Herria aldizkariak bere testu-masa osoa Euskaltzaindiaren eskuetan jarriko du, Lexikoaren Behatokia elikatzeko

Euskaltzaindiaren eginkizun garrantzitsuenetako bat euskaldunei euskara txukun erabiltzen laguntzea da

Zer da, ordea, hizkuntzaren erabilera txukuna? Nola jakin dezakegu zer den erabilera txukuna eta zer ez?

Garai batean jakintsu batzuen esku zegoen hori. Haien erudizioa zen egokitzat hartzen zen hizkeraren oinarria. Gaur egun, ordea, beste irizpide batzuk erabiltzen dira.
Hizkuntzalaritzan jarduteko, ezinbestekoa da hizkuntzaren benetako erabileraren berri ematen duten testuez osaturiko corpusak eratzea eta ustiatzea.

Corpusak benetako testuen bildumak dira. Hau da, liburu, egunkari eta dokumentuetatik hartzen diren testu multzo egituratu handiak dira. Horiek arakatuz ikusten da hitzen edo egitura gramatikalen erabilera zein izan den. Zer den ohikoa eta zer ez. Zer den idazle askok erabiltzen dutena. Zer den denboran zehar irauten duena eta zer aldatzen dena. Zer den hizkera arruntean erabiltzen dena eta zer hizkera jasoan erabiltzen dena, etab. Informazio hori guztia eskuetan dela, erabakitzen da gaur egun zer den txukuna eta zer ez.

Halaxe egin du Euskaltzaindiak *Hiztegi Batua* eta *Euskaltzaindiaren Hiztegia* lantzeko. Gaur egun Euskaltzaindiak hiru corpus ditu: *Orotariko Euskal Hiztegia*rena, antzinadanik XX. mendea arteko testuen bilduma dena, *XX. mendeko euskararen corpus estatistikoa* eta hemen aurkezten ari garen *Lexikoaren Behatokia*rena.

Corpusak eta informatika

Ordenagailua asmatu baino lehen ere erabiltzen ziren corpusak, esate baterako Bibliako aipuen bildumak egiten zituzten Antzinatean eta Erdi Aroan ere, baina eskuz biltzen eta aztertzen zituzten, lan eskerga eginez. Ordenagailua agertu eta gutxira hasi ziren AEBetan ingelesezko testuak corpus elektronikoetan biltzen eta lantzen, eta 1980ko hamarkadatik aurrera hainbat hizkuntzatako corpusak sortu dira.

Lexikoaren Behatokiaren ezaugarriak

Corpusak hainbat eratakoak izan daitezke helburuaren arabera. Lexikoaren Behatokiaren tankerakoei corpus monitoreak esaten zaie. Corpus hauek etengabe hazten dira, urtero testuak gehituz, hizkuntza denboran zehar nola aldatzen ari den ikusteko. Batez ere hiztegigintzan erabiltzen dira. Hazkunde horretan, testuak jasotzeko garaian, irizpide jakin batzuk erabiltzen dira, garai desberdinetako emaitzak konparagarriak izan daitezen. Dena den, jasotzen diren testu motek ez dute nahitanahiez erabilera errealean duten pisua islatu behar izaten.

Bada beste corpus mota bat, erreferentzia-corpusa deitzen dena. Erreferentzia corpusetan hizkuntzaren erabilera-mailek, generoek eta gaiek benetako erabileraren araberako pisua behar izaten dute eta oso handiak izaten dira, gutxienez 50 milioi hitzekoak.

Lexikoaren Behatokia corpus monitorea izan arren, erreferentzia-corpus bilakatzeko moduan lantzen da. Horretarako, testu guztiak katalogatzen dira eta ezaugarri batzuekin markatzen, etorkizuneko erreferentzia-corpusak behar duen oreka izan dezan.

Aurten 43,4 milioi testu-hitz izango ditu eta urtean-urtean 8 milioi inguru hazten da.

Corpus batek gaur egun, tamainaz aparte, izan behar dituen beste ezaugarri batzuk baditu Lexikoaren Behatokiak. Corpus anotatua da, hau da, dauzkan testu guztiek, ezkutuko marka batzuen bidez, bereiziak ditu egitura-elementuak (paragrafoak, aipuak, arrotz-hitzak, puntuen erabilera, etab.) eta hitz bakoitzaren ezaugarri linguistikoak (lema, kategoria, azpikategoria eta kasua).

Lan hori guztia egiteko, Euskaltzaindiak hitzarmenak sinatzen ditu, alderdi batetik, testuen hornitzaileekin, hala nola, Deia, Berria, Argia, ETB, eta beste hainbat hornitzailerekin, bereziki hedabideekin, eta, bestetik, corpusa lantzen duten hiru erakundeekin: UZEI, Elhuyar eta EHUko Informatika Fakultateko IXA taldearekin.

Orain Euskaltzaindiak hitzarmen garrantzitsu bat sinatu du, Lexikoaren Behatokiari ekarpen aberatsa egingo diona: *Herria* aldizkariak bere testu-masa osoa Akademiaren eskuetan jarriko du Lexikoaren Behatokia elikatzeko.

Oso garrantzitsua da lan hori egiteko erabiltzen den teknologia. Izan ere, ia prozesu osoa modu automatikoan egiten da, eta bada aipagarria den beste alderdi bat ere: corpus horrek, hizkuntzalaritzarako ez ezik badu interesa hizkuntza-teknologiako tresnak garatzeko eta probatzeko ere. Hau da, teknologiaz baliatzen da eta teknologia hobetzeko balio du.

Esan dezadan bukatzeko, edonork kontsulta dezakeela Lexikoaren Behatokiaren edukia, Euskaltzaindiaren webguneko helbide honetan:
http://lexikoarenbehatokia.euskaltzaindia.net

Andoni Sagarna,
Lexikoaren Behatokia egitasmoaren zuzendaria