Testu corpusak eta hizkuntza-plangintza

Xavier Gomez Guinovart

Vigoko Unibertsitatea

Itzulpena: Pello Goikoetxea

Hizkuntzaren teknologian, testu corpusa testu bilduma bat da, ahozko nahiz idatzizko testuen bilduma, euskarri informatikoan gordeta izaten dena. Testu horiek, hizkuntza-aldaera baten edo aldaera multzo baten adierazgarri izaten dira, hizkuntza edo filologia ikerketarako bilduak izaten dira eta erabilera bat baino gehiago izan dezakete hizkuntzaren industrietan.

Corpus bakoitzaren edukia corpusa sortzekoan jarritako helburuen baitan egongo da. Oro har, ordea, badira alderdi linguistiko batzuk, corpus baten ezaugarriak eratzeko orduan erabakigarriak izan daitezkeenak. Esate baterako, corpusean jasotako testuen ekoizpen modua (ahozkoak diren, idatzizkoak edo bietakoak), testuen kronologia (historikoak, gaurkoak edo diakronikoak), hizkuntza mota (hizkuntza orokorrekoak edo aldaera berezituetakoak), testuetan diren aldaerak (testu elebakarrak, hainbat hizkuntzatako testuak edo testu itzuliak), hizkuntzaren gaineko informazio erantsirik duten (oharrik gabeko testuak edo oharrak dituztenak) edota biltzeko orduan erabilitako hautapen teknikak (corpus irekiak eta itxiak, corpus orekatuak, lagin corpusak edo estatistikoak).

Beste alde batetik, hizkuntza-plangintza lurralde bateko errealitate soziolinguistikoa aldatzeko ahalegin bat da, hizkuntza politikako ekintza koherente sail bat aplikatuta. Hizkuntza gutxituak edo normalizatu gabeak dituzten herrialdeetan, hizkuntza-plangintzaren helburu nagusia hizkuntzaren normalizazioa izaten da.

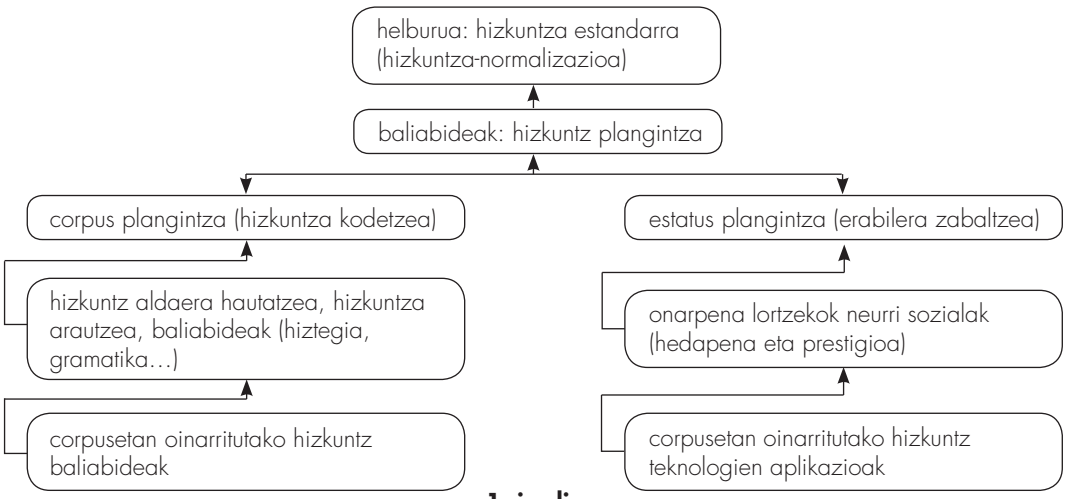
Hizkuntza-plangintza batez ere bi eskuhartze arlotan erabiltzen da: hizkuntzan eta gizartean. Hizkuntzaren alorrean, hizkuntzaren estandarizazioarekin lotuta izaten da, eta ortografia eta gramatika araudi bat egitea izaten da, hiztegi erreferentzial bat eta

oinarrizko beste baliabide batzuk. Soziolinguistikaren terminologian, "corpusaren plangintza" esaten zaio plangintza horri, eta "corpusa" "hizkuntza-aldaera arauemailea" izaten da. Corpus plangintzaren helburua hizkuntza-aldaera estandar bat egitea izaten da, hizkuntzaren gizarte eginkizun guztietarako ondo balio izateko modukoa.

