Lexikoaren Behatokia: leiho bat XXI. mendeko hedabideetako euskarari

Xabier Artola (Ixa Taldea, UPV/EHU),
Nerea Ezeiza (Elhuyar Fundazioa), Antton Gurrutxaga (Elhuyar Fundazioa),
Andoni Sagarna (Euskaltzaindia), Miriam Urkia (UZEI)

Sarrera

Euskaltzaindiak bere zereginen artean ditu hizkuntza lantzea eta aztertzea, eta erabilerari dagozkion arauak ematea. Hizkuntza zer bizi bat da, eta, arauak ganoraz emango badira, erabilera errealari erreparatu behar zaio. Bestalde, ez dago zalantzarik corpusak behar-beharrezkoak direla gaur egun hizkuntza baten erabilera erreala monitorizatzeko, ez baitago beste modurik hiztunek erabiltzen duten hizkuntza hurbiletik jarraitzeko.

Lexikoaren Behatokia (LB) proiektua Euskaltzaindiaren ekimenez abiatu zen, 2007an, Hiztegi Batuko Lantaldeak egindako proposamen bati erantzunez. Proiektuaren emaitza egun izen bera duen testu-corpusa da, zeina web bidez kontsultatu baitaiteke (Euskaltzaindia, 2008).

Proiektuak indarrean jarraitzen du, eta, hamar urte hauetan, ia 60 milioi hitzeko corpusa eratu da. Corpusa automatikoki prozesatuta eta linguistikoki etiketatuta dago, eta, beraz, hizkuntza-corpusek ohikoa duten kontsulta-funtzionalitatea eskaintzen dio erabiltzaileari.

Artikulu honen egitura honako hau da: sarrera honen ondoren, bigarren atalean, corpusaren beharra zergatik sortu zen aipatu, eta proiektuaren helburu nagusiak zein diren azalduko da. Hirugarren atalean, corpusaren ezaugarriak deskribatuko dira, labur-labur. Laugarren atalean, berriz, corpusa eratzeko lan-prozedura xehatuko da, hau da, testuen eskuratzea eta katalogazioa nola egiten den, corpuseratzearen nondik norakoak, eta prozesamendu linguistikoa zertan datzan. Bosgarrenean, corpusaren erabilera-kasu nagusia azalduko da: Euskaltzaindiaren Hiztegi Batuko Lantaldeak nola eta zertarako baliatzen duen corpusa bere lanean. Bukatzeko, corpusaren urtez urteko hitz kopuruak xehatu, eta web bidez egiten diren kontsulten berri emango dugu seigarren atalean, eta etorkizunerako asmoak zertan diren azalduko dugu zazpigarrenean.

Motibazioa eta helburuak

Gaur egun, zalantzarik gabe, hiztegia aztertzeko eta lan arauemailea egiteko, corpusetan oinarritu behar da, eta halaxe jokatu du Euskaltzaindiak azken urteotan. Testu klasikoak *Orotariko Euskal Hiztegia*ren (Euskaltzaindia, 2017) corpusean daude bilduta, eta XX. mendeko lagin aski adierazgarria dugu *XX. mendeko Euskararen Corpus Estatistikoa*n (UZEI, 2002). Bi iturri horiek izan dira *Hiztegi Batua* (Euskaltzaindia, 2016) egiteko oinarri nagusiak.

XXI. mendean sartuta gaude, ordea, eta euskara, zorionez, bizirik dago. Inoiz baino biziago, euskarazko testuen ekoizpenari dagokionez, eta, bizirik dagoenez, aldatuz ere joango da.

Hizkuntzaren bilakaerari hurbiletik jarraitu behar zaio; batetik, hitz eta adierazmolde berriak ezagutzeko, arauak zenbateraino betetzen edo urratzen diren jakiteko, eta arauak finkatzeko hori guztia ezagututa, eta, bestetik, hizkuntzaren erabilera sakonago ezagutzeko gramatika edo estilistika aldetik, erregistroen ezaugarriak aztertzeko edo ikuspegi soziolinguistikotik ikertzeko, nahiz edukien ideologia, historia eta abar aztertzeko.

Horrek guztiak corpusgintzan lanean jarraitzea eskatzen du. Euskaltzaindiaren ametsa erreferentzia-corpus handi, orekatu, lematizatu, etiketatu eta linguistikoki anotatu bat izatea da, eta badu esperantza amets hori hezurmamitzeko. Baliabide asko behar dira horretarako, eta denbora ere bai.

Bitartean, Lexikoaren Behatokia egitasmoa jarri du abian Euskaltzaindiak. Komunikabide eta argitaletxeek argitaratzen duten edo aireratzeko idazten duten materialarekin corpus monitore bat eraikitzea da helburua, hau da, hizkuntzaren erabilerari hurbiletik jarraitzeko corpus bat elikatzea, eta automatikoki lematizatzea eta etiketatzea pixkana-pixkana.

Gaur egun, Euskal Herrian badira tresnak eta ezagutza lan hori automatizatzeko, bai baitira aspaldidanik arlo horretan ikerketan ari diren lantaldeak. Euskaltzaindiak bidelagun ditu talde horietako batzuk egitasmo honetan. Corpusa elikatzeko testuak lortzeko, berriz, Euskaltzaindiak hitzarmenak sinatu ditu zenbait argitaldarirekin.

Corpusaren ezaugarriak

LB automatikoki prozesatutako hizkuntza-corpus monitore eta oportunista bat da. Gorago esan bezala, komunikabide eta argitaletxeek argitaratutako edo aireratzeko prestatutako testuekin osatua da. Proiektuan parte hartzen dute, Euskaltzaindiarekin batera, EHUko Donostiako Informatika Fakultateko Ixa Taldeak, Elhuyar Fundazioak eta UZEIk.

Testuetatik corpusera: lan-prozedura

Testuen eskuratzea eta katalogatzea

Corpusak eraikitzeko garaian, testuen jabegoarekin zerikusia duten kontuak zaindu beharra dago. Copyright-pean dauden testuak baimenik gabe erabiltzea ez da egokia, ezta ustiapen ko-

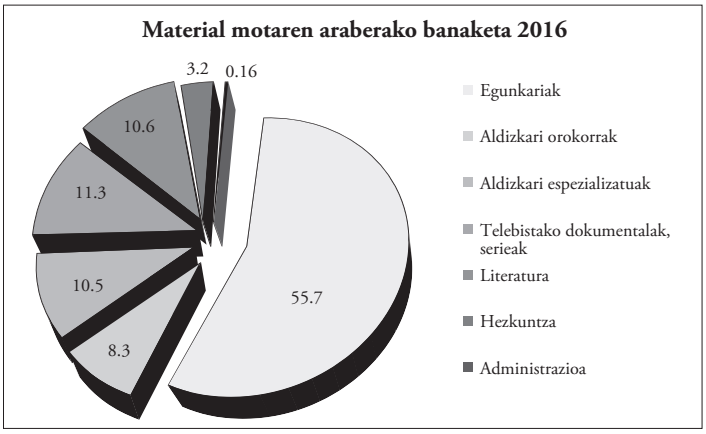
mertzialik ez duten corpusen kasuan ere. Horregatik, Euskaltzaindiak errespeturik handienaz erabili nahi ditu euskarazko testuak, eta, printzipio hori betetzeko, hitzarmenak sinatzen ditu argitaldariekin. 2017ko maiatzera bitartean, komunikabide, argitaletxe eta erakunde hauek laga dizkiote testuak Euskaltzaindiari: *Berria*, *Deia*, *Diario Vasco* eta *El Correo* egunkariak; EiTB; *Argia*, HABE, *Jakin* eta *Karmel* aldizkariak; AEK; Elkar argitaletxea; Elhuyar Fundazioa; Eroski; Laneki elkartea; Eusko Legebiltzarra; Arabako, Bizkaiko eta Gipuzkoako Batzar Nagusiak; eta Nafarroako Parlamentua.

Corpusa osatuko duen materiala (edukia, eta, ahal denean, baita metadatuak ere) hitzarmen bidez eskuratzen du, bada, Euskaltzaindiak. Informazio hori UZEIk jaso eta aztertzen du, informazio bibliografikoa eta edukiari dagozkion metadatuak inportazio automatikorako prestatuz. Izan ere, katalogoa gerora begira pentsatua dago, era horretan neurtzen baita erreferentzia-corpusa osatzeko informazioaren oreka.

Artikulu, liburu, saio zein araudi guztiak banaka katalogatzen dira eskuratu orduko, TEI egituran (TEI, 2002); alegia, datu bibliografikoak eta testualak jaso, eta automatizatzeko bidea abian jartzen da. Metadatuak, automatikoki eskuratu direnean, eskuz berrikusten dira beti, egokitasuna bermatzeko. Kasuren batean, baina, eskuz landu behar izaten dira, egileek ez baitute halakorik eskuratzerik. Prozedura ezarrita dagoenez, etengabe jasotzen dira argitalpenak, eta katalogoan lantzeko inportatzen. Eta, lanok bukatu ondoren, metadatuak eta testuak Elhuyarren esku uzten dira, etiketatu ditzan.

Lexikoaren Behatokia lantzeko, zenbait datu ez dira behar, baina proiektuaren helburua geroari begira pentsatua dago; bideratu beharreko erreferentzia-corpusari, alegia. Horregatik, katalogazio-lan orokorra egiten da —eremua eta gaia, adibidez—, Sailkapen Hamartar Unibertsalean oinarrituta (McIlwaine, 2010), gerora baliagarri izan dadin.

Orain arte landutakoaren banaketa grafiko honetan ikus daiteke:



1. irudia. Generoaren arabera, fikzioa % 16 da, eta ez-fikzioa, berriz, gainerako % 84a. 2016. urtearen bukaeran, 181.115 obra (artikulu zein liburu) zeuden datu-basean landuta.

Testuak corpuseratzea TEIren arabera: formatu-bihurketak eta egitura-etiketatzea

Corpus gordina eta katalogoko metadatuak corpuseratzea da urrats honen helburua, eta *egitura-etiketatzea* deritzo.

Dokumentuak corpuseratzeko lehen urratsa haien jatorrizko formatua LB corpusaren XML formatu bihurtzea da. Corpusak kodetzeko eta etiketatzeko proposatu diren ereduen artean, TEI P4 eredua (TEI, 2002) hautatu dugu.

Jatorrizko dokumentuaren formatua corpusaren XML formatura bihurtzeko, berariazko bihurtzaileak garatu ditugu. Haien bidez, .html, .xhtml, .rtf, .doc eta .odt formatuko dokumentuak prozesatzen dira. Batzuetan, jatorrizko formatua .pdf, .indd edo .qxd izaten da, eta, XMLra bihurtu aurretik, formatu hori duten dokumentuak aurreko bost formatuetako batera bihurtu behar dira. Formatu horietako batzuek arazoak sortzen dituzte formatu-bihurketa automatikoa egiteko, eta eskuz zuzendu behar izaten dira.

Formatu-bihurketa egitean, dokumentuaren egitura-elementuak etiketatzen dira (atalak, atalburuak, paragrafoak, taulak, buletdun zerrendak, etab.), eta zenbait ezaugarri tipografiko ere bai (letra-estiloak, hala nola etzana eta lodia, komatxoak, etab.).

Azkenik, XML dokumentu bakoitzaren goiburuan, dagokion obraren metadatuak biltzen dira (izenburua, egilea, argitaratze-urtea, argitaratzailea, arloa, erregistroa…). Metadatu horiek zuzenean ekartzen dira goiburura katalogoaren datu-basetik.

Egitura-etiketatzea egindakoan, corpusaren aurreprozesamendu linguistikoa egiten da. Prozesu horretan, hizkuntza-teknologia erabiltzen da.

Uneko datu-base lexikalean ez dauden eta Eustagger lematizatzaileak (Ezeiza *et al.*, 1998) ezagutzen ez dituen hitz berriak aurkitu, eta lexikoi osagarrian sartzen dira. Hori eginez gero, hurrengo urratsean (etiketatze linguistikoan), lematizatzaileak hitz horiek ezagutuko ditu, eta lan lexikografikorako ustiaketa eraginkorragoa izango da. 2016ra bitartean, 10.544 hitz sartu ditugu lexikoi osagarrian.

Euskarazkoak ez diren pasarteak etiketatu egiten dira, hurrengo urratsean linguistikoki analizatzea saihesteko. Prozesu hori erdiautomatikoa da (hizkuntza-identifikazio automatikoa eta ondorengo orrazketa). Bestalde, batzuetan, komeni da akats tipografikoak zuzentzea eta aldaera ez-estandarrak normalizatzea; esaterako, testua orraztu gabea denean edo euskara ez-estandarrean idatzia denean. Horrelakoak etiketatzea interesgarria da etiketatze linguistikoa errazteko eta eraginkorragoa izateko. Kasu horietan, baina, jatorrizko testu-hitza beti gordetzen da bilaketaren emaitzetan bistaratu ahal izateko.

Atal honetan deskribatutako lanak egiteko, Corpusgile tresna erabiltzen dugu. Corpusgintza kudeatzeko eta prozesuak exekutatzeko aukera ematen du.

Prozesatze linguistikoa

Egitura-etiketatzea eta aurreprozesatze linguistikoa egin ondoren, corpusa Ixa Taldearen esku uzten da linguistikoki prozesatzeko. Prozesatze linguistiko horren barruan daude tokenizazioa

(testuko hitzak eta puntuazioa ezagutzea eta banatzea), segmentazio morfologikoa (morfema-banaketa) eta analisi morfologiko osoa; azken biak Eustagger tresnaren bidez gauzatzen dira. Prozesu horren ondoren, testuak lematizatuta, kategoria aldetik etiketatuta eta automatikoki desanbiguatuta geratzen dira. Prozesuaren emaitza, beraz, corpus linguistikoki etiketatua edo anotatua da; anotazio horien formatua eta egitura AWAren arabera egiten da (AWA, Annotation Web Architecture: Anotazio Amaraunaren Arkitektura; Artola *et al.*, 2009).

AWAren arabera anotatutako corpusean, anbiguotasuna dagoen kasuetan informazio guztia gordetzen da, hau da, hitzaren analisi guztiak gordetzen dira. Jakina da hitz batzuk morfologikoki anbiguo suerta dakizkiokeela makinari, analisi bat baino gehiago onar baititzakete; adibidez, *pilotari* formak *pilotari* lema edo *pilota* lema (*pilota* + *-ari*) izan ditzake. Desanbiguazio-prozesuan, Eustaggerrek testuingurua aztertu eta analisi horietako bat lehenetsiko du automatikoki, eta hori erabiliko da bilaketa-sisteman. Hala ere, zuzentzat jotakoa ez ezik, gainerako analisi guztiak ere gordetzen dira, eta, hala, nahi izanez gero, beti egongo da gero eskuz landu eta zuzentzeko aukera (leheneste automatikoa okerra izan den kasuetan, noski).

Corpusa edizio bakoitzean berriro osorik etiketatzen denez (urtero), etiketatzea uniformea da; alegia, azken edizioan hitz bat lexikoi osagarrian sartu bada, aurreko edizioetako agerpenak ere ezagutuko ditu Eustaggerrek.

Lexikoaren Behatokia webean

Corpusa weberatzea

Etiketatze linguistikoaren emaitza prozesatuz, corpusaren kontsulta-aplikazioaren datu-basea eratzen da. Kontsulta-aplikazioa Lucene teknologian (Hatcher eta Gospodnetic, 2004) oinarrituta dago. Ideia nagusia zera da: agerpen bakoitza dokumentutzat hartzen da, eta kontsulta bat egitea, beraz, *query* edo galdera bati erantzuten dioten dokumentuak eskuratzea da. Formen eta lemen maiztasunak kalkulatzen dira, eta MySQL taula batzuetan gordetzen, bilaketa eraginkorragoa izan dadin.

Web-aplikazioaren ezaugarriak

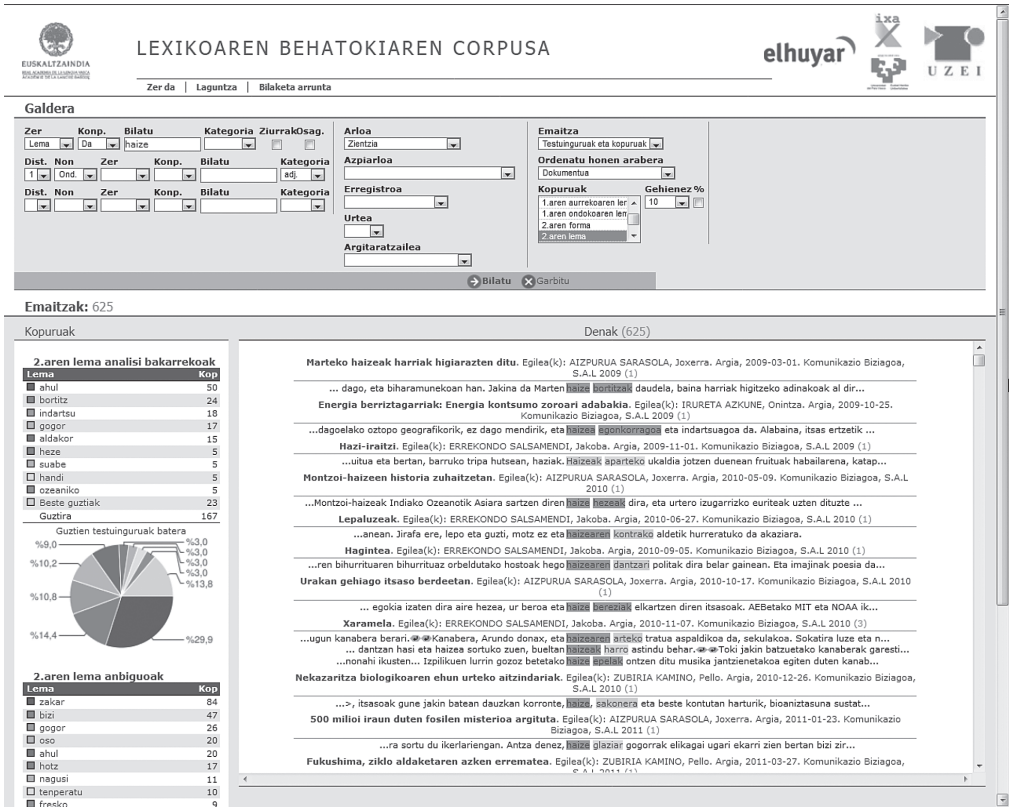
LBren webguneko laguntza-atalean, corpusa kontsultatzeko argibideak eta adibideak erantsi ditugu. Badira bi bilaketa mota: bilaketa arrunta eta bilaketa aurreratua. Labur beharrez, funtzionalitate jakin batzuk aipatuko ditugu hemen.

Bilaketa arruntean, hitz baten forma edo lema izan daiteke bilagaia (zehatza zein halako karakterez hasia zein bukatua). Kategoria iragazkitzat erabil daiteke, eta emaitzak ordenatzeko irizpide batzuk eskaintzen dira. Hitz baten lema bilatu eta emaitzarik ez dagoenean, komeni da 'forma + hasi' bilatze-modua erabiltzea; izan ere, nahiz eta lexikoi osagarriaren bidez datu-base lexikalaren estaldura handitu dugun, badira oraindik Eustaggerrek ezagutzen ez dituen lemak.

Bilaketa aurreratuan, aukera gehiago ditugu, eta corpusa kontsultatzeko ahalmena handiagoa da.

Adibidez:

- Badira hiru bilagai-errenkada, eta lema edo forma batzuen segida bilatu dezakegu.
- Bilaketa obraren metadatuen arabera mugatu daiteke. Arlo, azpiarlo, erregistro, urte edo argitaratzaile jakin baten emaitzak soilik erakustea aukeratu dezakegu.
- Emaitzen atalean, testuinguruak (KWIC edo konkordantzia eran) eta kopuruak (maiztasun-taulak eta grafikoak) konbina ditzakegu. Kopuruen kasuan, zer datu bistaratu nahi dugun zehaztu dezakegu (adibidez, lemen maiztasun-taula edo arloen edo urteen araberako banaketa-grafikoa). Gainera, emaitzak CSV formatuan esportatzeko aukera ere badago.
- Aurreko hiru parametro mota horiek erabiliz, honako bilaketa konplexu hau egin genezake: *haize* lemaren ondorengo izenondoen agerpenak bistaratzea, *Zientzia* arloaren barnean, eta emaitzetan izenondoen maiztasun-taula erakutsi.



2. irudia. Bilaketa aurreratuaren adibide bat: *haize* izenaren ondorengo izenondoen agerpenak *Zientzia* arloan, dokumentuka ordenatuta, eta izenondoen lemen maiztasun-taula.

Erabilera-kasu bat: Hiztegi Batua

Proiektuaren berehalako helburua Hiztegi Batuko Lantaldearentzat aztergaia prestatzea eta egungo erabileren hutsunea betetzea da, bai formen aldetik, bai adieren aldetik. Lehen urrats honetan, formak landu dira.

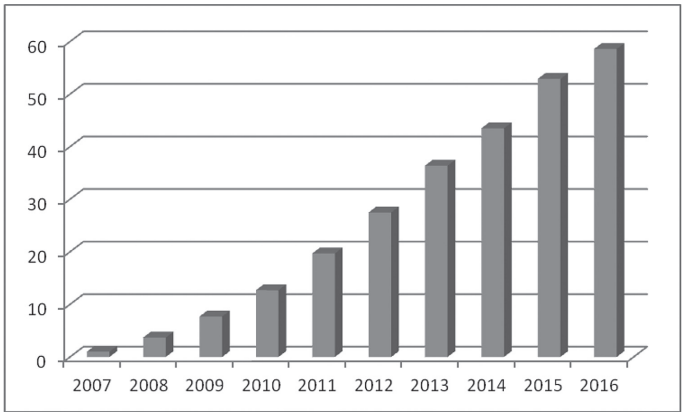
Lehen saio batean, Euskaltzaindiak dagoeneko aztertuak (eta isilduak edo ondoko urratsetarako utziak) zituen formak corpusaren emaitzarekin erkatu ziren, eta haietako batzuk, gaur oso erabiliak direnak, berreskuratu.

Ondoren, Euskaltzaindiak jaso ez baina corpusean maiztasun handi samarrez azaltzen ziren formak aztertu ditu lantaldeak, zenbait urratsetan. Tradizioa alde batera utzi gabe, egungo erabilerek ere lekua behar dute Euskaltzaindiaren Hiztegian.

Corpusaren tamaina eta bisitak

Corpusa, urtez urte

Grafiko honetan ikus daiteke LB corpusaren bilakaera urtez urte, hitz kopuruari dagokionez:



3. irudia. LB corpusaren bilakaera urtez urte (milioi hitzetan).

2016. urtearen bukaeran, beraz, 58.576.635 hitz zeuden guztira corpusean, kontsultatzeko moduan.

Bisita kopuruak

Sarreran esan den bezala, LB corpusaren orria Euskaltzaindiaren webgunean dago kontsultagai (Euskaltzaindia, 2008). 2017ko maiatzeko lehen hiru asteetan izandako bisita kopurutik estrapolatuz, oker handirik egiteko beldurrik gabe esan daiteke hilean 3.000 bat bisitarik (IP desberdinak) jotzen dutela LBren orrira kontsultaren bat egitera; bisitari horietako bakoitzak

1,41 bisita egiten ditu batez beste, eta bisita bakoitzeko 3,82 orri kontsultatzen ditu. Bisita kopuruak astegunetan jotzen du goia, eta, asteburuetan, jaitsi egiten da. Egunean zehar, berriz, bisita gehienak 09:00etatik 20:00etara bitarteko tartean egiten dira.

Etorkizunerako asmoak

LB proiektuaren motibazioak azaltzean, Euskaltzaindiaren ametsa aipatu da: erreferentzia-corpus handi, orekatu, lematizatu, etiketatu eta linguistikoki anotatu bat izatea. Helburu horretara hurbildu ahala, euskararen erabileraren argazki gero eta zehatzagoa izango dugu. Gaur egun, hiztegigintzarako erabiltzen dugu Lexikoaren Behatokia, baina erreferentzia-corpusek beste hainbat erabilera izaten dituzte: hizkuntzalaritzaren beste arlo batzuetako ikerketan, hizkuntzaren irakaskuntzan, hizkuntzaren soziologian, literaturaren ikerketan eta hizkuntzaren prozesamendu automatikoan. Espero dugu Lexikoaren Behatokiaren corpusaren erabilera ere hedatuko dela arlo horietara.

Erreferentziak

Artola, X.; A. Díaz de Ilarraza; A. Soroa & A.Sologaistoa (2009). «Dealing with complex linguistic annotations within a language processing framework». *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5): 904.-915. or.

Ezeiza, N.; I. Aduriz; I. Alegria; J.M. Arriola & R. Urizar (1998). «Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages». COLING-ACL'98, Montreal (Kanada).

Euskaltzaindia (2008). Lexikoaren Behatokia. http://lexikoarenbehatokia.euskaltzaindia.eus

——— (2016). *Hiztegi Batua* (orain, *Euskaltzaindiaren Hiztegia*). http://www.euskaltzaindia.eus/hiztegibatua

——— (2017). *Orotariko Euskal Hiztegia.* http://www.euskaltzaindia.eus/oeh

Hatcher, E. & O. Gospodnetic (2004). *Lucene in Action*. Co. Greenwich, CT, AEB: Manning Publications. https://lucene.apache.org/

McIlwaine, I. C. (2010). «Universal Decimal Classification (UDC)». In: *Encyclopedia of Library and Information Sciences*, 3. arg. New York: Taylor & Francis. Vol. 1:1, 5432.-5439. or.

