IGOR LETURIA AZKARATE
Informatikaria eta ikertzailea

ELHUYAR I+G-K EUSKARAZKO WEBEKO TESTUEN BILTEGI ERRALDOIA JARRI DU ON LINE

Web-corpusen Ataria

Elhuyar Fundazioko hizkuntza-teknologien I+G taldekook Web-corpusen Ataria jarri dugu on line. Bertan hiru baliabide jarri ditugu eskura: euskarazko corpus bat, euskara-gaztelania corpus paralelo bat, eta lehenetik automatikoki erauzitako hitz-konbinazioen kontsulta. Corpusak euskaraz dauden handienak dira, bakoitza bere motakoen artean. Baliabide hauek aurrerapauso handia dira euskararentzat, erabilgarriak baitira ez soilik hizkuntzalaritzarako, baizik eta baita hizkuntza-teknologien garapenerako ere.

Hizkuntza batentzat oso garrantzitsua da testu-corpusak edukitzea (azterketa linguistikoak egiteko balio duten testu-bildumak). Hizkuntzalaritza-ikerketetarako, edo hizkuntza-estandarizazioan erabakiak hartzeko datuak lortzeko ezinbestekoak dira. Eta oso baliagarriak dira testu sorkuntzan edo itzulpengintzan ere: hiztegietan agertu ez edo adibide nahikorik ez duten hitzak nola erabili edo nola itzuli izan diren argitu diezagukete.

Baina, horrez gain, corpusek berebiziko garrantzia dute hizkuntza-teknologien munduan. Gaur egungo telefono mugikor adimendunek ekartzen duten ahots-ezagutzako sistemetan, adibidez, corpusak erabiltzen dira guztiz ongi ulertu ez den hitz bat asmatzen saiatzeko, aukeren artean testuinguru horretan probableena zein den corpusetan begiratuta; edo itzulpen automatikoko sistemek, esaterako, corpus paraleloak (elkarren itzulpen diren testuez osatutako corpusak) erabiltzen dituzte ikasteko, 2009ko azaroko zenbakian kontatzen genizuenez.

ZENBAT ETA CORPUS HANDIAGOAK, HOBE

Artikulu hartan bertan azpimarratzen genuen corpus hauek zenbat eta handiagoak izan hobe dela. Hitz arraro baten erabilera kontsultatzeko, agerpen ezberdin gehiago, edo agertzeko aukera gehiago egongo dira corpusa handiagoa bada. Itzulpen automatikorako ere tamaina ahalik eta handieneko corpusak behar dira; horregatik da erreferentzia Google hizkuntza askotako itzulpen automatikoan, bilatzailerako indexatzen dituen testuekin corpus paralelo ikaragarriak osatzen dituelako.

Beste arlo askotan bezala, corpusenean ere baliabide gehiagoko beste hizkuntza batzuen oso atzetik dabil euskara, bai tamainan, bai denboran. Errepara diezaiogun ingelesaren egoerari: corpusgintza modernoaren abiapuntutzat jotzen den ingelesezko Brown corpusa 1964an sortu zen eta milioi bat hitz zituen; 100 milioi hitzeko British National Corpus 1995ekoa da; eta gaur egun, badaude ingelesezko milaka milioi hitzeko corpusak. Ingelesa barne hartzen duten corpus paraleloei dagokienez, Googlek 2005ean abiatutako itzulpen automatikoko sistema 200.000 milioi hitzeko corpus baten gainean entrenatu zen.

Euskaraz, aldiz, lehen corpusa (Euskaltzaindiaren *Orotariko Euskal Hiztegia*ren testu-corpusa) 1984an egin zen eta 4,6 milioi hitz ditu. Euskaltzaindiak berak egindako XX. Mendeko Euskararen Corpus Estatistikoa 2002an amaitu zen, 6 milioi hitzekin. Elhuyar Fundazioak eta EHUko IXA Taldeak Zientzia eta Teknologiaren Corpusa atera zuten 2006an, 9 milioi hitzekoa. EHUk ere urte horretan egin zuen Ereduzko Prosa Gaur deituriko corpusa, gaur egun 25,1 milioi hitzez osatua. Euskaltzaindiaren Lexikoaren Behatokia 2010ean abiarazi zen, eta egun 26,5 milioi hitz ditu. Corpus paraleloei dagokienez, itzulpen-enpresek dituzte ziurrenik horrelako handienak euren itzulpen-memorietan. Baina publikoarentzat eskuragarri eta hizkuntza-teknologietan erabiltzeko moduan oso gutxi daude; erakunde publiko (HAEEren Itzulpen Zerbitzu Ofiziala, Gipuzkoako Foru Aldundia, Bizkaiko Foru Aldundia...) edo bokazio sozialeko elkarte (EIZIE, Librezale) batzuetako itzulpen-zerbitzuen itzulpen-memoriak eta Eroskiren *Consumer* aldizkariko corpusa dira erreferentzia bakarrak, baina denak 5 milioi hitzen azpitik daude.