Gizartearen arloan, berriz, hizkuntza-plangintzaren helburua hizkuntza baten estatus soziala aldatzea izaten da, eta, beraz, ez da harritzekoa "estatus plangintza" izatea soziolinguistikak eginkizun horretarako aukeratutako terminoa. Estatusaren plangintzak bi gauza lortu nahi izaten ditu: aukeratu eta arautu den corpusa onartzea gizarteak, eta corpus hori hizkuntzaren gizarte funtzio guztietan erabil dadin sustatzea.

Hizkuntzaren alderdiei edo gizartearen alderdiei, gehiago zeri eragiten dieten begiratu eta bi sailetan sar daitezke testu corpusek hizkuntza-plangintzan izan ditzaketen erabilerak. Hizkuntzari dagokion arloan, hau da, corpusaren plangintzari dagokionean, testu corpusak oinarri enpiriko nahitaezkoak dira hizkuntza-baliabideak egiteko (gramatikak eta hiztegiak), eta baita hizkuntza arautzeko ardura duten erakundeek erabaki beharreko arauak erabakitzeko ere. Bestalde, plangintzaren alderdirik sozialenean, hau da, estatusaren plangintzan, corpusen aplikazioen aldakortasuna nabarmendu behar da, nola hizkuntzaren ikaskuntzan eta ikerkuntzan hala hizkuntza-teknologien garapenean (itzulpen automatikoan, esate baterako), eta teknologia horien papera funtsezkoa da erabilera handitzeko nahiz gizarteak hizkuntza baten ospearen pertzepzioa izateko. Azkenik, estatus plangintzaren baitan, hor dugu corpus linguistiko nazionalek hartzen duten funtzio sinbolikoa ere, hizkuntzaren batasunaren, independentziaren eta indarraren ikur gisa.

Laburpen modura, diagrama honen bidez ikus dezakegu testu corpusen eta hizkuntza-plangintzaren arteko harremana (1. irudia).



1. irudia

Hizkuntza-corpusak eta corpusaren plangintza

Hizkuntzaren lexikoak eta gramatikak hizkuntza-baliabide oinarrizkoak dira hizkuntza estandarizatzeko. Testu corpusak ustiatuta nabarmen hobetu daitezke baliabide horiek. Adibidez, hizkuntza baten lexikoaren ortografia arautzeko orduan, hitz bakoitzaren formak sinkronikoki eta diakronikoki izan dituen aldaerak ikusteko erabil daiteke corpusa, eta, ondoren, behaketa enpirikoaren bidez, egokiena jotzen dena aukera daiteke, forma bakoitzaren maiztasuna eta banaketa kontuan hartuta, generoka, aroka, lurraldeka eta are autoreka ere. Lehian diren formen artean forma lexiko arauemaileak aukeratzeko, hizkuntza eta etimologia irizpideei ez eze, erabilera arruntari eta kanon literarioko autoreen erabilerari ere begiratzen zaio. Bi erabilerei begiratzeko ere oso egokia izan daiteke corpusa kontsultatzea, eta baita testu bakoitzean zer forma aldaera den ikusteko ere.

Hiztegiak eta gramatikak egiteko, berriz, testu corpusen ustiaketak aukera ematen du hitz batek erabiltzen den testuinguruaren arabera zer esanahi dituen zehazten saiatzeko, praktikan beste zein hitzekin konbinatzen den ikusteko, zein eraikuntza gramatikaletan erabiltzen den ikusteko eta testuetan zer lexiko harreman eta definizio kodetu diren aurkitzeko. Gainera, corpusean kodetutako datu testualen azterketa kuantitatiboa eginez gero, hitzen, hitzen adieren, itzulpenen eta eraikuntzen benetako erabileraren maiztasuna ere atera daiteke.

