eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Ingeniaritzako Gradua

Konputazioa

Gradu Amaierako Proiektua

# Lehen urratsak euskarazko testu anonimoen egile-esleipenean

Egilea

*Ioanes Ceballos*

informatika fakultatea
facultad de informática

2014

# Laburpena

Hizkuntzaren Prozesamenduan bada alor bat, testu anonimo bat emanda, eta anonimoak ez diren garai beretsuko testu multzo bat izanda, testu anonimoaren egile probableena zein den ebazten saiatzen dena. Lan honetan, egileen esleipenean egin diren aurrerapenak aztertu eta euskarara ekarri nahi izan dira. Zehazki, Bertsolari Txapelketa Nagusietako bertsoekin metodoa fintzea izango da helburua, eta gaur egungo bertsoen egileak zein diren asmatuko dituen sailkatzailea garatzea.

# Gaien aurkibidea

# Irudien aurkibidea

# Taulen aurkibidea

# 1. KAPITULUA

# Sarrera

Gaur egun, Internet bidez informazio ugari atzi daiteke, eta horren zati handi bat testu moduan aurki dezakegu. Hala ere, testu horiek analizatzeko tresna egokirik gabe eskaintzen diguten informazioak ez du ezertarako balio. Besteak beste, azken urteetan testu horiei probetxua atera nahian, egile-esleipenaren arloak interesa piztu du. Arlo honetan, helburua egile bat edo egile batzuen estiloa ikasiko duen algoritmo bat garatzea da, ondoren, egile horien etorkizuneko dokumentuak automatikoki identifikatu ahal izateko.

Egile-esleipenaren problemaren ebazpenak, konputazioan eta hizkuntzaren prozesamenduan eginiko garapenetan oinarritzen dira. Hala ere, pausorik garrantzitsuena testuko ezaugarriak erauzi eta dokumentuak errepresentatzean datza. Horretarako, egilearen estiloa ondoen islatzen duten ezaugarriak erauzi eta erabiltzen dira. Aplikatzen diren teknikei dagokienez, hasieran teknika estatistiko soilak erabiltzen ziren arren, pixkanaka ikasketa automatikoko metodoak gero eta arrakasta handiagoa dute.

Ezertan hasi baino lehen, gure lanaren helburua aztertu eta txostenaren egitura azalduko dugu.

## 1.1 Helburua

Proiektu honen helburu orokorra, azken urteetako bertso bat edo etorkizunean sortuko den edozein bertso emanda, bertsolari probableena zein den esango digun sistema diseinatu eta garatzea da.

Helburu zehatzagoetan barneratuz, lan honetan ondorengoak defini ditzakegu:

- Bertsokera adieraziko duten atributu egokienak aurkitu.
- Atributu horiek konbinatu, bertsolaria hobeto identifikatzeko.
- Bertso-esleipenean emaitza onenak ematen dituen ikasketa-algoritmoa bilatu.
- Garatutako sistemaren gaitasuna frogatu, hainbat saio eginez bertsolari kopuru desberdinak erabiliz.

## 1.2 Txostenaren egitura

Lehenengo, 2. kapituluan proiektuaren helburuen dokumentua aurkitzen da, proiektuaren plangintzari buruz diharduena. Jarraian, 3. kapituluan, egile-esleipenaren arloan azken hamarkadan egindako lanak aztertu eta gure lanerako ondorioak biltzen saiatuko gara. Honen ostean, 4. kapituluan, aurrekarietatik ikasitakoa praktikan jarriz, proiektuaren garapenaren nondik norakoa azaldu eta lortutako emaitzak aurkeztuko ditugu. Azkenik, 5. kapituluan, proiektu honetatik ateratako ondorioak bildu eta etorkizuneko lanak azalduko ditugu.

# 2. KAPITULUA

## Proiektuaren Helburuen Dokumentua

Azken hamarkadan testu anonimoen egile-esleipenean egin diren aurrerapenez baliatuz, hainbat lan aztertu eta euskarazko testuekin lehen urratsak eman nahi dira. Zehazki, Bertsolari Txapelketa Nagusiko bertsoekin metodoa findu nahi da, gaur egungo bertsoen egileak zein diren asmatuko duen sailkatzailea garatuz.

Problema honi hurrengo pausoekin egingo diogu aurre:

1. **Aurrekarien azterketa**. Aipatu dugun moduan, arlo honetan azken urteetan egindako aurrerapenak aztertuko ditugu.

2. **Corpusaren prestaketa**. Azken urteetako Bertsolari Txapelketa Nagusietako bertsoen corpusa lortu beharko dugu, eta gure lanerako egokitu.

3. **Ezaugarri linguistikoak**. Egile-esleipenetan pausorik garrantzitsuena da. Testu batetik egilearen estiloa adieraziko duten ezaugarriak erauztea da helburua, eta horiekin bertso bat errepresentatzea.

4. **Ikasketa-algoritmoak**. Egile-esleipena egiteko erabiltzen diren metodoak ikasketa automatikokoak izan ohi direnez, guk ere bide hori jarraitu eta hainbat algoritmo probatuko ditugu.

5. **Esperimentuak**. Proba ugari egingo dira, garatutako sailkatzailerik onena aurkitu nahian.

Kapitulu honetan, 2.1 atalean proiektua deskribatu eta helburuak ikusiko ditugu. Jarraian, 2.2 atalean proiektuaren plangintza aurkeztuko dugu. Honen ostean, 2.3 atalean proiektua aurrera eramateko lan-metodologia deskribatuko dugu. Ondoren, 2.4 atalean proiektuaren bideragarritasuna bermatzen saiatuko gara. Azkenik, 2.5 atalean proiektuaren garapenean sor daitezkeen arriskuak identifikatu, eta horien aurrean hartu beharreko neurriak erabakiko ditugu.

## 2.1 Proiektuaren deskribapena eta helburuak

Aipatu dugun moduan, proiektu honen helburua egileen esleipenean egin diren aurrerapenak aztertu eta euskarara ekartzea da. Zehazki, Bertsolari Txapelketa Nagusietako bertsoekin metodoa findu nahi da, 1989.urtetik aurrera egindako txapelketetako bertsoen egileak zein diren asmatuko dituen sailkatzailea garatuz.

Lana aurrera eramateko, Bertsolari Txapelketa Nagusietako bertsoen corpusa beharko dugu. Corpus hau gure beharretara egokitzea izango da lehenengo lana. Nahasketak ekiditeko, bertsolariek binaka eginiko saioak alde batera utziko ditugu. Horrela, 110 bertsolarik banaka botatako bertsoak dauzkagu. Egile-esleipen lanetan egile kopurua oso garrantzitsua da, zenbat eta egile gehiago, problemaren konplexutasuna orduan eta handiagoa baita. Horregatik, proiektu honetan hainbat bertsolari kopururekin egingo ditugu probak. Gauzak horrela, proba bakoitzean dagokion bertsolari kopuruko corpusa eraiki beharko dugu.

Egile-esleipena egiteko gehien erabiltzen diren metodoak ikasketa automatikokoak dira, eta guk ere bide hori jarraitu eta hainbat algoritmo probatuko ditugu.

Horrela, egile-esleipena bertsolaritza mundura ekarriko dugu, eta probetarako aukeratutako edozein bertsolariren bertso bat emanda, egile probableena zein den esango digun sailkatzailea garatuko dugu.

## 2.2 Proiektuaren plangintza

### 2.2.1 Emangarriak

Proiektuan ondorengo emangarriak identifikatu dira:

1. **Memoria**

   Proiektuaren inguruko zehaztasun guztiak biltzen dituen dokumentua da. Bertan proiektuaren deskribapena, aurrekarien azterketa, plangintza, garapena, emaitzak eta ondorioak azalduko dira.

2. **Proiektuaren helburuen dokumentua**

   Memoriaren zati bat izango da, proiektuaren hasieran garatua. Bertan proiektuaren plangintzari buruzko informazioa azalduko da.

3. **Aurkezpena**

   Proiektua amaitzean proiektua bera defendatzeko prestatuko den aurkezpena. Bertan proiektuaren nondik norakoa azaldu beharko da.

4. **Corpus egokitzailea**

   *Perl* lengoaian idatzitako bi programa izango dira. Lehenengoarekin, nahi dugun bertsolari kopurua aukeratu, eta Bertsolari Txapelketa Nagusietako bertsoen corpusetik, bertsoak bertsolarika erauzita corpus berri bat sortuko dugu. Bigarrenarekin, sortutako corpus horretatik abiatuz *Train/Develop/Test* banaketa egingo dugu.

5. **Estilo-ezaugarriak ateratzeko programak**

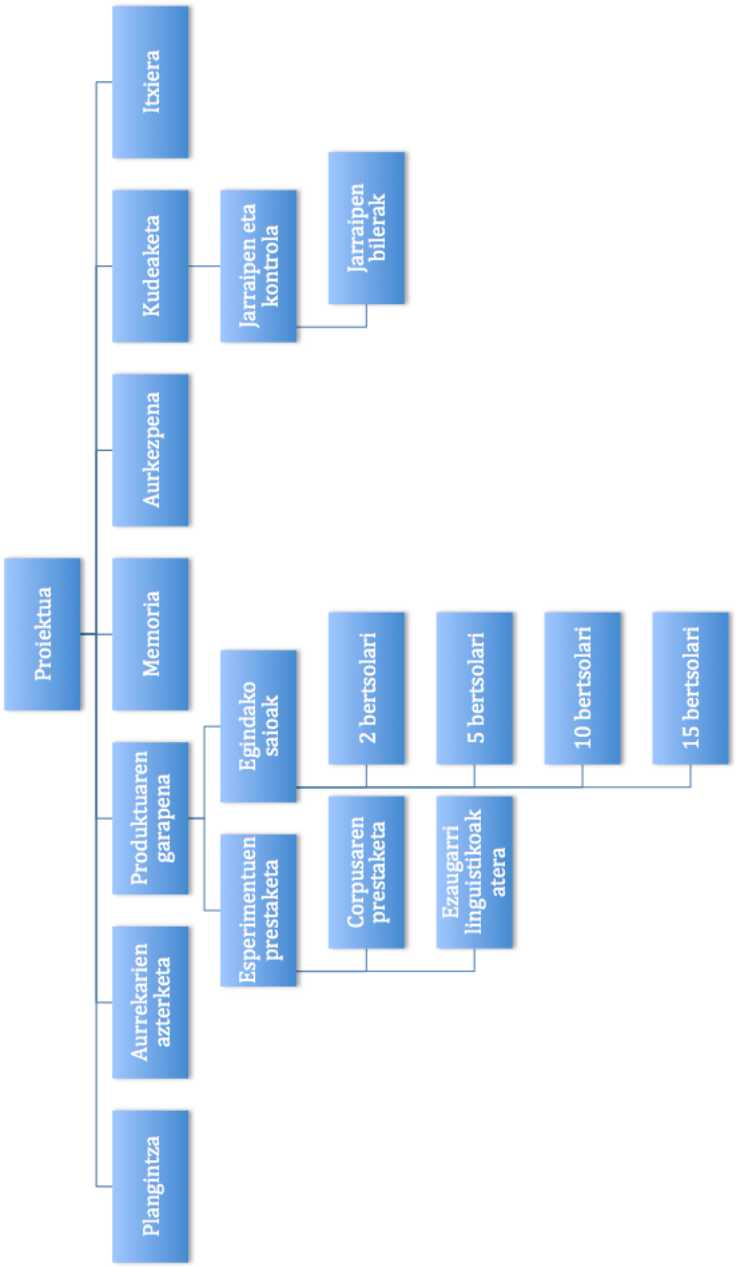
   *Train/Develop/Test* corpusetatik ezaugarri linguistikoak atera eta 4. kapituluan azalduko diren *ARFF* fitxategiak sortzeko *Perl* lengoaian idatzitako programak izango dira.

### 2.2.2 Lanaren deskonposaketa egitura (LDE)

2.1 irudian proiektuaren lanaren deskonposaketa egitura (LDE) aurkezten da. Bertan, proiektuan zehar burutu beharreko lan-karga ataza eta azpiataza garrantzitsuenen arabera erakusten da.

### 2.2.3 Proiektuaren atazak

LDE diagramaren baitan eginiko lanaren deskonposaketaren arabera, beronen burutzapenerako honako atazak zehaztu dira:




**2.1 Irudia:** *Lanaren deskonposaketa-egitura*

- **Plangintza**

  Ataza honetan proiektua aurrera eramateko plangintza garatuko da. Bertan proiektuaren helburuak zehaztu, eta hauek betetzeko atazak definituko dira. Honez gain, atazen denbora estimazioa ere egingo da. Azkenik, proiektuaren bideragarritasun eta arrisku plana ere garatuko dira.

- **Aurrekarien azterketa**

  Ataza honetan proiektua garatzeko beharrezkoa den ezagutza jasoko da. Horretarako egileen esleipen-arloan azken urteetan argitaratutako zenbait lan aztertuko dira.

- **Esperimentuen prestaketa**

  Ataza honetan esperimentuak prestatuko dira. Alde batetik, corpusa lortu eta egokitu beharko da. Bestetik, corpus hauetatik beharrezko ezaugarri linguistikoak lortzeko programak garatu beharko dira.

- **Egindako saioak**

  Ataza honetan esperimentuak burutuko dira hainbat bertsolari kopururekin, hainbat ezaugarri linguistikorekin eta hainbat ikasketa-algoritmorekin.

- **Memoria**

  Ataza honetan proiektuaren dokumentazioa idatziko da.

- **Jarraipen eta kontrola**

  Ataza honetan proiektuko helburu eta mugarri guztiak betetzen direla bermatuko da. Ustekabekorik gertatuz gero, proiektuak aurrera jarraitzen duela kontrolatuko da. Horretarako zenbait jarraipen bilera burutuko dira. Ataza hau proiektuaren bizitza-ziklo osoan zehar egongo da aktibo.

- **Aurkezpena**

  Azken ataza honetan, proiektuaren aurkezpena prestatuko da defentsarako.

Jarraian ataza eta azpiataza bakoitzari esleitutako denbora estimazioak aurkitzen dira:

| Ataza | Estimatutako denbora (orduak) |
|---|---|
| Plangintza | 25 |
| Aurrekarien azterketa | 120 |
| Produktuaren garapena | 150 |
| Esperimentuen prestaketa | 40 |
| Egindako saioak | 110 |
| Memoria | 80 |
| Aurkezpena | 10 |
| Jarraipen eta kontrola | 30 |
| **Guztira** | **415** |

**2.1 Taula:** *Atazen denbora estimazioa*

A.1 taulari erreparatuz gero, proiektuko denbora gehiena aurrekariak aztertu eta produktua garatzeko dedikatuko dela argi ikus daiteke. Produktuaren garapenean, inplementazioari baino denbora gehiago eskainiko zaio hainbat bertsolari kopururekin esperimentuak burutu eta emaitzak lortzeari. Jarraian, memoria, jarraipena eta plangintza datoz, hurrenez hurren. Azkenik, proiektua defendatzeko aurkezpena prestatzeari ere behar adina denbora dedikatu beharko diogu.

### 2.2.4 Mugarriak

Proiektua modu arrakastatsuan aurrera eramateko zenbait mugarri identifikatu dira:

#### 2.2.4.1 Barne mugarriak

- **2014/02/28 - Aurrekariak aztertzeko epea**

  Egun honetarako egile-esleipenean azken urteetan egin diren aurrerapenak aztertuko ditugu.

- **2014/03/14 - Aurrekarietatik gure lanerako ondorioak atera**

  Aurrekariak aztertu ondoren, egun honetarako gure lanerako baliagarriak izango diren ideiak atera eta ondorioak bilduko ditugu.

- **2014/04/04 - Corpusa lortu eta gure lanerako egokitu**

  Egun honetarako Bertsolari Txapelketa Nagusietako bertsoen corpusa lortuko dugu eta gure beharretara egokituko dugu.

- **2014/04/18 - Lehenengo emaitzak lortu ohiko ezaugarriak erabiliz**

  Egun honetarako aurrekarietan aztertutako estiloa adierazten duten ezaugarriak erabiliz lehen emaitzak lortu beharko dira.

- **2014/05/02 - Lehenengo emaitzak lortu berariazko ezaugarriak erabiliz**

  Egun honetarako bertsolaritzako ezaugarriak erabiliz lehen emaitzak lortu beharko dira.

- **2014/05/30 - Hainbat proba hainbat bertsolari kopururekin**

  Egun honetarako hainbat bertsolari kopuru erabiliz lortutako emaitzak bukatuta egon beharko dira.

- **2014/06/02 - Dokumentazioarekin hasi**

  Egun honetarako proiektuko ataza nagusiak bukatuta egon beharko dira eta dokumentazioarekin hasi beharko gara.

- **2014/06/27 - Proiektua bukatu**

  Egun honetarako proiektuko ataza guztiak bukatu beharko dira, aurkezpenari dagokiona izan ezik.

#### 2.2.4.2 Kanpo mugarriak

- **2014/06/30 - Proiektua entregatu**

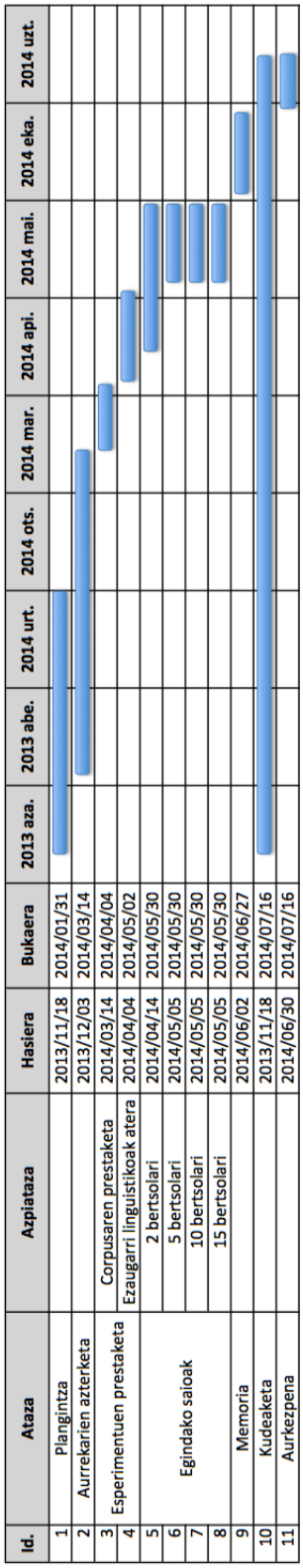
  Egun hau izango da proiektua entregatzeko azken eguna. Horregatik, egun honetarako proiektuko ataza guztiak bukatu beharko dira, aurkezpenari dagokiona izan ezik.

- **2014/07/16 - Proiektuaren defentsa**

  Egun honetarako aurkezpena prestatua egongo da.

### 2.2.5 Gantt-diagrama

Atal honetan, 2.2.3 atalean aipatutako ataza eta azpiatazek egutegian duten kokapenaren adierazpen grafikoa ikus dezakegu 2.2 irudiko Gantt-diagramaren bitartez.



| Id. | Ataza | Azpiataza | Hasiera | Bukaera |
|---|---|---|---|---|
| 1 | Plangintza | | 2013/11/18 | 2014/01/31 |
| 2 | Aurrekarien azterketa | | 2013/12/03 | 2014/03/14 |
| 3 | Esperimentuen prestaketa | Corpusaren prestaketa | 2014/03/14 | 2014/04/04 |
| 4 | | Ezaugarri linguistikoak atera | 2014/04/04 | 2014/05/02 |
| 5 | Egindako saioak | 2 bertsolari | 2014/04/14 | 2014/05/30 |
| 6 | | 5 bertsolari | 2014/05/05 | 2014/05/30 |
| 7 | | 10 bertsolari | 2014/05/05 | 2014/05/30 |
| 8 | | 15 bertsolari | 2014/05/05 | 2014/05/30 |
| 9 | Memoria | | 2014/06/02 | 2014/06/27 |
| 10 | Kudeaketa | | 2013/11/18 | 2014/07/16 |
| 11 | Aurkezpena | | 2014/06/30 | 2014/07/16 |

**2.2 Irudia:** *Estimatutako Gantt-diagrama*

Bertan azaltzen den moduan, proiektu honetan aurrekarien azterketari garrantzi handia eman zaio, berau izango baita gure lanaren oinarri. Egindako saioen garrantzia ere berebizikoa da, bertatik irtengo baitira proiektuko emaitzak eta ondorioak.

### 2.2.6 Kronograma

2.2.4 eta 2.2.5 ataletan aurkeztutako mugarriak eta Gantt-diagrama osatzeko asmoz, eta atazek elkar duten eragina erakutsi nahian, 2.3 irudiko kronograma eraiki da.

Bertan agerikoa den bezala, ataza gehienek dute hurrengo atazetan eragina. Lehenengo, aurrekariak aztertu eta gure lanerako baliagarriak izango diren ondorioak eskuratuko ditugu. Ondorio horiekin, corpus egokia sortu eta egile-esleipeneko ezaugarri linguistiko egokienak aterako ditugu. Ezaugarri hauetaz baliatuz eta aurrekarietan aztertu ditugun metodoekin, hainbat esperimentu burutuko ditugu. Azkenik, aurretik burututako ataza guztiak erabiliko ditugu memoria idazteko.

## 2.3 Lan-metodologia

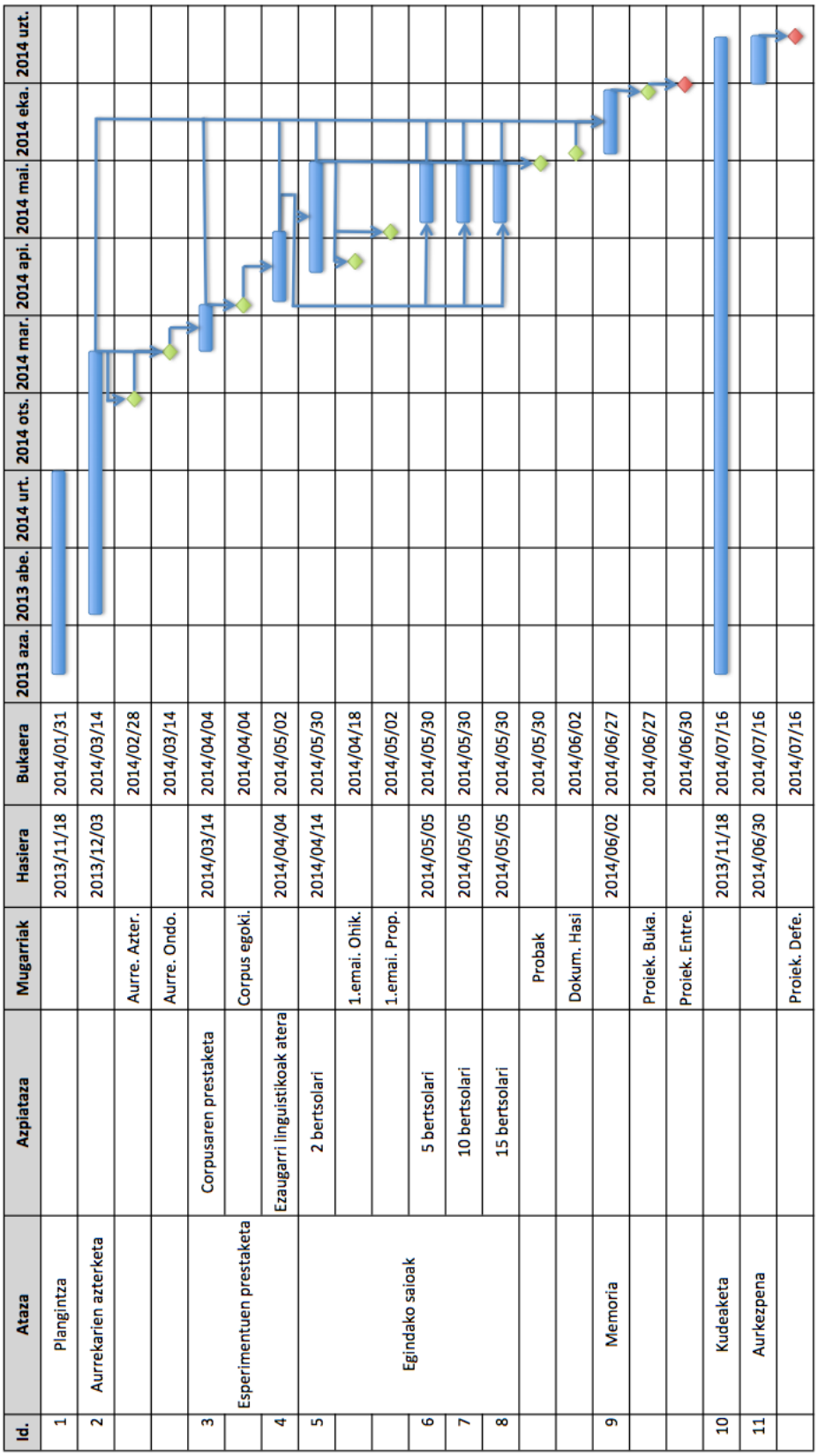
Proiektua arrakastaz garatu ahal izateko, hasieratik jarraian azalduko diren baldintzak zehaztu dira. Hala ere, proiektua aurrera joan ahala baldintza hauek aldatuz joan daitezke.

### 2.3.1 Bilerak

Barne mugarri bat bukatu orduko bilera egingo da, asteazkenetan goizeko hamarretan. Bilera hauek bi ordu ingurukoak izatea estimatzen da. Hala ere, ikasle edo tutorearen ezusteko bat dela eta, hitzorduak aldaketak jasan ditzake. Aldaketa hauek posta elektronikoz komunikatuko dira. Bertan, orduraino egindako atazei buruzko jarraipena egingo da, eta baita jarraian egin beharko diren atazei buruzko hausnarketa ere.

### 2.3.2 Planifikatutako ordutegiak

Planifikatutako epe eta ordutegiak behar bezala errespetatzen saiatu beharko da. Hala ere, aipatu beharra dago proiektua garatu bitartean ikaslea graduko bi irakasgaitan eskolak jasotzen ari dela, eta horiei lehentasuna emango zaiela.




**2.3 Irudia:** *Estimatutako kronograma*

### 2.3.3 Prestakuntza

Proiektuaren lehenengo fasea, ikaslea egile esleipenaren arloan murgiltzera, eta horri buruz ikasi eta prestatzera bideratuta dago. Horrela, tutoreak proposaturiko zenbait aurrekari aztertuko dira, eta horietatik ikasitakoa bertsolari esleipenera ekartzen saiatuko da.

### 2.3.4 Garapena

Proiektuaren bigarren fasean, lehenengoan ikasitakoa praktikan jartzea izango da helburu. Horrela corpusa egokitzeko eta corpusetik bertsokera edo bertsolarien estiloa adierazten duten ezaugarriak ateratzeko programak garatu beharko dira. Azkenik, txapelketetako bertsoen egileak zein diren asmatuko dituen sailkatzailea garatu, eta azken proba batzuk egin beharko dira sistemaren zehaztasuna neurtzeko.

## 2.4 Bideragarritasuna

Proiektua aurrera eramateko beharrezkoak diren baldintzak aztertu ondoren, proiektuaren bideragarritasuna bermatzen saiatuko gara.

### 2.4.1 Baliabideen kostua

Proiektuan beharrezkoak izango diren baliabideak doakoak direla bermatu da.

### 2.4.2 Baliabideen funtzionamendu bermea

Erabiliko diren baliabideak proiektuaren garapenean prest eta atzigarri egongo direla bermatu da.

### 2.4.3 Denbora

Plangintzan proiektua aurrera eramateko nahiko denbora izanen dela bermatu da.

### 2.4.4 Komunikazioa

Ikasle eta tutorearen artean komunikazio eraginkor bat jarraitzeko asmoz, 2.3.1 azpiatalean lan-metodologia egokia finkatu da.

## 2.5 Arriskuen analisia

Proiektuaren garapenean zenbait arrisku sor daitezke, honen arrakasta baldintzatu dezaketenak. Horregatik, lehenengo lana arrisku hauek identifikatzea izango da, eta ondoren, horien aurrean hartu beharreko neurriak erabaki beharko dira.

### 2.5.1 Identifikatutako arriskuak

Ondorengoak dira proiektuan zehar sor daitezkeen arriskuak:

1. Proiektuaren garapenean zehar atazen bat burutzeko ezintasuna, denbora faltagatik, irakasgaietako lan zamagatik edo bestelako arazoengatik.
2. Proiektuko zatien galera.

### 2.5.2 Kontingentzia-plana

Aurreko azpiatalean aipatutako arriskuen aurrean hartu beharreko neurriak honakoak izango dira:

1. Proiektuaren garapenean zehar atazen bat burutu ezin denean, arazorik ez sortzeko, planifikazio malgu bat diseinatu da. Horrela, halako arazoen aurrean lan-programa egokitzeko aukera egongo da. Bestalde, arazoa oso sakona bada, proiektua irailean aurkezteko aukera ere badago.
2. Proiektuko edozein zatiren galera saihesteko, astero proiektu osoaren segurtasun kopiak egin eta hodeian gordeko dira.

# 3. KAPITULUA

# Aurrekarien azterketa

Kapitulu honetan, azken hamarkadan egile-esleipenaren arloan eginiko hainbat lan aztertuko dira, eta bertatik ikasitakoa euskarara ekartzea izango da asmoa, bertsoen mundura hain zuzen ere.

Horrela, 3.1 atalean, testuetako egile-estiloa adierazten duten ezaugarri linguistikoak aztertuko ditugu. Ondoren, 3.2 atalean, arlo honetan erabiltzen diren ikasketa-algoritmoak ikusi eta ikertuko ditugu. Honen ostean, 3.3 atalean, atributu-aukeraketaren nondik norakoa eta gure lanean izan dezaken eragina azalduko dugu. Jarraian, 3.4 atalean, egile-esleipeneko lanetan aurki daitezkeen bi esleipen-metodoak ikusiko ditugu. 3.5 atalean, egile-esleipenean kontuan hartu beharreko zenbait ideia azalduko ditugu. Azkenik, 4.4 atalean, egindako ikerketatik gure lanerako baliozkoak izango zaizkigun ideia nagusiak eta horien inguruan hartutako erabakiak bilduko ditugu.

## 3.1 Ezaugarri linguistikoak

Hizkuntza bera eta hizkuntzarekin lotura duten fenomenoak aztertzen dituen zientzia da linguistika edo hizkuntzalaritza. Zientzia honek, informatikarekin uztartuta, hizkuntzaren prozesamenduaren (HP) alorra osatzen du.

Hainbat eratako ezaugarri linguistikoak bereizten dira hizkuntzalaritzan; nagusiki, morfologikoak, sintaktikoak, semantikoak eta pragmatikoak. Jarraian, egile-esleipeneko lanetan gehien erabiltzen diren ezaugarri linguistikoak azalduko dira.

### 3.1.1 Ezaugarri lexikoak

Testu bat esaldietan elkartutako token sekuentzia modura ikus dezakegu. Horietako bakoitza hitz, zenbaki edo puntuazio ikur bati egokituko zaio.

Egile-esleipenerako, estiloa adierazteko metodo bat hiztegiaren aberastasuna –*vocabulary richness*– erabiltzea litzateke. Horren helburua testu batean erabilitako hiztegiaren aberastasuna neurtzea da. Neurketa honetarako metodo desberdinak erabiltzen diren arren, ezagunena *mota-token* ratioa kalkulatzen duen metodoa[1] da. Hiztegiaren aberastasuna *hapax legomena* delakoaren bidez ere neurtu dezakegu. Honekin, corpus batean behin bakarrik agertzen den hitz, hitz-forma edo adierazpidea jasoko dugu. Horregatik, askotan, *hapax*-a idazkera akatsa baino ez da izaten.

Hala ere, [Stamatatos, 2009]-k dioenez testuak errepresentatzeko modurik egokiena hitzen maiztasunen bektoreak erabiltzea da, hitz-poltsak –*bag of words*– deritzenak, hain zuzen ere. Hau da, testua maiztasun zehatz bat duten hitzen multzoa izango da, testuinguruaren informazioa kontuan hartu gabe. Hala ere, helburua gaiak sailkatzea bada, gehien erabiltzen diren hitzak (adizlagunak, izenordainak, determinatzaileak, lokailuak, preposizioak...) baztertu egingo dira; hau da, eduki-hitzak –*content words*– hartuko dira aintzat. Eduki-hitzak hitz ezegonkorrak direla esan ohi da, eta gaiari edo edukiari buruzko informazioa ematen digutenez, esan daiteke egileen estiloa adierazteko ez direla oso egokiak. Hitz ezegonkor bat “konparatu” litzateke adibidez, “alderatu”-rekin ordezka baitezakegu. Hala ere, [Koppel et al., 2009]-ek esaterako, beraien egile-esleipen lanean erabili dituzte, emaitza onak lortuz. Eduki-hitzak aukeratzeko baztertu ditugun hitzak egonkorrak direla esan daiteke, ezin baitira beste hitzekin ordezkatu; “eta” adibidez, ezin da beste hitz batekin ordezkatu. Horregatik, hitz egonkorrak egile-estilo adierazle onak dira. Hitz egonkor hauei ez-funtsezko hitzak –*function words*– deritze. Ondorioz, askoz ere hitz gutxiago beharko ditugu ezaugarri lexikoak erabiliz testu bati egilea esleitzeko, gaiaren arabera sailkatzeko baino. