

6.- PPT-A PRESTATZEKO BETE BEHARREKO ERANSKINA

1. KONTRATUAREN HELBURUA

Segidan aipatzen diren lanabesen mantentze, elikatze eta garapena, kasuan kasuko, zehazten diren xehetasunen arabera: Corpusen kontsulta-tresna ezartzea; *Egungo Euskararen Hiztegia (EEH)*; *Egungo Euskararen Corpusa (EEC)*; *Ereduzko Prosa Dinamikoa (EPD)*, eta *Hizkuntzen Arteko Corpusa (HAC)*.

1) *Egungo Euskararen Hiztegia (EEH)*.

Hiztegi honek 59.741 sarrera biltzen ditu dagoeneko, 218.520 azpisarrera, 450.855 adibide aukeratu eta 484.223 hitz multzo. Hiztegia garatzen eta elikatzen segitu behar da. Lehenik, E letra bukatu behar da, aurretik orain arte Elk argitaratu duen lanak erakusten dituen irizpideak eta baliabideak (EPG corpusean ageri den materiala) segituz. Eta A letrarekin hasi gero. Zehazki, «E» letraren zati bat («errazio» sarreratik aurrera) eta «A» letra osorik gelditzen dira egiteko. «E» letraren osaketa eta «A» letraren erdiaren hiztegi-sarrera sortzea obra honen parte izango da. Ezin du aurretik egindakoarekin homogeneitatea galdu.

2) Corpusen kontsulta-tresnaren ezartzea

Hautagaiak corpusak kontsultatzeko tresna bat garatu behar du, berea izango dena, eta hautagaiak datu-baseekin elikatuko duena. Tresna horren erabilera EHUn esku jarriko du eta EHUK mugarik erabili ahal izango du. Tresna horrek gutxienez ezaugarri hauek bete beharko ditu:

3) Egun Euskara Institutuak dituen corpusen mantentzea eta elikatzea

Euskara Institutuak, bere webgunean zenbait corpusen kontsulta egiteko modua eskaintzen dio erabiltzaileari. Proiektu honek, tresna horiek abian mantendu eta eguneratuta mantentzeko lanak ere bere barruan hartzen ditu.

Egiten den lanak egungo interfaze eta aurkezpenaren itxura-aldaketa badakar, orain arte Euskara Institutuak argitara eman dituen corpusak bere osotasunean egokitu beharko dira, aurreko lanak eta orain egin behar denak itxura bera izan dezan.

Mantentze lana, corpus kontsulta-tresna hauetan egin beharko da:

1. *Egungo Testuen Corpusaren (ETC)*

2. *Ereduzko Prosa Dinamikoa (EPD)*

3. *Hizkuntzen arteko Corpusa (HAC)*

2. BETEBEHARREKOEN ZEHAZTASUNA

1) *Egungo Euskararen Hiztegia (EEH)*.

Zehazki, hauek izan dira entregatu beharreko hiztegi puskak, eta epeak:

- 2019ko maiatza: "errazionamendu-espageti"
- 2019ko iraila: "espainiar-ezusteko"
- 2020ko ekaina: "a-aire"
- 2020ko abendua: "aireberritu-antsia"

Egungo Euskararen Hiztegia, corpusean oinarritutako hiztegia da. Zehazki, *Ereduzko Prosa Gaur* corpusean oinarritutakoa. Corpus horretan dagoena, eta hori bakarrik, izan daiteke hiztegi sarrera izateko hautagaia.

Sarrerak, orain arte landutako sarreren egitura bera mantendu beharko dute. Hiztegia, identifikatu beharreko elementuak hauek izango dira, gutxienez:

- Hiztegi-sarrera
- Definizioa
- Adierak
- Erabilera oharrak (beharrezkoa irizten zaion kasuetan, ETCko datuak ere gehituz)
- Adibideak. Adibideen kasuan, zein testutan agertu den. (Adibideen kopurua, agerpen kopuruaren arabera izango da).
- Hitz-multzoen maiztasuna (beharrezkoa denean).

Egiten den lanak egungo interfaze eta aurkezpenaren itxura-aldaketa badakar, orain arte Euskara Institutuak argitara eman dituen EEHaren gainerako zatiak ere egokitu beharko dira, aurreko lanak eta orain egin behar denak itxura bera izan dezan.

2) Corpusen kontsulta-tresnaren ezartzea

a) Kontsulta web bidez egingo da. Edozein web nabigatzailearen bidez egin ahal izango da tresna horrek biltzen eta kudeatzen dituen datuen kontsulta. Ez da, inongo kasuan, kontsulta egin ahal izateko erabiltzailearen ordenagailuan softwarerik instalatu beharko.

b) Kontsulta egiteko orduan, gutxienez, datuen bilaketa irizpide hauen arabera egin ahal izango da:

- hitza/lema
- noranahiko karakterea (") erabili ahal izango da hitzetan
- kategoria gramatikala
- bilaketa terminoen arteko distantzia
- testu-mota

c) Aipatu ezaugarrien arteko konbinazio desberdinak ere egin ahal izango dira bilaketa egiteko orduan.

d) Bilaketaren emaitzetan, gutxienez honako elementuek agertu beharko dute:

- Hitzen maiztasuna, modu grafikoan zein tabularrean.
- Emaitzen testuingurua.
- Sailkapen desberdinak, gutxienez testu-mota eta testuaren jatorria.

e) Ustiaketa tresnaz gain, corpusa osatzeko prozedura ere definituko da. Corpusaren elikatzea jarraia da. Testu berriak lortu ahala, corpusaren kontsulta-tresnan integratu ahal izateko prozedura eta programak eskainiko dira, elikatze lan hauek proiektu honen parte direlarik.

f) Kargatzen diren testu eta dokumentuek hutsak baldi badituzte, ahalik eta lasterren egokituko dira. Halaber, testuak prozesatzean hutsak aurkituz gero.

g) Interfazearen gaurkotze guztiak etengabe, sortu ahala, egin beharko dira.

h) Ezusteko arazoak sortuz gero, Elri jakinaraziko zaio, eta beharrezkoak diren konponketak diseinatu eta abian ezarriko dira.

3) Egun Euskara Institutuak dituen corpusen mantentzea eta elikatzea

1. Egungo Testuen Corpusaren (ETC)

Beranduenen 2020ko irailean, 2016tik 2019ra bitarteko testuak emendatu beharko dira, eta testu berrien kantitatea gutxienez 50 milioi testu hitzekoa izango da. Gehituko diren testuek, iturburu hauetatik hartuak izan beharko dute gutxienez:

- Berria egunkaria (osorik).
- Argia astekaria (osorik).
- Goienharia astekaria (osorik).
- Herria astekaria (osorik).
- Liburuak (gutxienez urteko 30 liburu) —berorien % 20 gutxienez itzulpenak izango dira—.
- Euskarazko wikipedia

2. *Ereduzko Prosa Dinamikoa (EPD)*

Urte urte eguneratuko da corpus hori, aurreko urteko testuak gehituta. Honela osatuko da urteroko corpus handitzea:

- Denera, urteko 3 milioi testu hitz:
 - Berria egunkaria (1,3 milioi).
 - Herria astekaria (0,2 milioi).
 - Liburuak (1,5 milioi) —berorien % 20 gutxienez itzulpenak izango dira—.

Handitze hori 2019ko abenduan (2018ko testuei dagokienez), beranduenik, eta 2020ko azaroan (2019ko testuei dagokienez), beranduenik, egingo da.

3. *Hizkuntzen arteko Corpora (HAC)*

Beranduenen 2020ko azaroan, gutxienez 25 liburu gehitu beharko zaizkio corpus horri. Hona bete beharreko baldintzak:

- Liburu guztiek lau hizkuntzatarako bertsio bana izango dute: euskara, gaztelania, ingelesa eta frantsesa.
- Lau hizkuntzatarako testuek esaldika lerrokatuta egongo dira (lerrokatze akatsak ezingo du % 10 baino handiagoa izan)

Corpus guztietan

Zerbitzuan aurreikusitako ohiko eragiketak:

- Tresnaren instalazioa eta konfigurazioa (EHUko zerbitzarietan).
- Tresna EHUren zerbitzu teknikoek egiten dituzten zerbitzari eguneraketei egokitzea.
- Testu berrien karga:
 - Testuen bilketarako koordinazio lanak.
 - Testuen etiketatzea.
 - Testuan lematizazioa.
 - Testuen karga produkzioan dagoen sisteman.
- Agertzen diren akatsen zuzenketa.
- Zerbitzu-etenen aurrean zerbitzu tekniko eskaini.

Datu basean behar diren gaurkotze guztiak egingo dira.

Lemak desanbiguatu behar izanez gero konponbide egokiak aurkituko dira.

Kontsulten abiadura dela eta, sor daitezkeen arazoen konponketa aurkituko da.

Tresneria erabili ahala sor daitezkeen arazoak aztertu eta konponduko dira.

Beharrezko suertatuko diren aldaketak egingo dira. Betiere, aldaketa hauek garrantzi handia badute, Elrekin batera hartuko dira erabakiak. Azken hitza Elk du.

Eguna: 2018.12.10

Sin.:
Zentroa/Saila/Zerbitzua:


EUSKARA INSTITUTUA
(PELLO SALABURU)