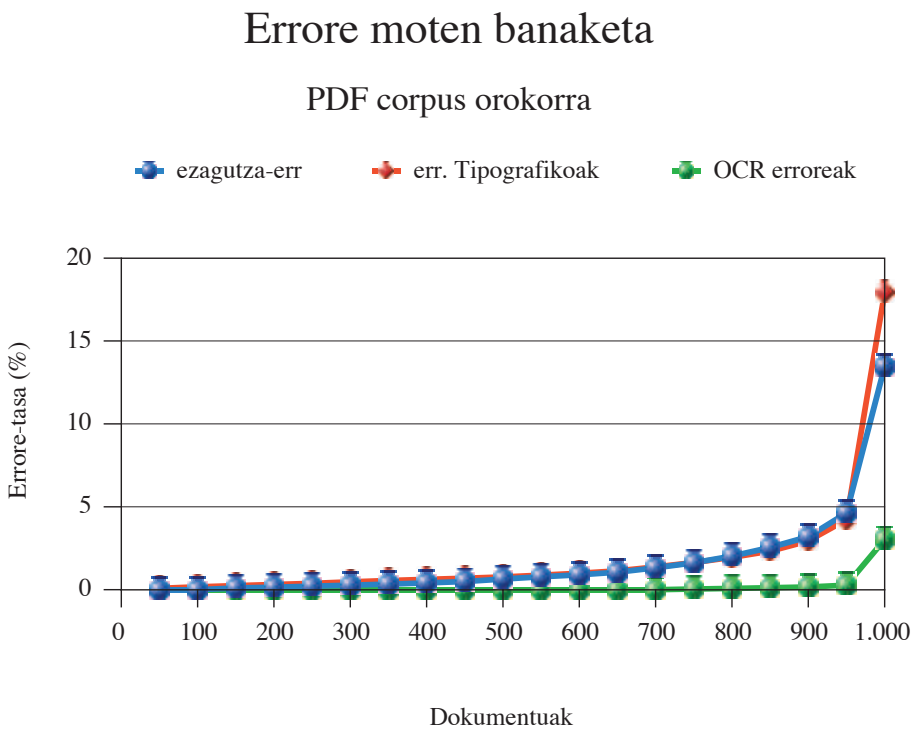
— Dokumentuen erdiak baino gehiagok ia ez du errorerik, eta errore-tasa altua duten dokumentuak corpusaren % 10 dira.
— Ezagutza-erroreek eta errore tipografikoek antzeko maiztasuna dute. Eskuz aztertu ditugu horien artean maizen gertatzen direnak eta ikusi ahal izan dugu askotan euskalkien erabilerari lot dakizkiokeen «erroreak» direla: *bainan*, *bezela*, *batetan*, *daben*, *bethi* eta abar.
— OCR erroreak oso gutxi gertatzen dira besteekin alderatuz gero. Gainera, errore hauen kopurua antzekoa da HTML eta PDF formatuetan, eta espero genuen PDF formatuan gehiago izatea. Hori dela-eta, ohartu gara webetik jaso diren dokumentuen artean ia ez dagoela errebisatu gabeko dokumentu eskaneaturik. Esperimentu erraz bat egin nahirik, webean bilatu ditugu maiztasun handiz hitzak OCR erroredun hitza eta emaitzek hipotesi bera baieztatu dute. Bestalde, azterketa sakonagoa egin beharko litzateke OCR aplikazioek sortutako dokumentuetan oinarrituta; horrela, neurtu eta doitu egingo lirateke OCR transduktorearen erregelak eta ziurtatu egingo litzateke detekzioaren zehaztasuna.

