IGOR LETURIA AZKARATE
*Informatikaria eta ikertzailea*

# Automatikoki eraikitako hiztegien ataria

**Elhuyarreko Hizkuntza-teknologien I+G sailean teknologia berri batekin ikertzen aritu gara azken urteotan, zubi-hizkuntza bat erabiliz hizkuntza-pare berrientzako hiztegiak automatikoki sortzeko. Eta ikerketa horren emaitzak jendearen eskura jartzeko, Hiztegi Automatikoen Ataria online paratu dugu, 5 hiztegi elebidun berrirekin.**

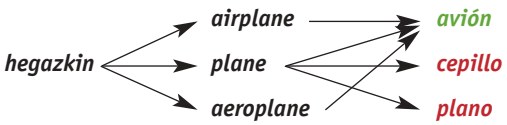
Hizkuntza-baliabide oinarrizko bat baldin badago, hori hiztegiak dira. Eta hiztegien artean, elebidunak oso beharrezkoak dira kasu askotan: hizkuntzak ikastean, itzulpengintzan... Elhuyarreko Hizkuntza-teknologien sailean ere hiztegi elebidunak ezinbestekoak ditugu hainbat arlotarako: itzulpen automatikorako, bilaketa eleanitzerako...

Alabaina, hiztegiak egitea garestia da. Hori dela eta, euskarazko hiztegi elebidunak ez dira guk nahi bezain ugariak, eta berdin gertatzen da baliabide urriko beste hizkuntza batzuekin ere. Normalean, kontaktuan dauden hizkuntzetarako hiztegiak egoten dira (tokian tokiko beste hizkuntza batzuk edo hurbilekoak), edo nazioarteko hizkuntza nagusietarakoak. Baina ez dira egiten beste hizkuntza txikietarakoak edo urruneko hizkuntza nagusietarakoak, eta horrek desabantailan jartzen ditu euskara bezalako hizkuntza txikiak bertako erdarekiko. Etorkinek hizkuntza ikasteko aukeretan, adibidez: ez da erraza euskara zuzenean norberaren hizkuntzatik abiatuta ikastea, beti gaztelania, ingelesa edo frantsesa zubi hartuta egin behar da; beraz, aurrez beste horietako bat ikasi behar da.

## ZUBI-HIZKUNTZAK HIZTEGIAK SORTZEKO

Zubi-hizkuntzarena desabantaila da hizkuntza bat ikasteko orduan, baina ideia hori baliatu daiteke hiztegi berriak erraz eta merke sortzeko. Izan ere, ia hizkuntza guztiek dute hiztegi elebidunen bat hizkuntza "handiren" batekin (normalean, ingelesa). Eta horrelako bi hiztegi baliatu ditzakegu, hizkuntza "handi" hori zubi moduan hartuta, bi hizkuntza-pare berriren arteko hiztegi bat eraikitzeko. Pibotajearen teknika deritzo horri, hizkuntza hori pibote moduan erabiltzean datzalako. Modu errazean azalduta, euskara-ingelesa hiztegi batean *etxe* → *house* jartzen badu eta ingelesa-alemana hiztegi batean *house* → *haus*, orduan *etxe* → *haus* dela ondorioztatzen dugu. Eta horrela eraiki dezakegu euskara-alemana hiztegi bat.

Elhuyarreko Hizkuntza-teknologien I+G sailean teknika horrekin ikertzen aritu gara azken urteetan, euskara eta beste hizkuntza batzuen arteko hiztegi berriak sortzeko asmoz. Goiko adibidea ikusita, balirudike oso teknika erraza dela; baina adibide hori oso sinplea da, errealitatean hitz batek esanahi anitz izan ditzakeelako, eta horietako bakoitzak hainbat ordain. Horrek dakar hiztegien kateatze sinple batek baliokidetasun oker asko sortzea, irudiko adibidean ikusten den bezala.




Beraz, kalitateko hiztegi bat sortzeko, nahitaezkoa da ekibalentzia oker horiek automatikoki detektatu eta ezabatzea, eta horretan datza teknika honen zailtasuna. Bi metodo erabiltzen dira horretarako. Lehenak bi hitzen artean zenbat bide dauden kontatzen ditu; zenbat eta bide gehiago, orduan eta probabilitate handiagoa baliokidetasuna zuzena izateko. Bigarrenak hizkuntza bietako corpusetan hitzek duten testuinguruen antzekotasuna neurtzen du; zenbat eta antz gehiago izan testuinguruek, orduan eta aukera handiagoa baliokideak izateko. Eta, jakina, testuinguruen antzekotasuna neurtzeko, hiztegi bat behar da, hizkuntza ezberdinetan baitaude; lehenengo metodoarekin lortutako ziurrak erabiltzen dira horretarako.

