CorpEus: Internet, corpusak eta euskara uztartuz

Kortabitarte Egiguren, Irati

Elhuyar Zientziaren Komunikazioa

Internet informazio-iturri ikaragarria da. Inork gutxik jartzen du zalantzan hori. Egun, informazioa bilatzeko ez ezik, gero eta gehiago erabiltzen da hizkuntza-kontsultak egiteko, corpusak osatzeko eta abarretarako. Hala, Internet baliabide linguistiko eta corpusen iturri aproposa bilakatzen ari da pixkanaka. Horren adibide bat da CorpEus, Internet euskarazko corpus erraldoi gisa baliatzeko aukera ematen duen tresna.

CorpEus sistemaren emaitzek horrelako itxura dute. Adibidean anorexia hitzari dagokion bilaketaren emaitza.

GAUR EGUN, HIZKUNTZA GUZTIEK BEHAR DITUZTE CORPUSAK. Corpusak formatu elektronikoan eta linguistikoki etiketatuta dauden testu-bildumak dira —linguistikoki etiketatuta egoteak esan nahi du hitz bakoitzari dagokion lema, kategoria... ematen zaizkiola— eta hizkuntzaren ikerketan eta hizkuntza-teknologien garapenean erabiltzen dira. Oso baliabide garrantzitsuak dira hizkuntza-teknologiak garatzeko, hiztegiak egiteko... Corpusak egitea, berriz, lan garestia eta neketsua da, eta zaila da beti eguneratuta edukitzea. Horregatik, euskarazko corpusak gutxi eta txikiak dira, beste hizkuntzetakoekin konparatuta behintzat.

Internetez baliatuz

Hor dago, ordea, Internet edo amarauna, testu-bilduma erraldoia, guztion eskura, euskarazko beste edozein corpusetan baino askoz testu gehiagorekin. Hori ere corpus bat da, nahiz eta linguistikoki etiketatu gabea den. Ondo legoke corpus gisa kontsultatu edo ustiatu ahal izatea. Hori da, hain justu, CorpEus-ek egiten duena.

Lehendik ere badaude horrelako zenbait tresna sarean (WebConc eta WebCorp, adibidez), baina horiek ere Interneteko beste tresna eta bilatzaileek euskararekin dituzten bi arazoak dituzte: batetik, forma zehatz bat soilik bila dezakete, eta ez hitz edo lema baten forma guztiak batera —esaterako, *lur* bilatzeko eskatu eta *lur, lurra, lurrean, lurrarekin...* aurkitzea interesatzen zaigu—; bestetik, euskarazkoak ez diren emaitzak ere eman ditzakete eta ematen dituzte, baldin eta hitz-forma bera bada beste hizkuntzaren batean (*software, anorexia* eta *sulfuroso* hitzen kasuan, adibidez).

Muga horiek gainditu ahal izateko sortu dute CorpEus. Tresna hori Elhuyar Fundazioko I+G taldeak garatu du, EHUko Informatika Fakultateko IXA Taldearen laguntzarekin, eta, esan bezala, Internet euskarazko corpustzat erabiltzeko aukera ematen du. Izan ere, Internet corpus erraldoi bat dela esan liteke, euskaraz dagoen edozein corpus baino askoz ere handiagoa. Gainera, beti ari da eguneratzen eta edukia gehitzen; beraz, hitz berrienak ere kontsulta daitezke.

Internet informazio-iturri ikaragarria da gaur egun, eta, bilaketa-tresna egokiekin, corpus erraldoi gisa ere erabil daiteke.

ARTXIBOKOA

CorpEus-ek Interneteko bilatzaileen APIak erabiltzen ditu (Google, Yahoo edo Microsoft-enekin ibil daiteke) hitz bat zein orritan agertzen den jakiteko –APIak (Application Programming Interface) zerbitzu batek beste programa batetik erabiltzeko eskaintzen dituen funtzioak dira–. Gero, orri horietan dauden hitzaren agerpen guztiak erakusten ditu, bere testuinguruan. Agerpen-kopurua ere erakusten du.

> *"Internet euskarazko corpus erraldoi bat balitz bezala kontsultatzeko aukera ematen duen tresna da CorpEus"*

Emaitzak hainbat faktoreren arabera ordena ditzake, eta emaitzen analisi linguistikoa ere erakusten du. Hainbat dokumentu-motarekin funtzionatzen du (HTML, XML, RSS, RDF,TXT, DBF, DOC, RTF, PDF, PPT, PPS, XLS). Gainera, bilaketa euskararen bi arazoak konponduta egiten du: lemaren arabera bilatzen du, eta euskarazko orriak soilik ematen ditu, Igor Leturiak, CorpEus proiektuaren arduradun eta Elhuyar Fundazioko I+G taldeko ikertzaileak, azaldu digunez.

