# XX. mendeko euskararen corpus estatistikoa

Miriam Urkia

UZEI

## 1. SARRERA

*XX. mendeko euskararen corpus estatistikoa* garai eta egoera zehatz batean sortua da, behar berezi batzuei erantzuteko pentsatua. Euskal corpusei buruz aritzean, Euskaltzaindia aipatzea ezinbesteko da, haren ekimenari esker sortu baitira egun eskura ditugun bi corpusak: *Orotariko Euskal Hiztegia*, tradizioa biltzen duena, batetik, eta *XX. mendeko euskararen corpus estatistikoa*, bestetik.

Txosten hau kokatzeko, zilegi bekigu corpus hauen historia labur bat egitea, aurrekariak aipatzea, gerora egindakoa argitzeko lagungarri izan daiteke-eta.

1950. hamarkadan sortu zen Euskaltzaindian hiztegiaren beharra, nahiz hiru hamarkada behar izan ziren ideia hori gorpuzteko; alegia, Koldo Mitxelena *Orotariko Euskal Hiztegia* eratzeko arduradun izendatu zutenean. Hiztegiaren oinarri izango zen corpus baten beharra ikusi zuten orduan eta, horretarako, Mitxelenaren taldeak 310 obra aukeratu eta bildu zituen, 1545. urtetik XX. mendearen erdiraino jasoko zituena, guztira 5.800.000 testu-hitz bilduz. Halere, erabilera mugatua izan zuen lehen urteetan behintzat, hiztegirako besterik ez baitzen erabili corpusa, nahiz gerora Gramatika Batzordeak eta Hiztegi Batuko Lantaldeak, besteak beste, baliatu izan duten.

Bestalde, tradizioaz gain, euskara modernoa biltzeko beharra ere ikusi zuen Euskaltzaindiak, garai hartarako irakaskuntza, komunikabideak eta administrazioa euskara baliatzen ari baitziren eguneroko bizitzan. Horrela sortu zen, hain zuzen, *EEBS* (Egungo Euskararen Bilketa-lan Sistematikoa) izenez ezagutuko zen XX. mendeko euskara biltzen zuen corpusa, estatistikoa. Lan horren ardura UZEIri eman zion Euskaltzaindiak, EEBS batzordearen irizpideei jarraituz betiere.

Hasierako plan batean, 1991. urterako *Orotariko Euskal Hiztegia* eta *EEBS* bukatzea aurreikusten zuen Euskaltzaindiak. Bi lanok abiapuntu hartuta, *Hiztegi Hiritar Arauemailea*[1] osatzeari ekingo zion, alegia, tradizioa eta egungo euskara biltzen zituzten corpusetan oinarrituko zen nagusiki, LEF Batzordeak finkatutako irizpide lexikologikoak baliatuz eta bestelako materialak ere kontuan hartuz (terminologia, besteak beste). Baina proiektuok ez ziren bukatu eta *Hiztegi Hiritar Arauemailea*ri heldu aurretik, tarteko —eta presazko— beste eginkizun bati heltzea erabaki zuen Euskaltzaindiak: 20.000-30.000 forma bilduko zituen hiztegi ortografiko bat

[1] Ikus *Euskera*, 1986,XXXI, 130. or.

osatzea, hau da, *Hiztegi Batua*. Eta, hauxe izango zen, hain zuzen, egunen batean bideratu beharreko *Hiztegi Hiritar Arauemailea*ren abiapuntu.

Testuinguru honetan kokatu behar da, beraz, ondoko orriotan aurkeztuko dugun *XX. mendeko euskararen corpus estatistikoa*, lehen *EEBS* izenez ezagutzen genuena.

## 2. XX. MENDEKO EUSKARAREN CORPUS ESTATISTIKOA

***XX. mendeko euskal lexikoaren erabileraren berri ematea*** da, beraz, azken mendeko euskararen corpus estatistikoaren helburu nagusia, hasierako definizioaren arabera. Helburu argi bezain mugatu honek eragina izan du corpusaren osaeran eta ustiapenean, gero ikusi ahal izango dugunez.

Corpusak zazpi ezaugarri nagusi ditu, ondoko ataletan azaltzen saiatuko garenak:

a) *Sinkronikoa*: XX. mendea hartzen du, 1900-1999 urteak, hain zuzen.

b) *Elebakarra*: euskara.

c) *Idatzia*: ahozkoa transkribatu eta argitaratu den neurrian jaso da.

d) *Lagina*: estatistikoa, XX. mendeko euskal argitalpenen inbentario sailkatuan oinarritua eta proportzionalki jasoa.

e) *Sailkatua*: epea, euskalkia, testu-mota eta obraren tamainaren arabera.

f) *Kodetua*: SGML formatu estandarra baliatuz.

g) *Lematizatua*: testu-hitz bakoitzari lema estandar bat esleituz.

Ondoko orrietan corpusaren osaera-prozesua azaltzen saiatuko gara, corpusaren abiapuntu izan den unibertsoa aipatuz, hau da, XX. mendeko euskal argitalpenen inbentario sailkatuaren berri emanez, eta, corpusa mugatu ondoren, osaera-faseak banaka azalduz. Horrela, bukaeran emaitzak ikusi ahal izango ditugu eta etorkizunean beharko genukeen corpuserako zertzelada batzuk ematen saiatuko gara, ondorioetan.

### 2.1. Abiapuntua: XX. mendeko euskal argitalpenen inbentario sailkatua

Corpus bat osatzen hasi aurretik, unibertsoa ezagutu behar da. Alegia, gurera etorriz, zer argitaratu[2] da XX. mendean euskaraz? Horren berri izateko, guztiaren inbentarioa egin behar da, inbentario sailkatua gainera. Horixe izango da, hain zuzen, corpusaren oinarri.

Bi sail nagusitan banatu zen unibertsoa, gerora bi azpicorpusen oinarri izango zena: liburuak eta aldizkari "nagusietako"[3] artikuluak, batetik; eta egunkariak —aldizkari ofizialak barne— eta aldizkariak zein komunikabideetako lanak, oro har, bestetik.

[2] "Argitaratu" diogu, corpus idatzia izango zela erabakia baitzegoen hasieratik: idatzi argitaratua. Alegia, corpus honetan ez da ahozkoa landu (transkribatu eta argitaratu den neurrian bai, ez bestela), ez eta material elektronikoa. Azken hau, gainera, ez zegoen eskura 1986an, corpusa abian jarri zenean.

[3] Aldizkari "nagusi" diogunean nolabaiteko pisua dutenak adierazi nahi dira. *Euskera* edo *Jakin* modukoak, artikulu landu eta mardulak dituztenak hartu ziren kontuan. Herri-aldizkariak, esaterako, artikulu laburragoak, askotan sinatu gabeak eta nolabait arinagoak (eta maiztasun handiagokoak) sailkatu gabeen multzoan bildu ziren, aurrekoekiko desberdintasun bat dagoela aitortuz.

### 2.1.1. Sailkapen-irizpideak

Lau irizpide nagusi aukeratu ziren, nahiz, ondoren ikusiko dugunez, ez ziren goian aipatutako bi sail nagusietan modu berean aplikatu. Baina, zabal hartuta, hala egin zen sailkapena:

a) *Garaia*: lau epe nagusitan banatu zen guztia.

1. *1900-1939*: mende-hasieratik gerrak artekoa.
2. *1940-1968*: gerraostean abiatu eta euskara batuaren sorrera artekoa.
3. *1969-1990*: euskara batuak ekarritako aldaketekin hasi eta Euskaltzaindiaren gomendioak eta arauak artekoa (eta Ibon Sarasolaren *Hauta-Lanerako Euskal Hiztegia* argitaratu artekoa).
4. *1991-1999*: araugintza berriaren ondokoa.

b) *Euskalkia*: sailkapen zabala egin zen.

*1.* *Bizkaiera*
*2.* *Gipuzkera*
*3.* *Zuberera*
*4.* *Lapurtera/Nafarrera*
*5.* *Euskara batua*
6. "*Nahasiak*" (sailkatu gabeak)[4]

c) *Testu-mota*:

*1.* *Saio-artikuluak*
*2.* *Administrazio-idazkiak*
*3.* *Ikasliburuak*
*4.* *Saio-liburuak*
*5.* *Prosa literarioa*
*6.* *Poesia*
*7.* *Antzerkia*
*8.* *Bertsoak*
*9.* *Ikerketa-lanak*
*10.* *Haur- eta gazte-literatura*
11. *Ahozkoak* (ahozko jardunen transkripzioak)
*12.* *Liturgia*
*13.* *Egunkariak*
*14.* *Astekariak/hamaboskariak*
15. *Hilabetekariak* (eta maiztasun urriagokoak)

d) *Obraren tamaina*:

1. 1-5 orrialde (1-1000 hitz)
2. 6-20 orrialde (1000-5000 hitz)
3. 21-50 orrialde (5000-10000 hitz)
4. 51-250 orrialde (10000-50000 hitz)
5. 251 orrialde baino gehiago (50000 hitzetik gora)

Eta, sailkatu gabeen atalean, argitalpen bakoitzaren testu-masaren totala bakarrik hartu zen kontuan, artikuluka banatzea ezinezkoa baitzen.

[4] Euskalkiaren arabera sailkatu gabe utziak: a) liburuetan (bertsoak, elkarrizketak), eta b) aldizkarietan (masa osoa jasota, artikuluka). Izan ere, askotan hainbat euskalki tartekatzen dira halako obratan.

### 2.1.2. Unibertsoa: inbentarioaren emaitzak

Irizpideok aplikatuz, hainbat liburutegi[5] haztatu genituen XX. mendean euskaraz argitaratutako guztia inbentariatu eta sailkatzeko.

