Montse Maritxalar Anglada
UPV/EHUko Donostiako Informatika Fakultateko irakasle eta ikertzailea

Corpusak

Euskal Wikipediako «testu corpus» sarreran esaten den moduan: «Hizkuntzalaritzan, corpus edo testu corpusa nolabait egituratutako testu sorta handi bat da. Hizkuntza deskribatzeko eta ikertzeko baliatzen da, hizkuntza datu-bilduma gisa, lexikoaren, morfosintaxiaren edo semantikaren aldetik. Latinetik datorren hitza da, eta euskaraz gorputza esan nahi du».

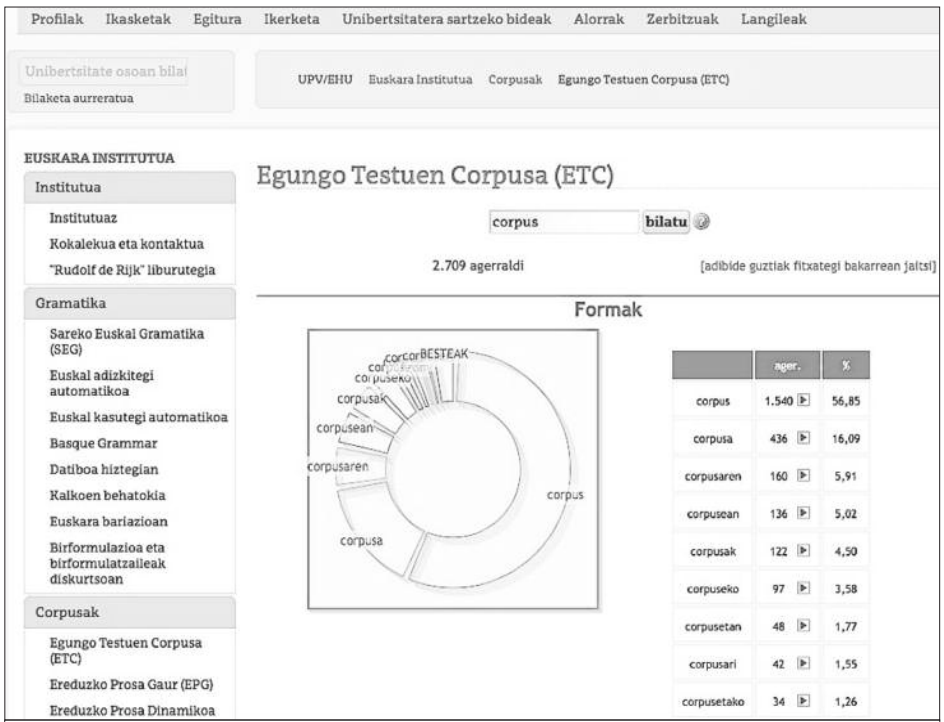
Corpusak formatu elektronikoan egon ohi dira, beraien analisia era automatiko edo erdi-automatikoan tresna informatikoen bidez egin ahal izateko. Eta hasierako erabilpenak hizkuntzaren ikuspegi linguistiko huts batetik begiratuta bazeuden ere, gaur egun azterketa kulturalak, hizkuntzalaritza historikoa, psikolinguistika, soziolinguistika eta abar luze bat aztertzeko ere erabiltzen dira corpusak. Corpusekin lan egitearen abantailarik handiena datu errealekin lan egitea da. Izan ere, hizkuntza egoera errealetan nola erabiltzen den irudikatzen dute. Corpusak egunerokotasunean ematen den hizkuntz erabileraren isla dira, bai hizkuntz jasoaren isla, baita esparru informal desberdinetan erabiltzen den hizkeraren isla ere.

Corpusen ustiaketaren bitartez corpusetan errepresentatuta dagoen informazioa era desberdinetan kontsultatzeko aukera eskaintzen zaigu. Alde batetik, *online* eskuragarri dauden corpusek, zenbaitetan, interfazeen bidez bilaketak egiteko aukera eskaintzen dute. Eta, beste alde batetik, badira tresnak erabiltzaileari aukera ematen diotenak bere corpus propioak kargatzeko eta unean bertan igotako corpusaren gaineko ustiaketa egin ahal izateko. Azken horien kasuan testuak automatikoki tratatzen dira eta makinak berak identifikatzen du testuan errepresentatuta dagoen informazioa. Horrek eskatzen du hizkuntzaren prozesamendu automatikoa egitea.

Euskararako badira hainbat corpus ustiatzeko baliabideak eskuragarri. Hona hemen Interneten atzigarri aurki ditzakegun batzuk: Zientzia eta Teknologiaren Corpusak (ZT), zientzia eta teknologiaren alorreko euskarazko testu-bilduma egituratu eta etiketatua gordetzen duena jakintza alorraren eta testu motaren arabera; eta Egungo Testuen Corpusa (ETC), XXI. mendeko liburuetan eta prentsa idatzian oinarritzen dena, besteak beste, hitzen agerpenak urteetan zehar nola gertatu diren ikus daitekeelarik. Bestalde, Interneten dugun "corpus" erraldoia ustiatzeko Corpeus Web aplikazioa erabilita, adibidez, "azkartu" hitza lema gisa bilatzeko eskatuz gero, "azkartzeko", "azkartzean", "azkartzen" eta "azkartuz" bezalako hitzek webean dituzten agerpenak bistaratzen dira.

Corpus elebakarrek hizkuntza bakar batean idatzitako testuak biltzen dituzten moduan, corpus eleaniztunek hainbat hizkuntzatako testuak jasotzen dituzte, adibidez, Eroskiko Consumer corpusa. Bertan, aldizkariko eduki guztia kontsulta daiteke corpusetan ohikoa den erara, besteak beste, agerpenak eta kopuruak. Gainera, corpus eleaniztuna izaki, hitza beste hizkuntzetan nola esaten den ere ikus daiteke. Horrelako corpusetatik abiatuta hiztegi antzeko baliabideak automatikoki sortzeko aukera ere badago. Horren adibide dugu atera berri den TextReference.com sareko hiztegia. Testuinguru hiztegi bat da, hitz eta esamoldeen ordainak itzulpen memoria handietan bilatzeko aukera ematen duena.

Bide berriak, eta testuak eta hiztegiak ulertzeko ikuspegi berriak eskaintzen dizkigu teknologiak. •



Irudian, Egungo Testuen Corpusa (ETC). GAUR8