KONPONBIDEA, WEBA

Arazo hori konpontzeko errezeta Adam Kilgarriff corpusetan adituak ematen zuen lehen aipatutako artikuluan: weba da corpus handiak modu erraz, merke eta azkarrean osatzeko modurik onena. Izan ere, ingelesezko aipatu ditugun azken urteotako corpus erraldoi horiek ere horrela osatu dira, ikusita corpusak era klasikoan osatzea (argitaletxeetara edo komunikabideetara joz) askoz garestiago eta neketsuagoa dela.

Corpusak webetik automatikoki osatzeak baditu bere aurkakoak ere. Haien objekzio nagusia da bertan kalitate eskaseko testu asko aurki daitezkeela. Baina beste ikuspegi batetik ikusita, hori da gaur egungo hizkuntzaren erabilera erreala, eta hori aztertzeko sortu ziren corpusak. Gainera, baliabide askoz gehiago dituzten hizkuntzek webera jo badute, euskararentzat ere hori da bidea atzean gelditu nahi ez badu.

EUSKARAZKO WEB-CORPUSEN ATARIA

Elhuyar Fundazioko hizkuntza-teknologien I+G taldekook badaramatzagu urte batzuk web-corpusen (webeko testuekin metodo automatikoak erabiliz eraikitako corpusen) arloa jorratzen. Mota askotako corpusak biltzeko metodoak landu ditugu: euskarazko corpus espezializatuak (jakintza-arlo jakin bateko testuz osatuak), corpus eleaniztun konparagarriak (jakintza-arlo bereko testuz osatuak), corpus paraleloak (elkarren itzulpen diren testuz osatuak), corpus orokor erraldoiak... Horrelakoak egiteko, beharrezkoa da hizkuntza-teknologietako beste teknika batzuk garatzea: bilatzaileen APIetatik hitz jakin batzuk dituzten web-orriak eskuratzekoak, testu baten hizkuntza ezagutzekoak, testu errepikatuak edo oso antzekoak detektatzekoak, web-orriak garbitzekoak (oinak, goiburuak, nabigazio-menuak, copyright-oharrak eta horrelakoak kentzeko eta testuaz soilik gelditzeko), spama apartatzekoak, testu baten jakintza-arloa detektatzekoak, itzulpenak ezagutzekoak...

Tresna horien bidez, aipatutako mota horietako guztietako corpus asko osatu ditugu. Orain, corpus horietako batzuk on line jarri ditugu Web-corpusen Atarian (http://webcorpusak.elhuyar.org/): 125 milioi hitzeko euskarazko corpus orokor handi bat (mota horretako orain arteko handiena, alde handiz) eta 18 milioi hitzeko euskara-gaztelania corpus paralelo bat (corpus paralelo publikoen artean handiena hau ere). Corpus horien gainean hainbat bilaketa-mota egitea ahalbidetzen da webgunean. Lema edo forma jakin bat edo haien hasiera edo bukaera emanda galde daitezke, gehienez 5 hitzeko distantziara dauden hiru hitz arterainoko konbinaziotan. Paraleloan bi hitz arterainoko konbinazioak galde daitezke, baina hizkuntza batean, bestean edo bietan izan daitezela eska daiteke. Oso baliagarriak dira, biak, hitzak nola erabili edo itzuli izan diren ikusteko.

Horrez gain, corpus elebakarraren gainean teknika linguistiko eta estatistikoak aplikatuta, gehien erabiltzen diren hiru motatako konbinazioak kalkulatu dira (izen-izen, izen-aditz eta izen-adjektibo) eta kontsultagai jarri da. Hala, sistemari galdetu diezaiokegu hitz jakin bat zein aditzekin konbinatu ohi den, edo zein adjektiborekin, eta abar.

Web-corpusen Atariaren argitaratzea jauzi kualitatibo bat da, lehen aldia baita webetik automatikoki erauzitako corpusak publikoaren eskura jartzen direna, eta baita kuantitatiboa ere, corpusen tamainan aurrerakuntza esanguratsua baita. Koldo Mitxelenak zioen euskararen benetako misterioa ez dela jatorria, iraupena baizik. Misterio handiagoa da etorkizunean iraungo ote duen. Guk ez dugu horren erantzunik, baina iraungo badu zalantzarik gabe hizkuntza-teknologietan presente egon behar du euskarak. Web-corpusen Atariarekin norabide horretan beste pausu bat eman dugula sinetsita gaude Elhuyarren.●

"Web-corpusen atari berriak 125 milioi hitzeko euskarazko corpus orokor handi bat eta 18 milioi hitzeko euskara-gaztelania corpus paralelo bat ditu"