Forma zehatz bat eta forma horren lematik eratortzen diren aukera posible guztiak agertzeko, Euskal Herriko Unibertsitateko IXA Taldeak garatu duen tresna bat erabiltzen dute. Hala, forma guztiak eskatzen zaizkio APIari, OR eragilea erabilita. Esaterako, erabiltzaileak *etxe* hitzaz galdetzen badu, honela jarriko zaio bilatzaileari: etxe OR etxea OR etxeak OR etxeari OR... Lehenengo arazoa konpondu dute, beraz. Noski, bilatzaileek ez dituzte nahi beste aukera onartzen, eta, hortaz, ez dira deklinazio guztiak bidaltzen, baina bai emaitza esanguratsuak lortzeko adina.

Euskarazko emaitzak

Arestian aipatu dugun bezala, ez dago euskarazko emaitzak soilik agertzen dituen bilatzailerik. Hori arazo bat da aurkitu nahi dugun hitza berdin esaten bada beste hizkuntza batzuetan. Horixe gertatzen da, hain justu, zenbait hitz teknikorekin –adibidez, *anorexia, sulfuroso* eta *byte*–, hitz labur batzuekin –*katu* eta *esne*, esate baterako– eta izen bereziekin –*Fiji* eta *Newton*, besteak beste–. Hain zuzen, hitz teknikoen bilaketak oso ohikoak eta erabilgarriak dira euskarazko corpusetan, terminologia ez baitago behar bezain normalizatuta euskaran.

Euskarazko emaitzak soilik eskuratzeko, CorpEus-ek iragazkiak erabiltzen ditu. Euskaraz gehien erabiltzen diren hitzak jarri dituzte iragazki gisa Elhuyar Fundazioko I+G taldeko ikertzaileek;

Banner hitzaren bilaketa CorpEus eta WebCorp corpusetan. CorpEus-ek euskarazko emaitzak soilik erakusten ditu.

guztiak AND batekin lotuta. Hitz erabilienak zein diren jakiteko, corpus bat erabili dute.

Zoritxarrez, euskaraz gehien erabiltzen diren hitzak (*eta, da, ez, ere*) motzak dira, beste hizkuntza batzuetan maiz erabiltzen dira, eta, zenbaitetan, laburdura eta akronimoak izan daitezke. Beraz, ez dago hitz *magikorik*, alegia, euskarazko testuetan soilik agertuko den eta iragazki gisa erabil daitekeen hitzik. Euskarazko hitz erabiliena *eta* da. Baina ETA akronimo bat ere bada, eta komunikabideetan maiz erabiltzen da hizkuntza askotan. Beste hitz erabilienetako bat *da* aditza da; baina errusieraz *bai* esan nahi du.

Beraz, hitz horietako zenbat erabili behar dira iragazki gisa bilaketa euskarazko orrietan soilik egiteko? Igor Leturiaren esanean "zenbat eta hitz gehiago erabili, orduan eta zehatzagoa izango da bilaketa, eta, beraz, euskarazkoak ez diren emaitza gutxiago agertuko dira. Dena den, euskarazko zenbait emaitza ere ez ditu erakutsiko, hitz horietakoren bat edo batzuk ez direlako agertzen horietan".

Zenbait muga

CorpEus orain arteko corpusen osagarri izango da. Alabaina, abantailak ez ezik, zenbait desabantaila ere baditu. Batetik, arestian aipatu den bezala, Internet linguistikoki etiketatu gabea denez, nolabaiteko ziurgabetasuna izango du beti lema bat baino gehiago dituzten hitzekin. *Pilotari* hitza bilatzean, adibidez; izan ere, *pilota* hitzaren datiboa ez ezik, pilotan jokatzen duen pertsona ere bada pilotaria.

E. CARTON

Elhuyar Fundazioko I + G taldeko kideak: ezkerretik hasita, Antton Gurrutxaga, Nerea Areta, Xabier Saralegi eta Igor Leturia.

> *"CorpEus-ek lemaren arabera bilatzen du eta euskarazko orriak soilik ematen ditu"*

ARTXIBOKOA

Formatu elektronikoan eta linguistikoki etiketatuta dauden testu-bildumak dira corpusak.

Beste desabantaila bat da orraztu gabea dela neurri handi batean —blogak, foroak, eduki pertsonala eta horrelakoak, batez ere—; abantaila gisa ikus badaiteke ere (ahozko hizkuntzatik hurbil dagoen eredua ematen delako), desabantaila ere bada, kalitatez txarragoa eta akastuna izan baitaiteke.

Bestetik, inoiz ezingo da ikusi dagoen guztia, bilatzaileek, normalean, mila orriko muga izaten dutelako; beraz, orri horietako emaitzak soilik erakuts daitezke. Eta, azkenik, bilatzaileekiko menpekotasuna du CorpEus-ek: alde batetik, haien emaitzen ordenaren menpekoak dira tresnaren emaitzak, eta, bestetik, APIetan egiten dituzten aldaketekiko eta APIei jartzen dizkieten mugekiko menpekotasuna ere badute.

Edonola ere, CorpEus izan da Internet, corpusak eta euskara uztartu dituen lehenengo saiakera. Segur aski, ez da azkena izango. Izan ere, beste hizkuntzetan ere hizkuntza-teknologietarako gero eta corpus handiagoak behar dira, eta horretarako Internet erabiltzeko joera nabarmen areagotzen ari da.

CorpEus proiektuaren web gunea:
http://www.corpeus.org