Gai espezifikoen corpusei dagokienez, batez besteko errore-tasak hauek dira: Informatikako corpusaren tasa 2,43 da, Turismoko corpusaren errore-tasa 1,88 eta Bioteknologiako corpusarena 1,48. Bioteknologiako corpusari dagokio beraz tasarik txikiena. Ez da ahaztu behar corpus hau txikiagoa dela, ez zelako dokumentu gehiagorik aurkitu gai honen inguruan. Gai berezituagoa denez, ez da erraza euskarazko dokumentu asko aurkitzea. Dena den, tasaren arabera, badirudi dokumentu horiek hobe idatzita daudela, besteak baino. Alderatzen badira errore-tasa hauek corpus orokorren tasekin, kasu guztietan txikiagoa da corpus espezifikoen errore-tasa.

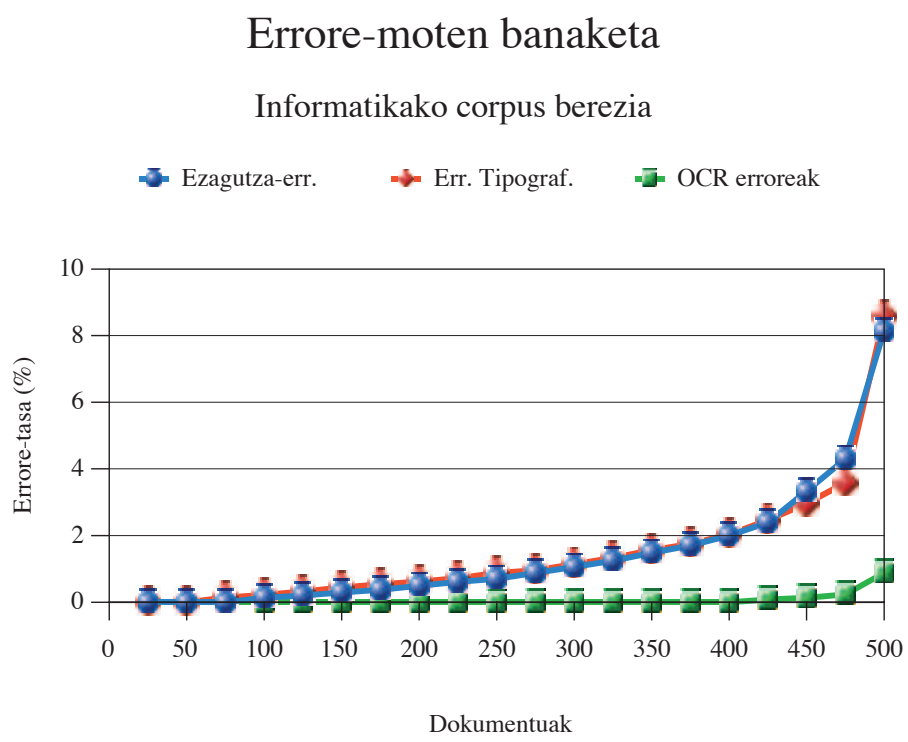
Errore-mota bakoitzaren banaketari dagokionez, Informatikako corpusaren grafikoa besterik ez dugu aurkeztu 3. irudian, kasu guztietan banaketa antzekoa delako: ezagutza-erroreak zein errore tipografikoak maila berean daude corpus guztietan, eta OCR erroreak, berriz, oso gutxi dira beste bien aldean.



1. irudia. HTML dokumentuen errore-tasa, errore motaren arabera



2. irudia. PDF dokumentuen errore-tasa, errore motaren arabera



3. irudia. Informatikako dokumentuen errore-tasa, errore motaren arabera

Oro har, alderatzen badira analisi honetan aurkitutako errore-tasak eta Ringlstetter et al., (2006) erreferentziazko lanean ingeleserako eta alemanerako aurkezten dituztenak, euskarazko dokumentuen errore-tasa askoz handiagoa da. Egia da euskara hizkuntzaren egoera soziolinguistikoa eta ingelesarena edo alemanarena oso desberdinak direla, eta balitekeela horrek diferentzia hori neurri handi batean bideratu izana, baina ez da hori desberdintasun bakarra ezta funtsezkoena. Hasieratik esan dugunez, erroreak detektatzeko erabiltzen dugun prozedimendua ez da bera bi lanetan. Haiek errore-zerrendak prestatzen dituzte erroreak detektatzeko, eta zerrenda horiek hainbat muga dituzte. Lortzen duten doitasuna handia da, baina estaldura, berriz, ez da hain ona. Gure lanean, aldiz, analizatzaile orokorrak erabiltzen ditugu erroreak detektatzeko, eta ondorioz, estaldura handiagoa lortzen dugu. Argi dago estaldura zabalagoa izateak eraman gaitzakeela errore gehiago detektatzera euskarazko testuetan.