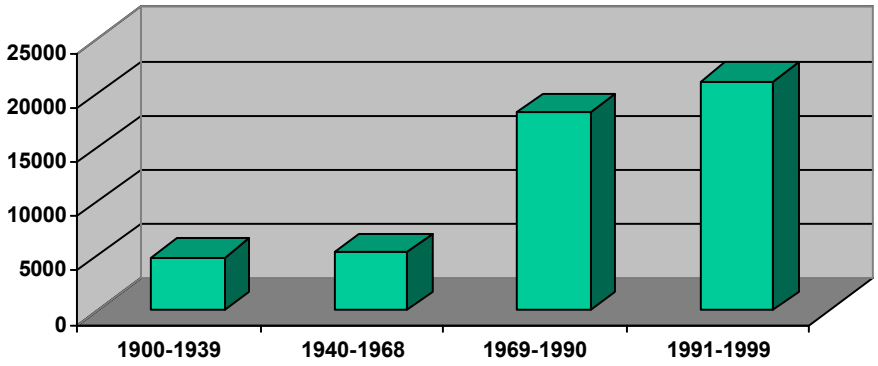
Unibertsoa bi sail nagusitan banatu genuenez —corpusa ere bi azpicorpus nagusik osatzen baitute—, inbentarioaren emaitzak sailka eskuratu genituen. Ikus ditzagun:

#### *2.1.2.1. Sailkatuak (liburuak eta aldizkari "nagusiak")*

Goian zehaztutako irizpideen arabera osatu zen inbentariatze-lana[6] eta datuok eskuratu genituen:

a) *Garaia, epea*:

1. 1900-1939: 4.910 dokumentu
2. 1940-1968: 5.415 dokumentu
3. 1969-1990: 18.283 dokumentu
4. 1991-1999: 21.139 dokumentu




Taulak erakusten duenez, lehen bi epeetan dokumentu-kopurua antzekoa da, baina euskara batuaren sorreratik gure egunetara gora egin du etengabe. Are gehiago, mendearen azken 9 urteetan aurreko 22 urteetan baino gehiago argitaratu da euskaraz, nahiz azken 9 urteetan euskal produkzio idatzia egonkortu egin dela esan daitekeen.

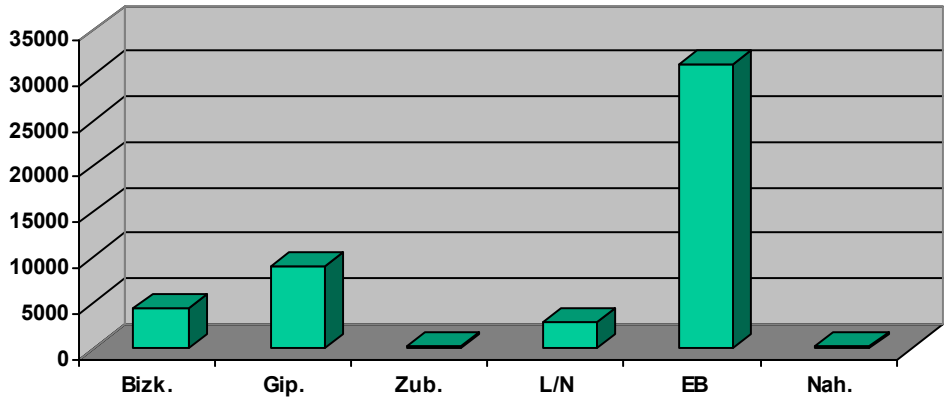
b) *Euskalkia*:

1. Bizkaiera: 4.506 dokumentu
2. Gipuzkera: 8.965 dokumentu
3. Zuberera: 203 dokumentu
4. Lapurtera/Nafarrera: 2.815 dokumentu
5. Euskara batua: 31.210 dokumentu
6. "Nahasiak": 177 dokumentu

---

[5] Bihoazkie gure eskerrik beroenak urtetan laguntza eskerga eman diguten guztiei: Euskaltzaindiko Azkue Liburutegia, Koldo Mitxelena Kulturunea, Donostiako Udal Liburutegia, Lazkaoko Beneditarrak, Donostiako Seminarioa, Zabaltzen, Bilintx Liburudenda, Jakin aldizkaria, Euskaldunon Egunkaria, Mendu, hainbat herri eta erakundetatik helarazi dizkiguten aldizkarietako arduradunak, eta beste hainbat lagun, beti laguntzeko prest izan ditugunak. Eskerrik asko guztiei.

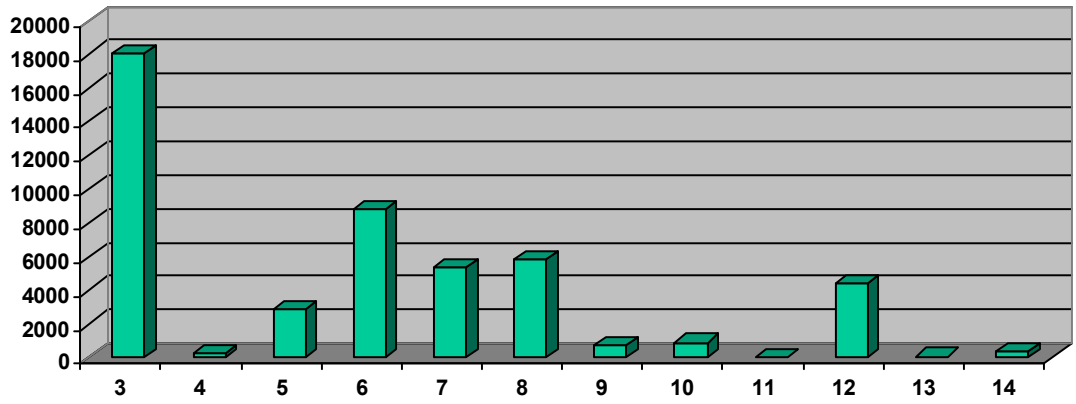
[6] Inbentarioan lehen edizioak besterik ez ziren kontuan hartu. Inbentarioan, obra batek beste autore baten hitzaurrea edo halakorik balu, edo autore berak obra berean testu-mota zein euskalki desberdinak erabili baditu, hori beste obra bat bezala jasotzen da, sailkapena aldatu egiten baita. Argi dezagun, bestalde, "obra" diogunean, fitxa edo erreferentzia esan nahi dugula: dokumentua, alegia.




Euskalkietan idatzitako lanak gutxi dira euskara batuarekin konparatuz, baina horrek azken 30 urteetan euskaraz idatzi dutenek sistema estandarrera jo dutela adierazten du: alegia, normalizaziorako ahalegina handia izan da. Mendearen azken urteetan bizkaiera gipuzkera baino gehiago erabiltzen da, baina datuek ez dute hori erakusten. Hain zuzen, urtetan gipuzkera ia batu gisa erabili izan zen, eta horixe da desberdintasunaren arrazoi nagusia. Bestalde, zubereraz oso gutxi idatzi da besteen aldean, 203 dokumentu horiek erakusten dutenez.

c) *Testu-mota*:

| | | |
|---|---|---|
| 3. Saio-artikuluak: | 18.083 | dokumentu |
| 4. Administrazio-idazkiak: | 280 | |
| 5. Ikasliburuak: | 2.913 | |
| 6. Saio-liburuak: | 8.851 | |
| 7. Prosa literarioa: | 5.355 | |
| 8. Poesia: | 5.822 | |
| 9. Antzerkia: | 711 | |
| 10. Bertsoak: | 919 | |
| 11. Ikerketa-lanak: | 81 | |
| 12. Haur- eta gazte-literatura: | 4.438 | |
| 13. Ahozko transkripzioak: | 104 | |
| 14. Liturgia: | 387 | |




Testu-motaren araberako sailkapenean saio-aldizkariak dira alde handiz nagusitzen direnak, hasieran aipatu ditugun aldizkari nagusietako artikuluak banaka jaso dira-eta. Administrazio-idazkiak gutxi dira, baina kontuan izan behar da sailkatu gabeen multzoan inbentariatu direla egunkari ofizialak eta, horiek, administrazioko euskararen erakusle garrantzitsuak dira.

d) *Tamaina*[7]:

| | | | |
|---|---|---|---|
| 1. 1-5 orrialde: | 25.418 | dokum. | (12.709.000 testu-hitz) |
| 2. 6-20 orrialde: | 12.875 | | (38.625.000 testu-hitz) |
| 3. 21-50 orrialde: | 4.754 | | (35.655.000 testu-hitz) |
| 4. 51-250 orrialde: | 4.041 | | (121.230.000 testu-hitz) |
| 5. 251 orrialde baino gehiago: | 766 | | (45.960.000 testu-hitz) |

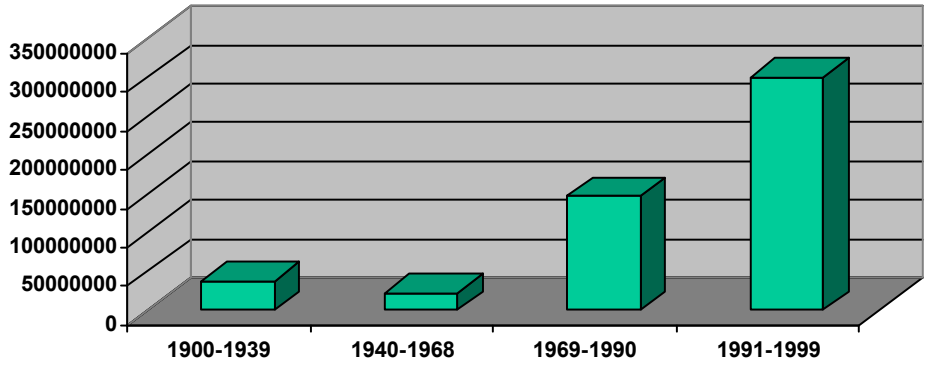
Datuokin, XX. mendeko euskal produkzio sailkatuaren kopurua **254.179.000** testu-hitzekoa zela jakin ahal izan genuen. Bestela esanda, XX. mendean 254.179.000 hitz argitaratu ziren euskal liburu eta aldizkari nagusien sailean.

### *2.1.2.2. Sailkatu gabeak (aldizkariak eta kazetaritzako lanak, oro har)*

Sailkatu gabeen artean ez zen euskalki-banaketarik egin, argitalpenak bere osotasunean hartu baitziren kontuan. Izan ere, testu-masa handiegia da artikuluka sailkatzeko, askotan sinatu gabeak dira eta, oro har, liburuak baino presazkoagoak —eta, ondorioz, gutxiago landuak— dira. Beraz, osotasunean jaso zen eta "nahasiak" multzoan bildu. Horrez gain, testu-masa handiak izan ohi direnez, argitalpen bakoitzaren hitz-kopuru totala zenbakitu zen.

a) *Garaia, epea*:

| | |
|---|---|
| 1. 1900-1939: | 37.583.775 hitz |
| 2. 1940-1968: | 22.666.000 hitz |
| 3. 1969-1990: | 147.812.782 hitz |
| 4. 1991-1999: | 302.264.239 hitz |

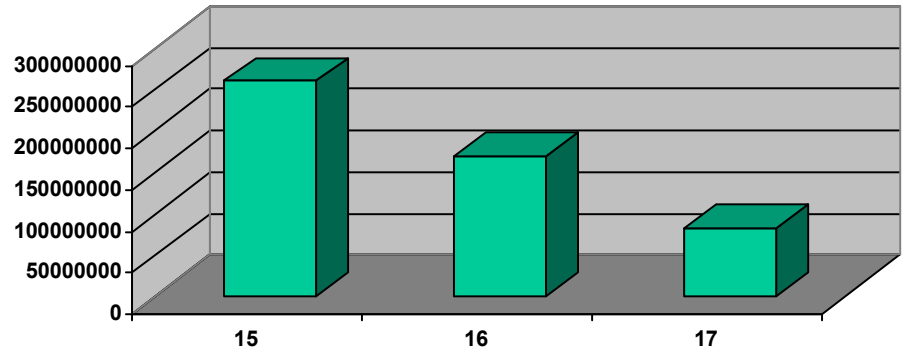



Mende-hasierako euskal produkzioa gerraostekoa baino handiagoa da, beheraxeago azalduko dugunez. Euskara batuaren sorreraren ondokoak gora egiten du nabari, baina aipagarria da mendearen azken 9 urteetako igoera. Erantzun erraza du honek: *Euskaldunon Egunkaria*k 1990eko abenduan kaleratu zuen lehen alea eta, ordutik, egunero —astean seitan— testu-masa itzela eman dio euskarari.

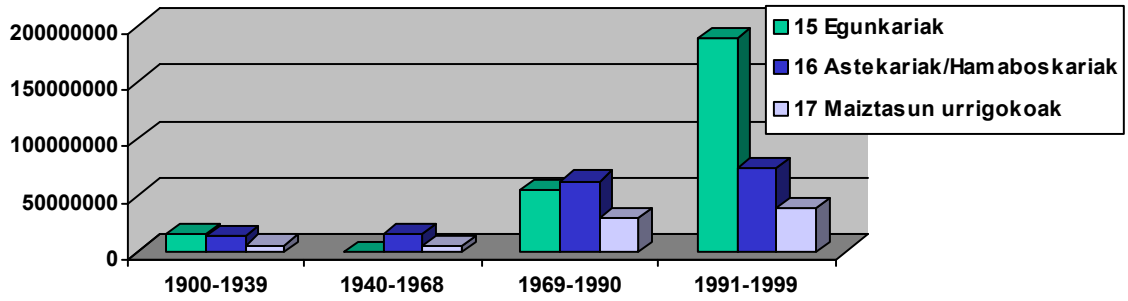
b) *Testu-mota*:

| | |
|---|---|
| 15. Egunkariak: | 259.840.586 hitz |
| 16. Astekariak eta hamaboskariak: | 168.773.954 hitz |
| 17. Hilabetekariak eta maiztasun urriagokoak: | 81.712.256 hitz |

[7] Formula bezala, "1" taldekoen media 500 hitzekoa da, "2"koena 3000, "3"koena 7500, "4"koena 30000 eta "5"ekoena 60000. Kopuru horietan oinarrituta atera da testu-masa osoa.




Argi dago, zenbateko maiztasunarekin argitaratu, halako testu-masa eskuratzen dela. Ikus dezagun garaia eta testu-mota konbinatuz egin daitekeen irakurketa:




Lehengo bi epeetako datuei begiratuz, mende-hasierako egunkariak (*Euzkadi'ko Agintaritzaren Egunerokoa*, *Euzkadi*, *El Día*) desagertu egin ziren gerraostean; 2. epe honetan, 1968 artekoa, *El Diario Vasco* besterik ez dugu (Basarrik idazten duen 'Mi atalaya montañera" saila batez ere). Baina ez da gauza bera gertatzen astekariekin eta bestelako argitalpenekin; izan ere, 1. epeko *Eskualduna*, *Gipuzkoarra*, *La Cruz*, *Jaungoiko-zale* eta *Armanak Uskara* (Ziberouko Egunaria) modukoak desagertzen badira ere, *Zeruko Argia*k, *Euzko Deya*k (orain Mexikon) eta beste batzuk jarraitu egiten dute, baina, horiez gain, *Herria* kaleratzen da Iparraldean, bai eta *Anaitasuna*, *Azkatasuna* eta *Eusko Gaztedi* (Caracas) ere, besteak beste. Alegia, mugaz bestalde edo atzerrian ikusten dute argia aldizkariek. 3. epean nabarmen egiten dute gora hiru sailek: egunkarietan euskara erabiltzen da, egunkari ofizialak kaleratzen dira euskaraz eta, azken urteetan batez ere, herri-aldizkariek hartzen dute indarra. Are nabarmenagoa da gorakada hori 4. epean.

Horiek horrela, sailkatu gabeen kopuru totala **510.326.796** testu-hitzekoa zela kalkulatu ahal izan genuen.

Eta, bi azpimultzoak batuz, hau da, sailkatuak eta sailkatu gabeak, **XX. mendeko euskal argitalpenen unibertso osoa 764.505.796 hitzekoa** zela jakin ahal izan genuen. Hori zen, beraz, abiapuntua. Datuokin has gintezkeen corpusa osatuko zuen lagin estatistikoa bideratzen.

## 2.2. Corpusaren osaera

### 2.2.1. Corpusaren tamaina zehaztea

Unibertsoa osatuta, zenbateko corpusa beharko litzateke XX. mendeko euskararen erakusgarri aski zabala izateko? Unibertsoa ezin da bere osotasunean jaso, hori argi dago, baina corpus erakusgarria eta egingarria behar genuen hasierako helburuari erantzuteko, alegia, mende osoko euskal lexikoaren ahalik eta irudirik zabalena islatzeko.

Horrela, garaiko beste lan batzuen emaitzak aztertu genituen, corpusaren tamaina eta emaitzen erakusgarritasuna erkatuz[8]. Haiek eskaintzen zituzten datuen argitan, 2.000.000 testu-hitzeko corpusa aski erakusgarri izan zitekeela kalkulatu zen, horrekin 40.000 lema desberdin —edo hiztegi-sarrera, modurik zabalenean ulertuta — lor zitezkeela aurreikusi baitzen. Nahikoa zen gure eginkizunerako, eta hala egin zen.

Gogora dezagun, halere, 1987an egin zirela lanok eta, beraz, goian aipatu dugun unibertsoa ez zegoela bere osotasunean bildua, 1900-1986 urteak hartzen zituena bakarrik. Labur esateko, ia gaur osatua dagoen unibertsoaren erdia besterik ez zen. Azken emaitzak, 1900-1999 osorik hartuta, gero ikusiko dugunez, nabarmen aldatu dira hasierako aurreikuspen horretatik.

Beraz, unibertsoko sailkapen-ataletako bakoitza jaso behar genuen, proportzionalki, bi milioi hitzeko lagin estatistikoa osatzeko. Horretarako formula estatistikoak baliatu behar genituen; eta horixe egin zen, hain zuzen, EUSTATen —zehatzago, Anjeles Iztueta andrearen— laguntzarekin. Parametro batzuk definitu eta atal guztiek izan zuten, zegokien proportzioan, bere lekua corpusean. Testu-mota bakoitzari, esaterako, kuota bat esleitu zitzaion: prosan (prosa arrunta, ikasliburuak, prosa literarioa eta antzerkia), 50 hitzeko bat jasoko zen; poesian eta bertsoan, 200 hitzeko bat; sailkatu gabeko kazetaritza-lanetan, 350 hitzeko bat. Arrazoia sinplea da: prosazko lanek hizkuntzaren erabilera arrunta islatzen dute, poesiak eta bertsoak askotan ez bezala, hizkuntzaren berariazko desbiderapenak egiten baitira[9]. Kazetaritza-lanetan, lehen esan dugunez, produkzioa handia da eta liburuetako prosa baino presazkoagoa —eta, ondorioz, ez hain landua— askotan.

Aukeratze-lana ausazkoa izan zen, autoreei eta obrei begiratu gabe: estatistikoki jasoak. Unibertsoan fitxa bibliografiko osoak jaso baziren ere, guk kodea besterik ez genuen erabili horiek aukeratzeko, adierazgarritasuna zozketaren esku utziaz. Adibide batekin esateko, mende-hasierako (1. epea) bizkaieran argitaratutako prosa literarioko obra txikiak (1-5 orrialde artekoak), unibertsoan, 162 dira. Horietako 4 pasako ziren corpusera, formulak aplikatuta. Edo, 3. epeko (1969-1990) saio-artikuluak, euskara batuan, 6-20 orrialde artekoak, unibertsoan 2.152 zirenak, 159 izango ziren corpusean lekua izango zutenak. Zenbat dokumentu ziren bagenekien, zeintzuk ez, hori zozketak emango zuen-eta.

Egokitutako obra horiek, gainera, ez ziren bere osotasunean jasoko, bakoitzetik orrialde gutxi batzuk baizik. Hartara, lexiko-aberastasuna berma zitekeela uste zen. Obrak bezala, orrialdeak ere ausaz jaso ziren, hala jokatuz: sailkatuetan, 400, 800 edo 1.600 hitz obra bakoitzeko, tamainaren arabera; sailkatu gabeetan, 2.000 edo 4.000 hitzeko multzoak (argitalpen bereko hainbat artikulutan banatuak).

---

[8] Lau hizkuntzatako lanak hartu genituen kontuan: *Frequency Dictionary of Spanish Words* (500.000 hitzeko corpus batetik abiatuz, 5.024 hitz desberdin lortu ziren); Ibon Sarasolaren *Gaurko euskara idatziaren maiztasun-hiztegia* (800.000 hitzeko corpusetik, 24.537); *Computational Analysis of Present-day American English* (1.014.232 hitzetik, 50.406); eta, azkenik, *Trésor de la Langue Française* (70.317.234 hitzetik, 71.415 desberdin).

[9] Prosaren barruan, literaturan ere egiten dira halakoak, egia da. Hizkera literarioa eta informatiboa ez dira maila berean jasoko, aurrerago ikusiko dugunez.

Beraz, 1900-1987 urteak hartuko zituen corpusaren lehen bertsioak 2.000.000 hitz bilduko zituen. Dena den, corpus ireki gisa planteatu zen eta, gerora, mendea bukatu arte, urtero eguneratuz joan da, betiere irizpide berak errespetatuz, oreka mantentzeko. Azken hamarkadan, halere, hain zuzen 4. epeari ekitean, "korrektibo" bat ezarri behar izan zen, testu-masak nabarmen egin baitzuen gora eta, ondorioz, corpusean desoreka handia sortu, mendearen lehen erdia ia desagerraraziz. Formulak berrikusi eta jaso beharrekoa erdira jaistea erabaki zen. Alegia, prosan, 100 hitzeko bat jasoko zen; poesian eta bertsoan, 400 hitzeko bat; eta sailkatu gabeko kazetaritza-lanetan, 700 hitzeko bat.

Horrela, mendearen azken bederatzi urteetako lagina biltzean, aurrekoen erdia besterik ez da jaso. Eta, hala ere, testu-masak gora egin du. Hasierako 2.000.000 hitz haiek bikoiztu egin dira, eta 40.000 lema desberdin beharrean, ehun milatik gora ditu corpusak.

Laburtzeko, esan dezagun 764.505.796 hitzek osatzen zuten unibertsotik 4.657.165 besterik ez direla corpuseratu: alegia, % 0.6 baino ez dugu landu. Corpusera pasa diren formen artean, 418.487 forma desberdin daude, eta 104.817 lema desberdin. Orotara, corpusa 6.352 dokumentuk osatzen dute, hala banatuta azpicorpusen arabera: 3.467 dokumentu sailkatuen multzoari dagozkionak, eta 2.885, berriz, sailkatu gabeen multzoari dagozkionak.

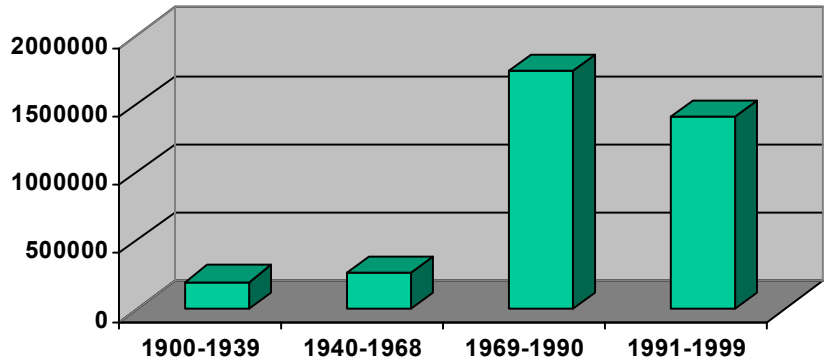
### 2.2.2. Corpusaren osaeraren emaitzak

Formula guztiak aplikatu ostean, eta goraxeago aipatu ditugun kopuruen argitan, ikus dezagun zer atera zen, alegia, corpus izatera zer pasa zen.

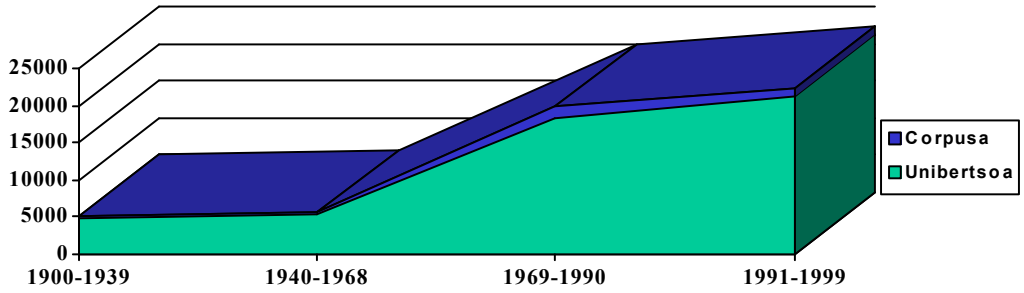
#### 2.2.2.1. *Sailkatuen azpicorpusa:*

*a) Epeka:*

| *Epea* | *Hitzak (dokumentuak)* | *%* |
|---|---|---|
| 1900-1939 | 198.057 (277 dok.) | % 5.5 |
| 1940-1968 | 263.932 (359 dok.) | % 7.3 |
| 1969-1990 | 1.751.703 (1702 dok.) | % 48.3 |
| 1991-1999 | 1.408.340 (1129 dok.) | % 38.9 |
| **Guztira** | **3.622.032 (3467 dok.)** | **% 100** |

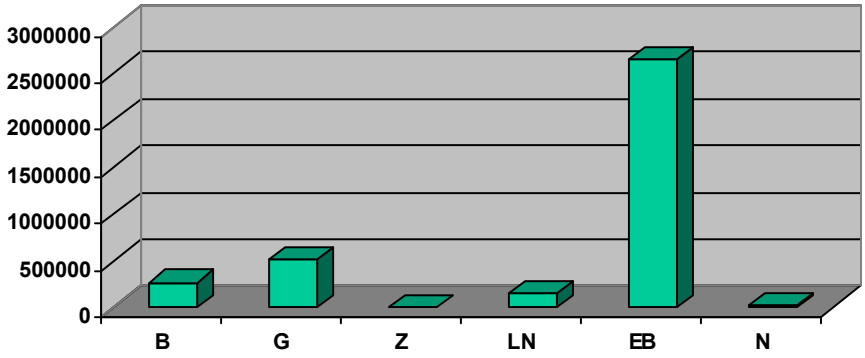



Unibertsoa osatzean ikusi ditugun datuekin erkatuz, benetan proportzionalki jaso dela erakusteko, hona irudi honetan unibertsoa, behean, eta hortik corpusera pasatako zatia, goian.

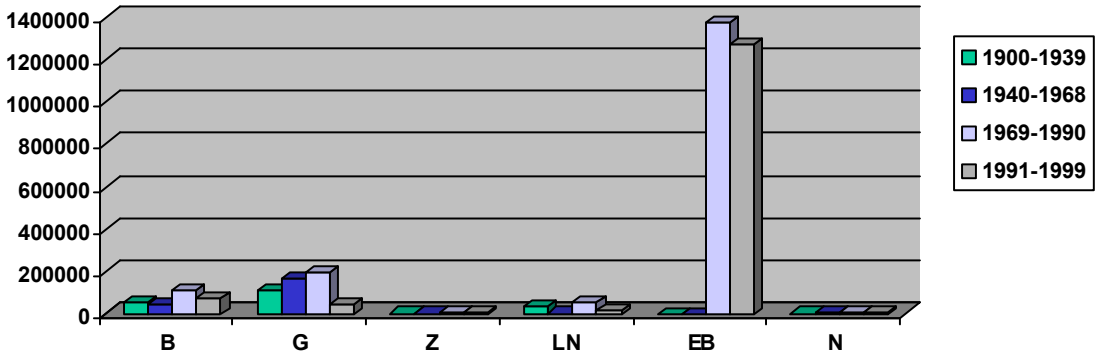



b) *Euskalkika*:

| *Euskalkia* | *Testu-hitzak (dokumentuak)* | *%* |
|---|---|---|
| Bizkaiera | 273.295 (332 dok.) | % 7.5 |
| Gipuzkera | 510.911 (614 dok.) | % 14.1 |
| Zuberera | 12.328 (20 dok.) | % 0.3 |
| Lapurtera/Nafarrera | 154.859 (240 dok.) | % 4.3 |
| Euskara batua | 2.651.935 (2235 dok.) | % 73.2 |
| "Nahasiak" | 18.704 (26 dok.) | % 0.5 |
| **Guztira** | **3.622.032 (3467 dok.)** | **% 100** |

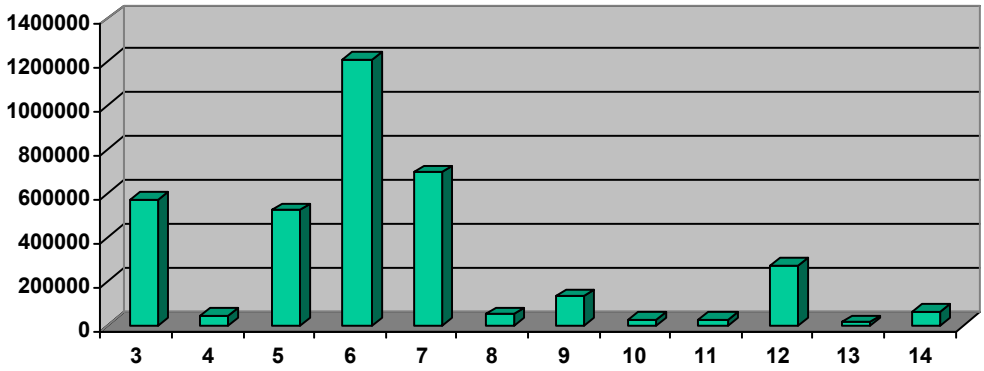



Epeak eta euskalkiak konbinatuz, berriz, hau da irudia:




c) *Testu-motaka*:

| Testu-mota | Testu-hitzak (dokumentuak) | % |
|---|---|---|
| Saio-artikuluak | 575.194 hitz (767 dok.) | % 15.9 |
| Administrazio-idazkiak | 47.550 (45 dok.) | % 1.3 |
| Ikasliburuak | 521.102 (390 dok.) | % 14.4 |
| Saio-liburuak | 1.207.055 (959 dok.) | % 33.3 |
| Prosa literarioa | 695.019 (630 dok.) | % 19.2 |
| Poesia | 47.882 (170 dok.) | % 1.3 |
| Antzerkia | 132.226 (129 dok.) | % 3.6 |
| Bertsoak | 24.207 (39 dok.) | % 0.6 |
| Ikerketa-lanak | 26.256 (17 dok.) | % 0.7 |
| Haur- eta gazte-literatura | 274.147 (255 dok.) | % 7.5 |
| Ahozko transkripzioak | 11.713 (11 dok.) | % 0.3 |
| Liturgia | 59.681 (55 dok.) | % 1.6 |
| **GUZTIRA** | **3.622.032 (3467 dok.)** | **% 100** |

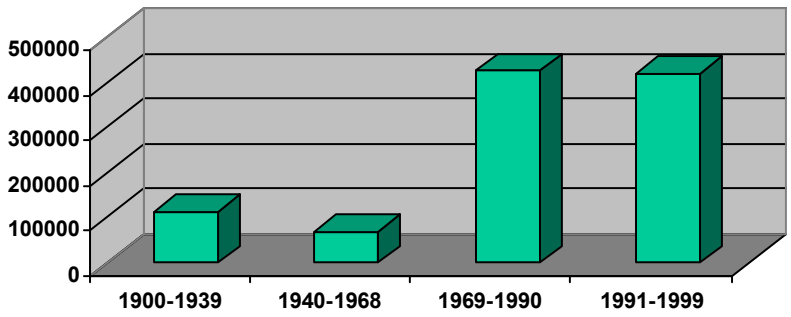



Aurreko datuon argitan ondoriozta dezakegu: azpicorpusaren % 32.5 literarioa da, eta ez-literarioa edo informatiboa % 67.5. Alegia, ia hirutik bat da literarioa, estandarrek markatzen dutenera hurbilduz. Baina datuok azpicorpusari bakarrik egiten diote erreferentzia, ez corpus osoari, gero ikusiko dugunez.

### 2.2.2.2. *Sailkatu gabeen azpicorpusa:*

a) *Epea*:

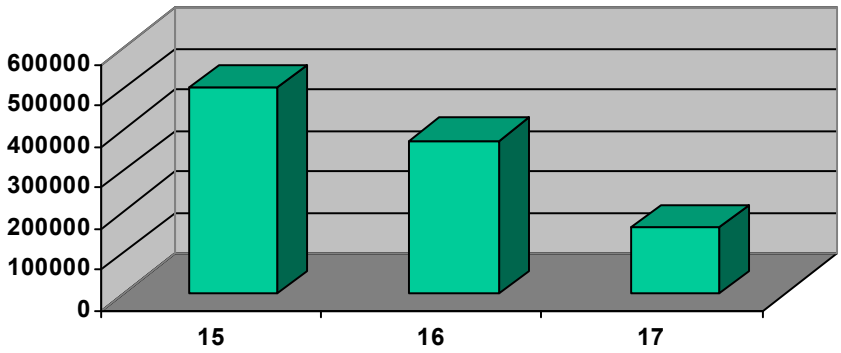
| | | |
|---|---|---|
| 1900-1939 | 114.294 hitz | % 11 |
| 1940-1968 | 70.221 hitz | % 6.8 |
| 1969-1990 | 429.429 hitz | % 41.5 |
| 1991-1999 | 421.190 hitz | % 40.7 |
| **Guztira** | **1.035.134 hitz (2885 dok.)** | **% 100** |



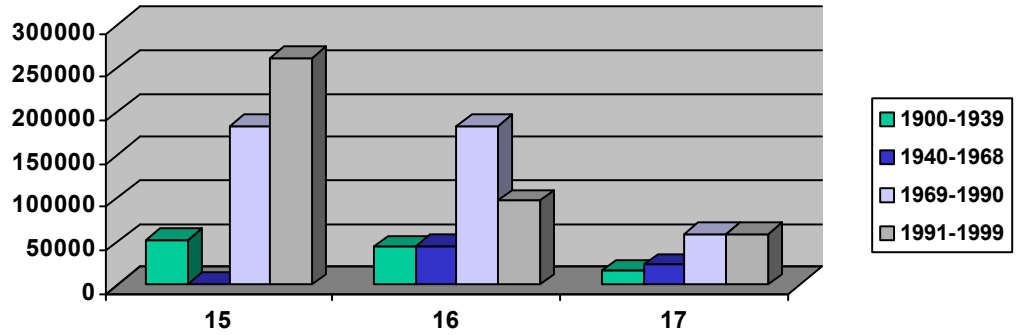

Hemen ere, sailkatuen atalean ikusi dugun bezala, unibertsoarekiko adierazgarritasuna eta proportzioa mantendu dira.

b) *Testu-mota*:

| | | |
|---|---|---|
| Egunkariak | 502.938 hitz (1467 dok.) | % 48.6 |
| Astekariak + hamaboskariak | 372.904 (1002) | % 36 |
| Aldizkariak | 159.292 (416) | % 15.4 |
| **Guztira** | **1.035.134 (2885)** | **% 100** |




Eta, aurreko biak, hau da, epea eta testu-mota konbinatuz:




Aurrekoa ikusita, hauek dira unibertsoa eta corpusaren arteko aldeak, corpusera pasa den proportzioa: sailkatu gabeen unibertsoak 510.326.796 hitz zituen eta, horietatik, azpicorpusera 1.035.134 hitz pasa dira, hau da, % 0.2 besterik ez. Adierazgarritasuna kontuan izan behar da; izan ere, kazetaritza-lanak ez dira hain "landutzat" hartzen normalean, testu-masa handia osatzen dute eta, ondorioz, corpusera pasatzen dena ere gutxiago izan ohi da.

### 2.2.3. Unibertsoa vs. corpusa

Unibertso osoa eta corpusa bere osotasunean erkatuz, hauek dira datuak:

| | *UNIBERTSOA* | *CORPUSA* | *%* |
|---|---|---|---|
| ***Sailkatuak*** | ***254.179.000*** | ***3.622.032*** | ***1.4*** |
| ***Sailkatu gabeak*** | ***510.326.796*** | ***1.035.134*** | ***0.2*** |
| ***Unibertsoa guztira*** | ***764.505.796*** | ***4.657.165*** | ***0.6*** |

Eta, unibertso osoaren % 0.6 horren barruan, sailkatu vs. sailkatu gabe banaketa hala geratu da:

| | |
|---|---|
| Sailkatuak: | % 77.7 |
| Sailkatu gabeak: | % 22.3 |

Corpusean, beraz, medioa osotara hartuta:

| | |
|---|---|
| Literarioa: | % 25.3 |
| Informatiboa: | % 74.7 |

nahiz hau ez den egia osoa. Izan ere, sailkatu gabeetan literatura ere ageri da, alegia, aldizkari literarioak testu-masa oso bezala jaso dira eta, ondorioz, datuok ez dira literaturaren sailean bildu. Beraz, literarioari gehixeago esleitu beharko litzaioke —ez gehiegi, sailkatu gabeetan ez baitira maiztasun handienekoak—.

Horiek horrela, corpusa orekatua dela esan dezakegu, bi arrazoirengatik gainera: 1) gaurko estandarren bidetik doa, eta 2) unibertsotik abiatua denez, XX. mendeko euskal argitalpenen errealitatea islatzen duen neurrian mantentzen du oreka.

Unibertsorako baliatu den sailkapena, gainera, lagungarri da kontsultak bideratzeko orduan, murrizketak ezar daitezke-eta. Baina, ezin dugu ahaztu inbentariatzea eta sailkatzea lan itzela dela ematen duen emaitza urrirako[10], urritzat har badaiteke behintzat.

### 2.3. Corpusa eskuratzea

Honaino, bada, corpusaren osaerarako irizpideak eta jarraitutako faseak. Baina, lagina aukeratu ondoren, eskuratze-lanari ekin behar zaio. Mende osoa landu dugula kontuan hartuta, bi lan-sistematara jo behar izan dugu: a) paperean eskuratu eta eskanerretik pasa, ondoren zuzenduz; eta b) formatu elektronikoan lortu eta egokitu. Azken urteetan bigarren aukerak gora egin du, baina sailkatu gabeen kasuan gehienbat. Izan ere, obra-zati txiki asko behar genuenez, askotan arinagoa zen eskanerretik pasatzea —dokumentuen kalitatea ona baitzen, eta eskanerrak azkarrak, lehen urteetan ez bezala— obra guztiak eskuratzeko bidea lantzea baino.

### 2.4. Corpusa egituratzea: kodetzea

Eskuratutako dokumentuak formatu bateratu batera ekarri behar dira. UZEIn SGML (*Standard Generalized Mark-up Language)* aukeratu dugu, elkartrukerako sistema erosoa baita, estandar izateaz gain. Corpusaren erabiltzaileak testuak letra-tipo desberdinetan (etzana, lodia edo azpimarratua) aurkituko ditu, testu originalean zegoen bezala. Baina, hauez gain, erdarak, aipamenak, metahizkuntza eta bestelakoak ere markatuta aurkituko ditu, egilearen erabilera bereziak edo beste norbaiti hartuak agerian utziz.

SGMLn Real Academia Española k osatu duen CREA[11] corpuserako baliatu zituen irizpideak gure egin genituen, behar ziren egokitzapenak eginez. Hala ere, 1900-1990 arteko corpusa kodetze-modu propio batean egin zen, gerora SGMLra pasa bada ere. Aldiz, 1991-1999 zuzenean SGMLn kodetu zen. Lau hau bi fasetan egin zen:

[10] Emaitza urria diogu helburua corpusa delako eta, beraz, aurrelan handia delako, unibertso sailkatutik abiatuz, corpusa osatzea.

[11] CREA: Corpus de Referencia del Español Actual. www.rae.es

1) azaleko markatzea: alegia, testua, orrialdea, paragrafoa eta halakoak, batetik, eta marka tipografikoak bestetik (**lodia**, *etzana*, azpimarra, "komatxoak", etab.).
2) informazio lexikografikoa: aurreko marka tipografikoak interpretatuz, erdarak, aipamenak, metahizkuntza eta halakoak ezarriz. Alegia, testua interpretatu egin zen, erabiltzaileak horren berri izan eta dokumentua modu egokian ustia zezan.

Eredu honekin, dokumentu originalera buelta daiteke, nahiz informazioa gehitua duen. Alegia, aberastua da, lexikografoak marka inplizituak esplizitu egin ditu-eta.

Adibide batekin ikusteko:

*(1) Testu originala:*

4506500012
00021
**I. Atala**
**HAURTZAROKO TRAGEDIA (1888-1905)**
**1. IRIARTEN HIRUZKIAK**
Orixeren drama, inoiz gainditu ez zuen drama mingotza, amarena izan da edo, agian hobeto esanda, amarik ezarena.
Gauza jakina da jadanik, askok askotan esana delako, Orixe hiruzkia zela.
Maria Manuela Inazia Pellejero amak, berak ama "Manazi" deitzen duenak, hiru sabelkide munduratu zituen 1888ko abenduaren 6an, eguerdiko hamabitan, Orexako Iriarte baserrian: bi mutil, Nikolas eta Martin(1), eta neska bat, Dionisia, berak etxeko izenez *Quito'n arrebarekin*-en "Denuxi“ deitzen duen arreba moja(2).

(1)Martin 2 urterekin hil zen, 1890-II-18an.
(2)"Damas Catequistas" izeneko Kongrazioan sartu zen 34 urterekin. Amerikan 47 urte egin ondoren, Loiolako komentura erretiratu zen zahartzaroan eta bertan hil 1979-VI-26an 90 urterekin, Institutuan 56 urte eginda gero. Komentuan esan didatenez (Elkarrizketa Loiolako "Damas Catequistas"eko Ama Nagusiarekin, 1897-II-13), erlijiosa oso jarraibidezkoa omen zen, santa fama zeukana. Baina, bide batez, sinestezinezko beste gauza harrigarri hau ere jakin dut Jose Mari Aranalderengandik: alegia, euskara erabat ahaztu zuela, hitz solte batzu ezik (Elkarrizketa J.M. Aranalderekin, Donostia, 1987-I-22).
Zentzu batean, konprenitzekoa ere bada, Amerikan 47 urte egin bait zituen, eta horietatik hainbat Quito-n. Horregatik jarri zion, hain zuzen, *Quito'n arrebarekin* izena idazlan honi.

*(2) Testu kodetua, SGML-markak agerian dituela:*

```
<!DOCTYPE CORPUS SYSTEM "/usr/EUSLEM/espz.dtd" >
<CORPUS>
<HEAD>
<TITLE>4506500012</TITLE>
</HEAD>
<BODY>
<PAGE>00021
<P><head rend="negr">I. Atala</head></P>
<P><head rend="negr">HAURTZAROKO TRAGEDIA (1888-1905) </head></P>
<P><head rend="negr">1. IRIARTEN HIRUZKIAK</head></P>
<P>Orixeren drama, inoiz gainditu ez zuen drama mingotza, amarena izan da edo, agian hobeto esanda, amarik ezarena.</P>
<P>Gauza jakina da jadanik, askok askotan esana delako, Orixe hiruzkia zela.</P>
<P>Maria Manuela Inazia Pellejero amak, berak ama <X rend="cdob">Manazi</X> deitzen duenak, hiru sabelkide munduratu zituen 1888ko abenduaren 6an, eguerdiko hamabitan, Orexako Iriarte baserrian: bi mutil, Nikolas eta Martin <note>1 Martin 2 urterekin hil zen, 1890-II-18an. </note>, eta neska bat, Dionisia, berak etxeko izenez <bibl rend="curs">Quito'n arrebarekin</bibl>-en <X rend="cdob">Denuxi</X> deitzen duen arreba moja <note>2 <foreign rend="cdob">Damas Catequistas</foreign> izeneko <corr sic="Kongrazioan">Kongregazioan</corr> sartu zen 34 urterekin. Amerikan 47 urte egin ondoren, Loiolako komentura erretiratu zen zahartzaroan eta bertan hil 1979-VI-26an 90 urterekin, Institutuan 56 urte eginda gero. Komentuan esan didatenez (Elkarrizketa Loiolako <foreign rend="cdob">Damas Catequistas</foreign>eko Ama Nagusiarekin, 1897-II-13), erlijiosa oso jarraibidezkoa omen zen, santa fama zeukana. Baina, bide batez, sinestezinezko beste gauza harrigarri hau ere jakin dut Jose Mari Aranalderengandik: alegia, euskara erabat ahaztu zuela, hitz solte batzu ezik (Elkarrizketa J.M. Aranalderekin, Donostia, 1987-I-22). Zentzu batean, konprenitzekoa ere bada, Amerikan 47 urte egin bait zituen, eta horietatik hainbat Quiton. Horregatik jarri zion, hain zuzen, <bibl rend="curs">Quito'n arrebarekin</bibl> izena idazlan honi.</note>.</P>
```

Adibidean ikus daitekeen moduan, izenburua markatua dago (kasu honetan sailkapen-kodea[12] besterik ez dugu bistaratua), bai eta orrialdea ere (21.a). Eta, horrekin batera, paragrafo-hasierak eta –bukaerak ere markatuta daude <P>ren bidez. Goiburukoak <head> gisa ageri dira, lodian daudela markatuz gainera. Testu barruko markei dagokienez, oin-oharrak dagokien hitzaren ondoren txertatu dira (<note>); erdal hitzak (<foreign>) eta erreferentzia bibliografikoak (<bibl>) berariaz kodetu dira, euskal lexikoaren ustiapenean kanpoan gera daitezen; akatsak detektatu eta zuzendu dira (<sic> eta <corr>), etab. Erabiltzaileak lehen begiratu batean ikusten dituen horiek tresnari erakutsi egin behar zaizkio, ustiapenean datuok behar bezala erabil ditzan eta, ondorioz, emaitza okerrik eman ez dezan.

[12] 4506500012 hala irakurtzen da, digituka: 4 (4. epea, 1991-1999), 5 (euskalkia, euskara batua), 06 (testu-mota, saio-liburua), 5 (obraren tamaina, 251 orrialde baino gehiagokoa) eta 00012 (obraren eta autorearen erreferentzia). Datuok ez dira hemen bistaratzen, beste taula batean azalduak daude-eta.

Testua egokitu, zuzendu eta markatu ondoren, guztia bateratua eta interpretatua dugunean, corpusa lexikografikoki lantzen hasteko moduan izango da. Bukatu dira, beraz, aurrelanak.

### 2.5. Corpusa lematizatzea

Euskal lexiko gisa utzi ditugun hitz guztiak (kodetzean baztertu beharrekoak baztertu ondoren, iragazki baten bidez horiek ez baitira hustuko) lematizatzea da hurrengo urratsa, alegia, testu-hitz bakoitzari lema estandar bat esleitu behar zaio, hiztegi-sarrera modukoa izango dena. Horrek, besteak beste, erraztu egingo du kontsulta. Adibide batekin esateko, forma deklinatuei eta aldaerei lema bakarra ezarriko zaie eta horixe izango dugu helduleku, erosoa gainera. *Etxe* lemaren kontsulta egiten badugu, *etxe*, *etxea*, *etxeetan*, *etxien*, *echeco*, *Etcheco*, *etchetik* eta beste ehunka aldaera eskuratuko ditugu berehala, aldaeraren bat ahazteko arriskurik gabe. Lematizazio-kasurik arruntena da hau.

Lematizazio hau, bestalde, ez da hitz bakunetara mugatzen: hitz soilez gain, hitz elkartuak, eratorriak eta bestelako hitz anitzeko unitate lexikalak ere markatu dira. Alegia, lema hiztegi-sarrera modukoa dela esan berri dugu, baina hori baino gehiago da: hiztegietan sarrera, azpisarrera edo are adibide gisa markatzen dena, hemen maila berean jasoa baitago, unean behar dugun hitza berehala aurkitzea ahalbidetuz. Aurreko adibidearekin jarraituz, *etxe* lema soilaz gain, *abeletxe*, *argitaletxe*, *bainuetxe*, *etxe orratz*, *etxe-abere*, *etxe-tresna*, *etxeko*, *etxeko jaun*, *etxeko andre*, *etxepe*, *etxezain* eta beste hainbat aurkituko ditugu, bilaketak mugatuz. *%etxe%* eskatuta, hau da, lemaren barruan, edozein posiziotan, *etxe* duena eskatuz, orain arte aipatu ditugun guztiak eta gehiago bistaratuko zaizkigu. Aldiz, *etxe%* idatziz, horrela hasiak, edo *%etxe* eskatuz, hala bukatuak eskuratuko ditugu. Baina hau da garrantzitsua: ez ditugu grafia eta kasu-marka guztiak jakin edo zehaztu behar forma bat bilatu ahal izateko.

Beste adibide batek erabilera gramatikaletan eman dezakeen laguntza erakutsiko digu: *hala* forma soilaz gain, *hala bada*, *hala edo hala*, *hala ere*, *hala eta*, *hala eta guztiz ere*, *hala moduz*, *hala... nola*, *hala nola* eta abar lematizatua aurkituko dugu eta, ondorioz, horietako bakoitzaren erabilerak bakarrik —baina corpusean dauden guztiak— eskuratzeko aukera izango dugu.

Adibideok lematizazioaren laguntza erakusten dute, baina, zergatik behar da corpusa lematizatua izan? Badaude lematizazioa ezinbesteko egiten duten hiru arrazoi behintzat:

a) Hizkuntzaren tipologia: euskararen kasuan, esaterako, eranskarien taldean sartzen den hizkuntza izanik, deklinabidea (*umiendako* -> *ume*) eta aditz jokatuak (*diñosku* -> *esan*) forma estandar batera ekartzea beharrezkoa da.

b) Hizkuntzaren egoera: egia da normalizazioa eta araugintza bideratzen ari direna, aurrerapauso handiak egin direna, baina corpusean biltzen diren dokumentuek erabilera errealari erantzuten diote eta, bistan da, aldaerak maiz azaltzen dira testuetan (ez dira, gainera, araugintza berriaren ondoko testuak bakarrik: XX. mende osoa hartzen dute, euskara batua sortu aurrekoak ere bai): *jardun / ihardun*, *arazi / erazi*, etab. Noski, besterik dira euskalkietako aldaera lexikoak, horiek bai baitute lekua, euskalkia islatzen duten neurrian.

c) Hitz anitzeko unitate lexikalen helduleku[13]: hitz elkartuak (*hizkuntza-corpus*), aditz-perifrasiak (*behar izan*), aditz konposatuak (*negar egin*, *min*

[13] Hauen osagaiak askotan ez dira elkarren segidan agertzen, eta beste osagai batzuk tarteka daitezke gainera (*ederki jo zigun adarra* -> *adarra jo*, esaterako).

*eman*), lokailuak (*hala eta guztiz ere*), lokuzioak (*adarra jo*, *hanka sartu*), terminoak (*kafea jotzeko tresna*), etab.

Halere, corpus honen lematizazioaren barruan zehaztapen batzuk egin behar dira. Izan ere, hasieran aipatu dugu corpusaren helburua lexiko arrunta zela, eta horrek ondorioak izan ditu lematizazioan: "lematizazio faltsua" da neurri batean:

a) aditz laguntzaileak ez dira lematizatu, ez dira-eta lexiko arruntaren erakusgarri. *Etorri naiz*-en *naiz* **$$$** bezala lematizatu da, baina *ni naiz*-en *naiz*, berriz, **izan**.

b) zenbatzaile kardinalak **$$** gisa lematizatu dira: *hiru* -> **$$;** ordinalak, aldiz, lema osoa dute: *hirugarren* -> **hirugarren**, eta *3.* -> **hirugarren**, atzizkia[14] dute-eta.

c) izen propioak: hauek ere **$$** gisa lematizatu dira. *Donostia* -> **$$,** *Agirre* -> **$$,** *Kutxa* -> **$$** (baina *kutxa* -> **kutxa**, izen arrunta). Izen propioetan salbuespen bakarrak euskal osagai arruntek osatuak dira: *Euskal Herria* -> **Euskal Herri**, *Erresuma Batua* -> **Erresuma Batu**.

Metahizkuntza lematizatu egin da, kodetua egotean, markatua baitago. Beraz, lema horrek SGML-marka duen forma bati erantzuten dionez, badakigu erabilera markatukoa dela.

Horiek horrela, XX. mendeko euskararen corpus estatistikoak **104.817 lema desberdin** ditu guztira. Baina datu hau zehaztea komeni zaigu. Ikus ditzagun maiztasun handieneko 23 lemek (20 erreal eta 3 "faltsu") eskaintzen dizkiguten emaitzak:

| *Lema* | *hitz-kopurua* | *%* |
|---|---|---|
| $$$ | 399.665 | 8.6 |
| $$ | 324.536 | 7 |
| k | 212.351 | 4.5 |
| eta | 192.035 | 4.1 |
| izan[15] | 124.036 | 2.6 |
| ez | 74.950 | 1.6 |
| hau | 44.914 | 0.9 |
| egin | 44.227 | 0.9 |
| bera | 43.239 | 0.9 |
| egon | 41.350 | 0.8 |
| hori | 39.780 | 0.8 |
| ukan | 32.993 | 0.7 |
| ere | 32.156 | 0.7 |
| esan | 25.397 | 0.5 |
| beste | 23.6741 | 0.5 |
| baina | 23.093 | 0.5 |
| behar izan | 22.785 | 0.5 |
| edo | 21.798 | 0.4 |
| hura | 18.814 | 0.4 |
| guzti | 18.066 | 0.3 |
| gu | 17.235 | 0.3 |
| eman | 17.032 | 0.3 |
| ikusi | 15.268 | 0.3 |

Esan dugunez, **$$$** aditz laguntzaileak dira, **$$** izen propioak eta zenbatzaile kardinalak eta, azkenik, **k** lemaren osagaia, sarrerako helduleku ez dena. Honek corpusaren helburua erakusten du berriro ere: lexiko arrunta da interesa duena, ez bestea. Halere, hiru "lema" hauek corpus osoaren % 20,1 hartzen dute, hau da,

[14] Alegia, euskal hitz-eraketako osagaia.

[15] *Izan* aditz nagusi soila da leman honetara bildua. Laguntzailea *$$$* leman jasoa dago, eta, aditz-perifrasietakoa, zein bere lekuan (ikus, zerrenda honetan bertan, *behar izan*, adibidez).

osoaren 1/5. Beraz, lexiko arruntaren aldetik, corpus "erabilgarria" 3.720.583 hitzekoa da. Hain zuzen, lema "erreal" horiekin zein osotasunarekin kalkulua eginez:

- Maiztasun handieneko 10 lemek corpus osoaren % 14 hartzen dute, lema errealen % 18.
- Maiztasun handieneko 20 lemek corpus osoaren % 18 hartzen dute, lema errealen % 23,4.
- Maiztasun handieneko 100 lemek 1.429.604 hitz hartzen dituzte, hau da, corpus osoaren % 30,7; lema errealen % 38,4. Lema errealak hala banatzen dira, hamarreko taldeetan:

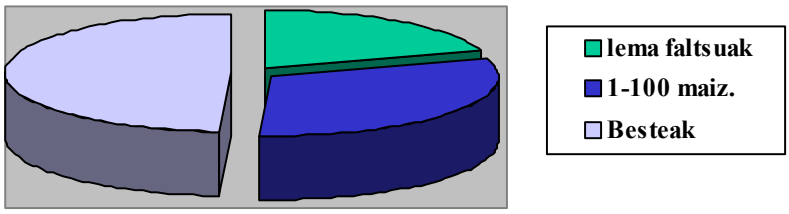
| ***Maiztasuna*** | ***Agerraldi-kopurua*** |
|---|---|
| 1-10 | 669.680 |
| 11-20 | 203.129 |
| 21-30 | 125.902 |
| 31-40 | 91.247 |
| 41-50 | 74.981 |
| 51-60 | 65.607 |
| 61-70 | 58.806 |
| 71-80 | 51.981 |
| 81-90 | 45.570 |
| 91-100 | 42.701 |

- Azkenik, agerraldi bakarreko 55.144 lema daude: corpus osoaren % 1,18 hartzen dutenak, errealen % 1,48. Baina, lema desberdinak kontuan hartuta (104.817, alegia), horien erdia baino gehiago dira corpusean behin[16] bakarrik jaso direnak: 52,6, hain zuzen.

| | |
|---|---|
| arte giro | kardiologia |
| arte hizkuntza | kardiologiko |
| arte idorokunde | kardiologo |
| arte ikatz | kardu ondoxka |
| arte ikerlari | kare beltz |
| arte ikuspegi | kare bizi |
| arte ipurdi | kare alga |
| arte jardun | kare altzionario |
| arte kondaira | kare apur |
| arte konposaketa | kare behar |
| | kare estruktura |
| | kare fabrika |

Aurreko guztia irudi batean bilduz argi ikusi ahal izango dugu orain arte azaldutakoa:

[16] Honek azalpen argia du: hitz elkartu automatikotzat har daitezkeen guztiak lematizatu dira. Halere, *kardiologia*, esaterako, behin baino ez da corpusean azaldu.




Honaino XX. mendeko euskararen corpus estatistikoaren lematizazioak emandako emaitzak; baina, nola egin du hori UZEIko lantaldeak?

a) 1900-1990 urteak hartzen dituen zatia Baionako HIZKIA etxeak prestatutako *RTerm* programa baliatuz lematizatu zen. Corpus-zati bat eskuz lematizatu ondoren, hitza—lema bikoteak irakatsi zitzaizkion lematizatzaileari eta, dokumentuak lematizatu ahala, zerrenda aberastu egiten zen. Ezagutza linguistikorik gabeko tresna zen, eta automatikoki ezartzen zituen lema guztiak lexikografoek berrikusten zituzten.

b) 1991-1999: automatikoki lematizatu zen, UZEIren eta IXA taldearen elkarlanaren ondorioz sortutako EUSLEM lematizatzaile automatikoa baliatuz. Baina, hemen ere, *RTerm*-ekin egin zen bezala —askoz ere modu azkarragoan, noski—, guztia berrikusi zen eta eskuz zuzendu, osatu edo desanbiguatu, beharrezko zenean.

## 3. CORPUSAREN APLIKAZIOAK

XX. mendeko euskararen corpus estatistikoa, bere muga guztiekin ere, hainbat lanetan aplikatu da eta baliagarri dela erakutsi du:

1. Hiztegi Batua: Euskaltzaindiko Hiztegi Batuko Lantaldeak, bere hileroko bileretan hainbat forma aztertu ohi ditu eta, horien gaineko erabakiak hartu ahal izateko, formen historiala dokumentatua behar du. Material hori UZEIko prestalaneko taldeak helarazten dio eta, hain zuzen, abiapuntu gisa corpusa izan ohi du:
   - aztergaia osatzeko, maiztasunen araberako hitz-zerrendak prestatu eta horietan oinarritzen da; alegia, zein forma landuko den corpusetik ateratzen da.
   - forma horiei buruzko historialak egiteko, dokumentatu egin behar da eta, horretarako, tradizioko hitzetan *Orotariko Euskal Hiztegia*-ko corpusa baliatzen den bezala, azken mendeko erabileren argazkia *XX. mendeko euskararen corpus estatistikoa*tik ateratzen da. Tradiziorik gabeko forma berrietan, erabakitzen zailtasunik handiena dutenetan, hain zuzen, haien nondik norakoa bil daiteke corpusean.
2. Gramatika: Euskaltzaindiko Gramatika Batzordeak ere, egungo erabileren berri emateko, tradiziotik datozen erabilerekin bat ez datozenak dokumentatzeko, XX. mendeko corpusa baliatu izan du.
3. Ikertzaileak, unibertsitatekoak batez ere:
   - hipotesiak frogatzeko: aditzen erregimena, espazioa eta denbora markatzen duten elementuak, edo osagaien urruneko mendekotasuna kontrolatzeko, besteak beste.

- erabileren helduleku gisa: hitz anitzeko unitate lexikalen multzo zabalak, hitz konkretuen familiak —eratorriak eta konposatuak landu ahal izateko —, eratorpenaren sistematizazioa aztertzeko, atzizki zehatzak, etab.
- tresnak trebatzeko: lematizatzailea, ahotsaren tratamendurako oinarrizko ezagutza bideratzeko, etab., horiek corpusetan oinarritu beharra baitute, alegia, testu-masa handiak behar dituzte.

4. Lexikografoak eta terminologoak, hau da, hiztegigintzan dihardutenak. Azkenean, galdera hauei erantzuteko behar izaten da corpusa: forma bat dokumentatua dago? zeinek erabili du? arlo zehatz bati dagokio? bestetarako balio dezake? gaur bizi da? euskalki- edo erabilera-markarik behar du?

Gaur corpusa sarean dago kontsultagai, edozeinen eskura, www.euskaracorpusa.net helbidean. Sarean jartzeak erabiltzaileen gorakada ekarri du, zuzenean erabil daiteke-eta. Erabiltzaile-mota zabala duela erakutsi du, maiztasunak besterik behar ez dituztenekin batera, edo adibide hutsak, ikerketarako datuak eskuratzen dituztenak ere bai baitaude[17].

## 4. CORPUSAREN MUGAK

Orain arte ikusitakoarekin, corpusa lagungarri eta oso erabilgarri dela erakutsi dugu; baina, baditu bere mugak, ondoren ikusiko dugunez. Dena den, bere garaian kokatu behar da, helburu zehatz batekin sortu baitzen: *XX. mendeko euskal lexikoaren berri ahalik eta zabalena eskaini*. Dudarik ez da, corpusak erabilera askotarikoa beharko lukeela, ikusi dugunez, baina 1986an jarri zen martxan proiektua, baliabide mugatuekin (lantaldea, dirua, ezagutza eta denbora, guztiak ere urriak ziren hasierako garai haietan). Aipa ditzagun muga nagusiak:

1. Ez da "erreferentea", alegia, txikia da, testu-masa urria du euskararen erabileren berri zabala emateko, adibidetegi mugatua da, maiztasun urriko adierak ez aurkitzeko arriskua dago. Horrez gain, estatistikoa da eta, ondorioz, autoreak eta obrak ez dira garrantzitsuak, ez dira ezeren erreferente.

2. Ez dago etiketatua, informazio morfosintaktikoa ez dago, modu esplizituan behintzat, eskuragarri. Hiru atal nagusi ukitzen ditu honek:

   a) Kategoria eta azpikategoria gramatikalik ez da eskaintzen. Ondorioz, erabiltzailearen esku geratzen da halakoak argitzea eta desanbiguatzea. Adibidez, *iritzi* lemak bi kategoria ditu: izena eta aditza.

   b) Adiera-banaketak ere ez dira eskaintzen. Esaterako, *arte* lemak bost adiera ditu, gutxienez, euskaraz: a) zuhaitza, b) ertia, c) trebetasuna, d) bitartea, e) denbora-hedadura. Horien berri ez da ematen corpusean.

      Argi dezagun, halere, ez dela internet bidezko kontsultan eskaintzen, nahiz UZEIko lantaldeak horiek guztiak lantzen dituen Hiztegi Batuko

[17] 2002. urtean, otsailean jarri zen kontsultagai eta lehen hilabeteetan, batez beste, eguneko 37 kontsulta izan zen, baina urrian 53ko media da eguneroko kontsultena. Kontsultagileak ez dira Euskal Herrikoak bakarrik, ez behintzat hemen bizi direnak. Ameriketako Estatu Batuetatik hainbat kontsulta egiten dira, Washingtondik (eta, zehatzago, Seattletik) eta Californiatik (San José) batez ere, nahiz besteetatik ere hainbat sarrera dauden. Mexiko, Argentina eta Txile ere bisitarien artean ageri dira. Europa mailan, ia herrialde guztietatik (Britainia Handia, Suedia, Polonia (Lodz), Herbehereak, Alemania, Italia,...), Kanadatik. Bestalde, bakanago bada ere, Australia eta Japongo kontsultak izan dira, bai eta Afrikatik ere, Nigeriatik, hain zuzen. Dena den, hainbat bisitaren jatorria ezezaguna da eta ez dago horiei buruzko datuak ematerik.

Orain arteko datuok erakusten dute Euskal Herritik kanpo ere kontsultatzen duenik badela, eta diasporan dauden euskaldunak aipatu behar dira gainera: EEBBetan hainbat euskal ikertzaile dago, Hego Amerikako euskaldunak ere sartzen dira web gunean, eta Europan zehar dauden euskaldunak izango dira, seguruena, bisitari ugari horiek. Hori da, hain zuzen, interneti zor diogun gauzatako bat: corpusa kontsulta dezakete atzerrian diren euskaldunek ere.

Lantaldearentzat prestalana egiten duen neurrian. Gainera, lematizazio automatikoak kategoriak helarazten ditu, nahiz kasuan kasuan berrikusi beharrekoak diren.

c) Menderagailuen informaziorik ez da azaleratzen lematizatzean eta testu-hitzera jo behar da hori berreskuratzeko. Adibidez, *gaudelako* forma *egon* gisa lematizatzen da eta, ondorioz, aditz-formaren datuetatik bilaketarik ezin da egin (aditz-mota, aspektua, pertsona, erlazioa, etab.). Ber gauza dugu *baletor* formarekin, *etorri* baita eskaintzen den informazio bakarra. Noski, honek corpusa morfosintaktikoki etiketatua izatea eskatuko luke, erabilera askotarikoa izan zedin.

Muga horiek kontuan izanda, kontsultarako hobekuntza batzuk bideratu dira UZEIn, Euskaltzaindiko Gramatika Batzordearekin adostuak. Izan ere, batzordeko kideei aurkeztu zitzaien corpusa eta, haien beharrak kontuan hartuz, kontsulta-sistema berrantolatu zen. Horren emaitza da:

a) lema + testu-hitza konbinatuz egin daitekeen kontsulta. Esaterako, *jakin* lema eta *%la* edo *%na* bukaera duen hitza, hitzaren posizioa eta tarteko elementuena ere zehaztuz, hala komeni denean.

b) autore edo obraren araberako bilaketarik ez da eskaintzen, obra-zati asko —beraz, autore eta obra asko—, baina ez-osoak direlako. Halere, sailkapenaren araberako bilaketak egin daitezke (epeak, euskalkiak eta testu-motak konbinatuz). Kontsultaren emaitzan, hori bai, erreferentzia bibliografikoa eta obraren orrialde oso bat ikusteko aukera eskaintzen da.

## 5. LANTALDEA

Corpusa osatu duen lantaldea UZEIko Lexikografia Sailekoa da, 7 lagunek osatua (5 hizkuntzalari/lexikografo, informatikari 1 eta idazkari 1). Lantaldeak, 1986tik corpusaren osaeraren fase guztietan —Hiztegi Batuko Lantaldearentzat prestalana egitearekin batera— jardun duenez, trebatzen joan behar izan du, tresna berrietara egokituz joan baita proiektua bere hasieratik gure egunetara.

Baliabide informatikoei dagokienez ere, egungo bertsioak ezaugarriok ditu: Solaris 2.6 sistema eragilea, Oracle 8.1.7 datu-base kudeatzailea eta, aplikazioetarako, Java programazio-hizkuntza erabili da.

## 6. ONDORIOAK ETA ETORKIZUNEKO LANAK

Txostenean ikusi ahal izan dugunez, *XX. mendeko euskararen corpus estatistikoa*k bere egitekoa bete du, hasierako *Hiztegi Hiritar Arauemailea*ri ez, baina *Hiztegi Batua*ri helarazi dio eman ziezaiokeen informazioa, eta hala jarraitzen du orain ere. Horrez gain, erabiltzaileen eskura jarri den lehen euskal corpusa da eta, interneten —doan— kontsultagai dagoenetik dituen kontsultak ikusita, horren beharra zegoela ikusi dugu. Aipagarria da, gainera, Euskal Herritik kanpo egiten den kontsulta-kopurua, atzerrian dauden euskaldunena batez ere. Beraz, lehen ondorio gisa, corpusa behar zela esan dezakegu. Une batean ia bost milioi euskal hitz kontsulta daitezke, murriztapenak egin daitezke bilaketan eta lematizatua dago. Hitz-zerrendak bakarrik kontsulta daitezke, hitzen edo/eta lemen arteko konbinazioak, haien testuinguruak (paragrafoa edo orrialde osoa), erreferentzia bibliografikoak. Alegia, laguntza berri bat izan dugu 2002. urtean euskalgintzan dihardugunok.

Ez dugu ahaztu behar, halere, txosten honetan aipatu ditugun corpusaren mugak, egungo —eta, batez ere, biharko— beharretarako ez baita erreferente gardena: txiki geratu da, estatistikoa baita, etiketatu gabe dago, baina bere garaian eta

testuinguruan kokatu behar da. Euskaltzaindiaren Hiztegigintza Sailerako, EUSKAL LEXIKO MODERNOAren berri emateko sortu zen eta, hain zuzen, horri erantzuten dio: *XX. mendeko euskal lexikoaren erabileren berri ematea* zuen helburu.

Dena den, gaur baliabide handiak ditugu material aldetik —sarean edo euskarri elektronikoan daude azken urteetako argitalpenik gehienak—, hizkuntza naturalaren prozesamenduan dihardutenek tresna egokiak sortu dituzte halako lanei ekiteko, eta ezagutza ere ez da 1986koa, esperientziadun adituak baitaude.

Beraz, aurrera begira, corpusaren alde onak eta hobe daitezkeenak ikusita, dudarik ez dugu etorkizunean —etorkizun hurbilean gainera— euskara modernoaren erreferente izango litzatekeen corpus etiketatu bat beharko genukeela, XXI. mendeko euskararen erreferentzia-corpusa dei genezakeena beharbada: *erabilera askotariko erreferentzia-corpusa,* tamainaz handia, erreferentea, egituratua, etiketatua, guztion eskura. Guztiona eta guztiontzat, ondoko lanetarako oinarri ezinbestekoa.