TEI (2002). *Text Encoding Initiative* (*TEI P4*)*: Language Corpora*. http://www.tei-c.org/Vault/P4/doc/html/CC.html

UZEI (2002). *XX. mendeko Euskararen Corpus Estatistikoa.* http://xxmendea.euskaltzaindia.net/Corpus/aurkezpena.html

Oharra: Aipatutako URL guztiak 2017-05-31n kontsultatu dira.

Resumen

La Academia de la Lengua Vasca, Euskaltzaindia, cuenta entre sus quehaceres investigar sobre la lengua y dictar normas sobre su uso. Por otra parte, es indudable que hoy en día los corpus son imprescindibles para monitorizar el uso real de una lengua.

El Observatorio del Léxico es un proyecto iniciado por Euskaltzaindia en 2007, en respuesta a una propuesta hecha por el grupo de trabajo de su diccionario unificado (*Hiztegi Batua*). El resultado del proyecto es el corpus del mismo nombre, que puede consultarse en su página web.

El proyecto continúa en vigor, y durante estos diez años se ha constituido ya un corpus textual de casi 60 millones de palabras. El corpus está procesado automáticamente y anotado lingüísticamente, y ofrece al usuario las funcionalidades de consulta habituales.

En el artículo se mencionan las razones que motivaron el proyecto, y se exponen sus objetivos y las características principales del corpus. Asimismo se detalla el procedimiento que se sigue en su formación: la adquisición de textos y su catalogación, los pormenores de su integración en el corpus, y el tratamiento lingüístico que se realiza sobre ellos. Finalmente se explica el uso que Euskaltzaindia hace del corpus y con qué objetivo, y se detallan los planes con vistas al futuro.

Résumé

Parmi les tâches d'Euskaltzaindia, Académie de la Langue Basque, figurent la recherche sur la langue et la promulgation des normes régulant son usage. Par ailleurs, aujourd'hui l'on ne met plus en doute le fait que les corpus sont indispensables pour contrôler l'utilisation réelle d'une langue.

L'Observatoire du Lexique est un projet créé par Euskaltzaindia en 2007, en réponse à une proposition faite par son groupe de travail sur le dictionnaire unifié (*Hiztegi Batua*). Le résultat de ce projet est le corpus du même nom, qui peut être consulté sur Internet.

Ce travail est encore en vigueur, et pendant ces dix dernières années l'on a constitué un corpus textuel de près de 60 millions de mots. Le corpus est traité automatiquement et annoté linguistiquement, de plus, il offre à l'usager les fonctionnalités habituelles de ce genre d'outil.

Dans cet article sont exposés les raisons pour lesquelles ce projet a été conçu, les objectifs à atteindre et les principales caractéristiques du corpus. Nous sommes également informés sur la procédure suivie pour l'élaboration dudit corpus : l'acquisition des textes et leur classement, les détails de leur intégration dans le corpus et le traitement linguistique auquel ils ont été soumis. Pour finir, l'on explique l'utilisation qu'Euskaltazindia fait de ce corpus et dans quel but, puis l'on mentionne les plans prévus pour l'avenir.

Abstract

Among the tasks of the Royal Academy of the Basque Language are investigating the language and dictating norms for its use. Furthermore, there is no doubt that corpora are indispensable today to monitor the real use of a language.

The Observatory of the Lexicon project was initiated by the Academy in 2007, in response to a proposal from its workgroup on the unified dictionary (*Hiztegi Batua*). The result is the corpus of the same name, which can be consulted on the Web.

The project is an ongoing work and, in the ten years of its existence, a text corpus of almost 60 million words has been compiled. The corpus is processed automatically and annotated linguistically, and offers the user all the usual functionalities of this kind of tool.

In this article, we present the reasons that motivated the project, explaining its main goals and the characteristics of the corpus. Likewise, we detail the procedures carried out to create the corpus: the acquisition of the texts and their cataloging, how they are integrated into the corpus, and the features of the linguistic treatment to which they are subjected. Finally, we explain how the Academy uses the corpus and what for, and discuss our plans for the future.