Hiztegigintza teknologiaz modernizatzen

Igor Leturia Azkarate
Informatikaria eta ikertzailea

Hiztegigintzan, beste ia edozein jardueratan bezala, aldaketa sakonak gertatu dira azken urteotan teknologiaren eskutik. Papera oinarri eta helburu izatetik eta eskuzko lan handia eskatzetik, pasatu gara testu eta corpus elektronikoak erabiltzera, prozesuaren zati handi bat automatizatzera eta argitaratzeko euskarri digitalak erabiltzera. Elhuyarreko hiztegigintzan ere eman dugu modernizazio-pauso hori, hizkuntza-teknologiak lagun.

Elhuyarreko lau sail nagusietako bat Hizkuntza eta Teknologia izenekoa da. Haren barruan, beste hiru azpisail daude: itzulpen-zerbitzuak, hiztegigintza eta hizkuntza-teknologiak. Hizkuntza-teknologiak asko dira, eta arlo ugaritan dira erabilgarriak. Eta guk ere arlo askotarako baliagarri direnak ikertu, garatu eta merkaturatzen ditugu; baina, normala denez, Elhuyarren beste arloetarako baliagarri direnak lantzen ditugu bereziki. Esaterako, itzulpen-zerbitzuetan lehia-abantaila eman dezaketen itzulpen automatikoa eta itzulpen-memorien teknologiak lantzen ditugu, eta hiztegigintzarako interesgarriak diren teknologia anitz ere bai.

LAN-PROZESUA ERRAZTEN

Hiztegigintzan egin beharreko lanetako bat hitz-hautaketa da. Horretan laguntzeko tresnak garatu ditugu. Horietako bat Erauzterm da. Euskarazko arlo jakin bateko corpus espezializatu bat emanik, Erauztermek han agertzen diren arlo horretako terminoak detektatzen ditu.

ElexBI tresnak antzeko zerbait egiten du, baina elebitan. Corpus paralelo batetik (elkarren itzulpen diren testuen bilduma) termino-baliokidetzak erauzten ditu, hau da, bi hizkuntzetako termino-bikoteak. Tresna hori web-zerbitzu gisa jarria dugu, Itzulterm izenarekin. Eta tresna hori erabiliz egin da Lanbide Heziketako hiztegia.

AzerHitz-ek ere ElexBI-ren gauza bera egiten du; baina lehengaitzat corpus paraleloak hartu beharrean corpus konparagarriak erabiltzen ditu. Horiek, elkarren itzulpen izan gabe, gai bera tratatzen duten testu-bilduma eleanitzak dira.

Testuetatik informazio lexikografikoa ateratzeko beste tresna bat Konbitz da. Hark euskarazko testuetatik ohiko konbinazioak, kolokazioak, fraseologia eta horrelakoak erauzten ditu.

PiboLex tresna ere badugu, hiztegi berriak sortzen dituena bi hiztegi eta zubi-hizkuntza bat erabilita. Harekin sortutako euskarazko bost hiztegi online jarri genituen automatikoki eraikitako hiztegien atarian (ikusi 299 zk.).

LANERAKO LEHENGAIA, CORPUSAK

Ikusi duzuenez, teknologia horietako askok corpusen beharra dute, eta horregatik da corpusgintza digitala asko lantzen ditugun arloetako bat. EHUko IXA Taldearekin batera, Zientzia eta Teknologiaren Corpusa sortu genuen; Eroski Fundazioarentzat *Consumer* aldizkariko corpus eleanitza osatu genuen; eta, Euskaltzaindiarentzat, Lexikoaren Behatokiko Corpusa osatzen ari gara IXA Taldearekin eta UZEIrekin batera.

Hala ere, corpusak egitea garestia denez, corpusak osatzeko weba erabili ahal izateko tresnak sortzen ari gara azken urteotan. Internet corpus gisa kontsultatu ahal izateko, CorpEus web zerbitzua jarri genuen online duela urte batzuk. Eta webetik automatikoki corpus orokor handiak, corpus espezializatuak, corpus paraleloak eta corpus konparagarriak sortzeko tresnak ere baditugu. Webetik automatikoki eraikitako euskarazko corpus orokor handi bat, euskara-gaztelania corpus paralelo handi bat eta lehen aipatutako Konbitz tresnaren bidez corpus orokor handitik erauzitako konbinazioak kontsultagai jarri genituen Web-corpusen Atarian (ikusi 294 zk.).

ELHUYAR HIZTEGIEN WEBGUNE BERRIA

Hiztegigintzako lan-prozesua errazteaz eta lehengaitarako corpus elektronikoz hornitzeaz gain, teknologiak, oro har, eta hizkuntza-teknologiek, bereziki, asko hobetu dezakete hiztegi-erabiltzaileen esperientzia. Duela urte batzuk hiztegiak webean jartzen hasi zirenetik, gehienetan eskaini izan da bilaketa-kutxen aukera, bilaketa azkarrak egin ahal izateko, alfabetikoki ordenatutako zerrenda batean bilatzera joan beharrean (nahiz eta badauden hiztegien PDFak online jarri besterik egiten ez dutenak). Baina bilaketa egin ondoren eskaintzen diren

emaitzak papereko hiztegiek eskaintzen dituztenak bezalakoak izaten dira. Berriki aurkeztu den Elhuyar Hiztegien webgune berrian (http://hiztegiak.elhuyar.org/), non euskara-gaztelania, euskara-frantsesa eta euskara-ingelesa hiztegiak baitaude, harago joan nahi izan dugu, eta aukera aurreratuagoak eskaini.

Adibidez, bilatu den hitz bat nola ahoskatzen den entzun daiteke, bi aukera baliatuta: Forvo webgunean erabiltzaileek grabatutako audioen bidez, edo TTS (text-to-speech edo ahots-sintesia) teknologiaren bidez, hau da, ordenagailuz sortutako ahots sintetikoa baliatuz. Erabiltzen dugun TTS sistema AhoTTS da, EHUko Aholab Taldeak garatutakoa eta guk merkaturatzen duguna.

Horrez gain, hitz bat bilatu nahi dugunean, hitza tekleatzen hasi ahala, hasiera hori duten hitzen zerrenda erakusten digu; hala, ez dago dena idatzi beharrik, eta oker idazteko aukerak murrizten dira.

Bestalde, hitzen adibideei dagokienez, hiztegian egileek sartzen dituzten ohikoez gain, lehen aipatu dugun webetik erauzitako euskara-gaztelania corpus paraleloan aurkitzen diren adibideak ikusteko aukera ematen du webgune berriak. Adibide horiek ez dira helburu-hizkuntzakoak soilik, elkarren itzulpen diren esaldipareak baizik.

Gainera, jatorri-hizkuntzako sarreren gaineko ohiko bilaketaz gain, helburu-hizkuntzako sarreretan bilatzeko aukera ere ematen da. Eta etorkizunean adibideetan ere bilatzeko aukera eskaini nahi da.

Hiztegia pertsonalizatzeko aukerak ere eskaintzen dira, hala nola egindako azken bilaketak gordetzea, bilaketa batzuk norberaren gogokoenen zerrenda batean gordetzea eta abar.

Oraingoz berrikuntza horiekin argitaratu badugu ere, etorkizunean poliki-poliki gauza gehiago sartzeko asmoa dago. Adibidez, zuzenean gorago aipatu dugun konbinazioen bilatzailera joateko aukera, beste hiztegi eta corpus batzuetako emaitzak ere erakustea, hitz bat gaizki idatzi denean zuzena proposatzea, bilatutako hitzaren deklinazioak edota inflexioak erakustea...

ETA ARE GEHIAGO ETORKIZUN!

Gainera, datozen urteetan are gehiago teknologizatu nahi dugu gure hiztegigintza-saila. Berrikuntza nagusia automatizazioaren arlotik etorriko da. Orain arte landu ditugun mota horretako teknologia gehienek hiztegirako hitz eta terminoak eta haien ordainak erauzten zituzten corpusetatik; baina, horiez gain, hiztegi batek definizioak, adierak eta adibideak behar ditu. Bada, orain hasi gara horiek modu automatikoan lortzeko modua ere lantzen, hau da, hitz baten definizioak, adierak eta horietarako adibide egokiak testuetatik edota webetik automatikoki erauzten.

Lehenagotik genituen hizkuntza-teknologiak ustiatzen jarraituz eta abiarazi berri ditugunak garatuz, Elhuyarren hiztegigintza punta-puntakoa izatea lortu nahi dugu, mundu gero eta globalizatuago honetan euskarak beste hizkuntzekin harremanetan jarraitu ahal izan dezan.

"definizioak, adierak eta adibideak testuetatik edo webetik automatikoki erauzteko modua lantzen hasi gara"