Garbiketarako teknika hauek aplikatuta ere, hizkuntza-teknologietako edozein metodo automatikorekin bezala, inoiz ez dira emaitza perfektuak lortzen, hau da; beti egongo da errore-tasaren bat. Lortzen den errore-tasa hori oso aldakorra da, hainbat faktoreren araberakoa baita (hizkuntzak, erabilitako hiztegiak, erabilitako corpusak...); baina neurketa batzuen arabera, emaitza zuze-


nen portzentajea % 60-80 artekoa izan daiteke. Argi dagoenez, ez dira hiztegi perfektuak; baina ezer ez izatea baino hobe da.

## HIZTEGI AUTOMATIKOEN ATARIA

Aipatutako metodoak erabiliz, euskarazko bost hiztegi elebidun sortu ditugu, hiru kontinentetako (Afrika, Asia eta Europa) hizkuntza nagusietako 5 aukeratuta: euskara-arabiera, euskara-swahilia, euskara-txinera, euskara-hindia eta euskara-alemana. Guztietan ingelesa erabili da zubi-hizkuntza gisa. Euskara-ingelesa hiztegi gisa Elhuyarrena erabili dugu, eta ingelesa eta beste hizkuntzen artekoentzat sarean libre zeuden bost hiztegi hartu ditugu. Eraikitako hiztegiak ez dira oso handiak: oinarrizko hiztegiak dira, 8.000 eta 21.000 sarrera artekoak. Izan ere, sarean lortutako hiztegi horiek ere halakoak ziren. Hiztegi guztiak bi norabideetan dira.

Hiztegi horiek guztiak jendearen eskura jarri ditugu orain, online jarri dugun Hiztegi Automatikoen Atarian: http://hiztegiautomatikoak.elhuyar.org. Eta jendearen eskura jarri ditugula diogunean, esan nahi dugu ez direla soilik kontsultarako. Batetik, hiztegi guztiak ataritik bertatik osorik deskargatu daitezke (jatorrian erabili ditugun hiztegiak libreak zirenez, horietatik eratorri ditugunak ere libre jartzen ditugu guk). Bestetik, eta arestian esan dugunez hiztegiak ez direla guztiz perfektuak eta akatsak dituztela, webguneak parte hartzeko aukera ematen die erabiltzaileei, hiztegiok zuzendu eta hobetzeko, zuzenak eta okerrak direnak markatzeko sistemaren bidez.

Webgunean, hiztegietan hitzak bilatzeko eremua dago batetik. Bestetik, emaitzen eremuan, ordainak zuzenak edo okerrak iruditzen zaizkigun adieraz dezakegu; izan ere, emaitza bakoitzeko, hitzaren ordainaz gain, corpusetako erabilera errealak erakusten dira, bai adibide gisa balio dezaten, bai erabiltzaileari emaitza ongi edo txarto dagoen erabakitzen lagundu diezaioten. Eta, halaber, ordain ziurrak eta zalantzazkoak ezberdintzeko aukera ere ematen du. Deskargen atal bat ere badago, hiztegiak osorik XML formatuan jaitsi ahal izateko. Eta, azkenik, foro bat ere badu webguneak, erabiltzaileek hitz zehatzen zuzentasunari buruz eztabaidatu dezaten, elkarri kontsultak egin diezazkioten eta abar. Webguneak interfazea 8 hizkuntzatan du, eta teklatu birtual bat eskaintzen du alfabeto latindarra erabiltzen ez duten hizkuntzetan bilaketak egiteko.

Egindako lana ez dugu bere horretan utzi nahi. Hiztegi gehiago sortu, eta horiek ere atarian sartzeko asmoa dugu. Baita kolaborazio-lana bozketaz haragokoa izan dadin aukera emateko ere: ordainak eta adibideak gehitu edo aldatzeko aukera emanez, esate baterako.

Hiztegi Automatikoen Atariarekin, lehenengo aldiz erlazionatu dugu euskara beste 5 hizkuntzekin. Urruneko hizkuntzak irudi lezakete, eta beharbada hala izango zen lehenago, baina globalizazioarekin eta Internetekin gero eta harreman handiagoa dute elkarrekin. Uste dugu baliabide garrantzitsua direla, eta are gehiago izango direla etorkizunean, denon artean hobetzen laguntuz gero.●

*"Webguneak parte hartzeko aukera ematen die erabiltzaileei, hiztegiok zuzendu eta hobetzeko"*


Saralegi, X.; Manterola, I.; San Vicente, I.: "Analizing Methods for Improving Precision of Pivot Based Bilingual Dictionaries". Conference on Empirical Methods in Natural Language Processing (EMNLP 2011). Edinburgh (2011).

Saralegi, X.; Manterola, I.; San Vicente, I.: "Building a Basque-Chinese Dictionary by using English as a Pivot". 8th international conference on Language Resources and Evaluation, LREC'12. Istanbul. (2012).