Orain arte, beste lexiko errepertorio batzuetan, kanon literarioko testuetako aipamen hautatuetan edota hizkuntzaren hiztun ziren aldetik zuten senean bilatzen zituzten lexiko errepertorioak egiten zituztenek hizkuntzari buruzko datuen informazio iturriak. Lan egiteko modu horrek muga handiak zituen lexikografia praktikarako. Alde batetik, lexikografoek lexikoaren erabileraren gainean egiten zituzten gogoetak ez zetozelako bat hizkuntzaren errealitatearekin; bestetik, lan kanonikoetako aipuak eskuz biltzea oso mantso egiten zen lana, eta etekin handirik ere ematen ez zuena, izaten zelako; eta, azkenik, inspirazio iturri erabiltzen ziren hiztegiak ez zirelako hiztegi eguneratuak izaten, eta, batzuetan (okerrenean), akatsak ere izan zitzaketelako urteak joan eta urteak etorri akats berbera kopiatuta.

Lexikografia lanean testu corpus informatizatuak sartzeak orain arte erabili izan den metodologiak zituen muga horiek gainditzea ekarri du, zalantzarik gabe, eta hizkuntza bateko lexikoari hizkuntzan benetan duen erabileran (bere testugintzan) begiratzeko modua ekarri du. Hiztegiak egiteko corpusak aurreneko aldiz baliatu zituztenak Birminghamgo Unibertsitatea eta Collins argitaletxea izan ziren, Cobuild hiztegia egiteko (*Collins Birmingham University International Language Database*); hango irakasle John Sinclair zenak zuzendu eta 1987an argitaratu zen hiztegia. Garai hartan, Cobuild proiektua oso proiektu berritzailea izan zen, lehenbiziko aldia baitzen lexikografia lana egiteko testu corpus adierazgarri bat erabiltzen zena, hitzen esanahiak, patroi sintaktikoen identifikazioa eta hizkuntza bateko (garaiko ingeleseko) kolokazioen eta fraseologiaren deskribaketa aztertzen laguntzeko. Cobuilden arrakasta ikusi eta gero,

beste argitaletxe handi batzuk ere hasi ziren lexikografia lana corpusetan oinarrituta egiteko metodologi hura jarraitzen, esate baterako Oxford University Press, Longman eta Larousse argitaletxeak (*British National Corpus* egin zuten elkarren artean) eta Cambridge University Press.

Gaztelaniari dagokionez, metodologia horrekin egindako lanaren adibide ditugu *Gran diccionario de uso del español basado en el Corpus lingüístico CUMBRE*, Sociedad General Española de Libreríak 2001ean argitaratu zuena, Aquilino Sanchez irakaslearen zuzendaritzapean; edota *Redes* kolokazio hiztegia, Ignacio Bosquerena, 2004an argitaratu zena eta SM argitaletxearen kazetaritza corpus batean oinarrituta eginda dagoena (250 milioi hitzeko corpus batean).

Corpusetan oinarritutako lexikografia lanaren metodologia katalanean ere ari dira erabiltzen. Hala ari da, adibidez, Institut d'Estudis Catalans *Diccionari descriptiu de la llengua catalana* egiten, *Corpus Textual Informatitzat de la Llengua Catalana* oinarritzat hartuta.