4. ONDORIOAK ETA ETORKIZUNEKO LANA

Lan honetan euskarazko ortografia-erroreen azterketa egin dugu webetik lortutako hainbat corpusetan, corpus horien kalitatea estimatzeko. Horren bitartez ikusi dugu posible dela dokumentuak iragazteko tresna auto-

matiko bat lortu eta gutxieneko kalitateko web corpusak osatu ahal izatea. Gutxieneko kalitate bat bermatuta egonez gero, hala osaturiko corpusetatik lor daitekeen informazioa egokiagoa izango da kasu gehienetan, batez ere, informazio linguistikoaren bila ari bagara. Bide batez, erroreen tipologia estimatzen saiatu gara: zenbat errore dagoen mota bakoitzeko eta nola identifikatu mota bakoitza.

Esperimentuen bidez ikusi dugu posible dela errore asko dituzten testuak detektatzea, eta posible dela errore mota bakoitzaren estimazioa egitea ere. Egin dugun lanaren metodologia finkatzeko, 2006an *Computational Linguistics* aldizkarian argitaratutako lan sakon bat hartu dugu erreferente gisa: Ringlstetter *et al*. (2006) lana; horretan oinarrituta gure lanerako hainbat erabaki hartu eta ezaugarri finkatu ditugu: corpus kopurua, corpusen ezaugarriak (tamaina, formatua, gaia...). Metodologiaz gain, ortografia-erroreak analizatu ahal izateko, ezinbestekoa zen datuak edukitzea, hots, webetik hainbat euskarazko testu eskuratu behar genituen. Horretarako, lankidetza ezin hobea izan dugu Elhuyarreko Igor Leturiarekin. Igor aditua da *Web as corpus* gaian euskaraz, eta bere aurreko lanetan (Leturia et al., 2008) osatuak zituen euskarazko hainbat corpusen lagina hartu dugu.

Metodologia finkatuta eta datuak izanik, tresnak behar genituen ortografia-erroreen analisia egiteko eta tresna horiek transduktoreak izan dira. Erreferentziazko lanean errore-hiztegiak sortzen zituzten erroreak identifikatzeko, baina euskararen kasuan ezin zen lana horrela planteatu hizkuntzaren ezaugarriengatik. Hori dela-eta, gure analisian erabili ditugun tresnak transduktoreak izan dira, egoera finituko teknologiak lortzen dituen oinarrizko tresnak alegia. Aukera hori egokia da flexio aberatseko hizkuntzekin lan eginez gero. Transduktore batzuk landuak ziren aurretik eta beste batzuk analisi hau egiteko garatu ditugu; guztiak sortzeko *foma* tresna erabili dugu. Kode irekiko tresna da *foma*, Mans Huldenek 2009an argitaratua. Errore-hiztegien ordez transduktoreak erabiltzeak hainbat onura zituen: (1) azkarragoa zen, ez baitzen zerrenda luzerik prestatu behar; (2) aurretik egindako hainbat lan berrerabiltzeko aukera eskaintzen zuen; (3) garatu behar izan diren transduktoreak azkar garatu ahal izan dira aurreko esperientziari esker; (4) erroreen detekzioan estaldura handia lortu dugu, ez baitugu lan egin errore zerrenda mugatuekin.

Emaitzei begira, gure analisiaren emaitzak erreferentziazko lanekoekin alderatuz gero, ondorioztatu dugu dokumentu gehienek errore-tasa txikia dutela, baina batzuek askoz tasa altuagoa dutela. Dena den, tasa horren balioa aztertuz gero, ingelesaren balioa nabarmen txikiagoa da: alde batetik, gure metodoak errore gehiago identifikatzen ditu, eta bestetik, ingelesa eta euskararen egoera ez da konparagarria, euskararen estandarizatze prozesua abian baitago oraindik.

Bestalde, OCR erroreen inguruan lortutako emaitzen arabera, etorkizunerako OCR erroreen azterketa sakonagoa egin behar da; izan ere, zaila da une honetan ditugun erregelak ongi funtzionatzen dutela ziurtatzea, halako dokumentuetan lortu ditugun emaitzak analizatu ondoren.

Bukatzeko, lortutako emaitzetan eskuzko analisi batzuk egin ditugu eta horiek aurretik genituen susmoak berretsi dituzte. Euskalkien erabilera ageri da webeko testuetan, baina euskara estandarraren ikuspuntutik euskalkietako hitz asko erroretzat hartzen dira. Horrek beste galdera bat sortzen digu: euskarazko testuen bila ari bagara webean, euskarazkotzat hartuko al ditugu esaterako zuberotarrez idatzitakoak? Ezezkoan, datu asko galtzen ari gara. Oso lan interesgarria izango litzateke euskalkien identifikazioa egin ahal izatea euskalkien hitzak erroretzat ez hartzeko. Gure ustez, gainera, euskalkien identifikatze hori baliagarria izan liteke testuaren erregistroa identifikatzeko: erregistro zaindua, kaleko erregistroa... Ildo horiek irekita daude une honetan eta etorkizuneko lan interesgarritzat jotzen ditugu.

ESKER ONAK

Egileek bereziki eskertu nahi diogu Mans Huldeni, transduktoreak *fomarekin* eraikitzeko emandako laguntzagatik.

Aldi berean, eskerrak eman nahi dizkiegu ikerketa hau aurrera eramateko diru-laguntza eman diguten erakundeei eta ikerkuntza-proiektuei: Eusko Jaurlaritzari, Berbatek proiektua (IE09-262) eta Zientzia eta Berrikuntzako Ministerioari, OpenMT2 proiektua (TIN2009-14675-C03-01).

BIBLIOGRAFIA

ALEGRIA, I., ARANZABE, M., EZEIZA, A., EZEIZA, N., URIZAR, R. (2002). Using Finite State Technology in Natural Language Processing of Basque. LNCS: Implementation and Application of Automata. (2002). Springer.

ALEGRIA, I., ETXEBERRIA, I., HULDEN, H., MARITXALAR, M. (2009). Porting Basque Morphological Grammars to foma, an Open-Source Tool. FSMNLP2009. Pretoria. South Africa.

BEESLEY, K.R. and KARTTUNEN, L. (2003). Finite State Morphology. CSLI Publications, Palo Alto, CA.

HULDEN, M. (2009). Foma: a Finite-State Compiler and Library. EACL 2009. Demo session. pp 29-32.

KILGARRIFF, A. and GREFENSTETTE, G. (2003). Introduction to the special issue on the web as corpus. Computational linguistics, 29(3): 333-347. MIT Press.

KUKICH, K. (1992). Techniques for Automatically Correc-ting Words in Text. ACM Comput. Surv. 24(4): 377-439.

LETURIA, I., SAN VICENTE, I., SARALEGI, X. and LOPEZ DE LACALLE, M. (2008). Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. Proc. of the 4th. Web as Corpus Workshop. LREC 2008.

RINGLSTETTER, C., SCHULZ, K.U. and MIHOV, S. (2006). Orthographic errors in web pages: Toward cleaner web corpora. Computational Linguistics, 32(3): 295-340. MIT Press.

SHAROFF, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Queries. WaCky! Working Papers on the Web as Corpus, 63-98. Ed. Marco Baroni and Silvia Bernardini. Bologna.

WHITELAW, C., HUTCHINSON, B., CHUNG, G.Y. and ELLIS, G. (2009). Using the web for language independent spellchecking and autocorrection. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, 890-899.