Euskal Herrian, 2009ko otsailean atera zuen *Orotariko Euskal Hiztegia*ren 16 liburukietako azkeneko alea Euskaltzaindiak, euskararen akademiaren 90. urteurrenarekin batera. *Orotariko Euskal Hiztegia* konbinazio bat da: batetik, corpusean oinarritutako hiztegi historikoa da (Oxford English Dictionary-ren modukoa), eta, bestetik, hiztegien hiztegi edo corpus lexikografikoa da. Hasieran, 1984an, Koldo Mitxelena izan zuen zuzendari, eta hura hil zenean (1987an) Ibon Sarasola. *Orotariko Euskal Hiztegia*k 6.000.000 hitzeko corpusa du oinarri, XVI. mendetik hasi eta 1970. urtera bitartean euskalkietan idatzizko lanez osatutako corpusa.

Galizian, galizieraren erreferentzia corpusa, *Tesouro Informatizado da Lingua Galega* (TILG), Anton Santamariaren zuzendaritzapean prestatzen ari diren *Dicionario Histórico da Lingua Galegare*n oinarria izango da; *Corpus CLUVI* itzulpen corpusa *Dicionario CLUVI inglés-galego* egiteko baliatu zuten; eta *Corpus Técnico do Galego* (CTG) da *Termoteca* galizieraren datu terminologikoen bankuaren oinarrian dagoena.

Denborak aurrera egin ahala, testu corpus informatizatuak geroz eta gehiago erabiltzen dira hizkuntzen ortografia, lexikoa eta gramatika arautzeko tresna lagungarritzat, eta, hala, hizkuntza arautzeko ardura duten erakundeek testuen erabilera dokumentatuan ere oinarritu ditzakete beren arauak. Dena dela, corpusak ez dira hizkuntza-baliabideak eraikitzeko edo arauen gaineko erabakiak hartzeko oinarriak bakarrik. Testu corpusak, berez, hizkuntza-baliabide oinarrizkoak dira, zuzenean kontsultatu eta baliatu daitezkeenak kontsultarako aplikazio egokiak erabilita. Hala, corpusak zuzenean kontsultatzeak, hitz batek testu batzuetan, testuinguru batzuetan, dituen zentzuetan, itzulpenetan edota eraikuntza gramatikaletan benetan zer erabilera duen batere bitartekorik gabe ikusteko modua ematen digu, eta, batez ere, hiztegietan ez dauden hitzen, zentzuen, adieren edo itzulpenen gainean dokumentatzeko bidea.

Hizkuntza-corpusak eta estatusaren plangintza

Estatusaren plangintzari dagokionez, testu corpusak funtsezko baliabideak dira hizkuntza-teknologiak garatzeko. Teknologia horiek, berriz, behar-beharrezkoak dira hizkuntzaren erabilera informazioaren gizarteko ekipamenduetan sustatzeko (esate baterako, ordenagailuetan, eskuko telefonoetan, agenda elektronikoetan, MP3/MP4 irakurgailuetan edota bideo-jokoen kontsoletan), eta erabilera testuinguru horietan erabilgarritasuna eta prestigioa irabazten du hizkuntzak hizkuntza-plangintzarako interes estrategikoa duen erabiltzaile sektore handi batean.

Testu corpusak osagai oinarrizkoak dira hizkuntzak prozesatzeko tresnak egiteko. Gramatika etiketatzaileak (*taggers*ak), entitateen izenak identifikatzako tresnak (*named entity recognition*ak), analizatzaile sintaktikoak (*parsers*ak), probabilistikoak, etiketatzaile semantikoak eta lexikoaren esanahiari anbiguotasuna kentzeko programak (*word sense disambiguation*ak) dira, besteak beste, corpusetan oinarritutako erreminta batzuk; eta horiek ez balira, ez legoke hizkuntzaren teknologietako beste aplikazio konplexuago batzuk eraikitzerik: ez itzulpen automatikorik ez eta hizkera ezagutzeko tresnarik ere, konparazio baterako. Era berean, hizkuntzaren teknologietako aplikazio mordo baten funtzionamendu algoritmoak corpusetatik ondorioztatutako datuetan oinarritzen dira zuzen-zuzenean. Adibidez, testu igarlearen aplikazioak (eskuko telefonoetan erabiltzen dira), ortografia, gramatika nahiz estilo zuzentzaileak, testuak laburtu eta informazioa berreskuratzeko programak, galderei erantzuteko sistemak, adibideetan oinarritutako itzulpen automatikoa, estatistikan oinarritutako itzulpen automatikoa (Googlerena, esate baterako) edota diktaketa sistemak.

Hizkuntzaren eskurapena planifikatzeko aplikazioak ere badituzte testu corpusek, horietan oinarrituta egiten baitira material didaktikoak, hizkuntzaren hedapena nazio eta nazioartean handitzeko. Hala berean, testu corpusena material inportantea da dena delako hizkuntzaren ikerkuntza sustatzeko hizkuntza politikari eusteko ere. Eta, azkenik, ez dugu ahaztu behar corpus nazionalek badutela beste eginkizun bat ere: hizkuntzaren batasunaren, duintasunaren, independentziaren eta indarraren irudi publiko sinbolikoa izatea. Funtzio sinboliko hori hizkuntzaren hiztegi "ofiziala"ren parekoa da, baina, teknologiari lotutako baliabidea denez, indar handiagoa ere izan dezake.

Ondorioak

Gaur den egunean pentsaezina da hizkuntza-plangintza bat egitea hizkuntzaren testu corpusak eraikitzea eta ustiatzea kontuan izan gabe, direla hizkuntza bakarreko edo gehiagotako corpusak, gaur egungoak, historikoak, orokorrak nahiz berezituak. Testu corpusek paper nagusia jokatzen dute hizkuntzaren plangintzan; paper funtsezkoa alderdi guztietan: corpusaren plangintzan, eskurapenaren plangintzan eta estatusaren plangintzan.

Gizartean duen eragin izugarria dela eta, hizkuntza bat teknologietan izatea erabakigarria da hizkuntza horrek hizkuntza nor-

malizatuaren estatusa lortzeko edo estatus horri eusteko. Hizkuntza bat teknologian izatea, berriz, hizkuntza horretako hizkuntza-teknologiak garatzearekin lotuta dago oso estu, eta teknologia hori garatuko bada, nahitaezkoa izaten da, askotan behintzat, testu corpusak izatea. Hori dela eta, hizkuntza normalizatzeko lanek dituzten helburuen artean, behar diren testu corpusak eta hizkuntza-teknologiak egitea eta eskura jartzea ere izan behar luke, eta horrekin batera dena delako hizkuntza informazioaren gizarteko ekipamenduetan eta zerbitzuetan izatea ere bermatu behar lukete.

Corpus textuales y planificación lingüística

En este artículo, presentamos una perspectiva panorámica de las aplicaciones de los corpus lingüísticos textuales en los procesos de normalización y planificación de una lengua, teniendo en cuenta sus diversas vertientes en relación con la normativización lingüística o planificación del corpus, con la planificación de la adquisición y con la planificación del estatus social de la lengua, incluyendo en este apartado las aplicaciones de los corpus textuales en las tecnologías lingüísticas y la función simbólica de los corpus nacionales como emblema de la fortaleza de una lengua.

Textual corpuses and linguistic planning

In this article we present a panoramic point of view of the applications of textual linguistic corpuses in the processes of standardisation and planning of a language, taking into account their different approaches regarding the linguistic regulation or corpus planning, with the planning of the acquisition and the planning of the social status of the language, including in this sections the applications of textual corpuses with regards to linguistic technologies and the symbolic function of national corpuses as a sign of strength of a language.

Corpus textuels et planification linguistique

Dans cet article, nous présentons un point de vue panoramique des applications des corpus linguistiques textuels dans les processus de normalisation et de planification d'une langue, en tenant compte de ses différents aspects de réglementation linguistique ou de planification du corpus, de planification de l'acquisition et de planification du statut social de la langue, incluant dans ce passage les applications des corpus textuels dans les technologies linguistiques et la fonction symbolique des corpus nationaux comme emblèmes de la force d'une langue.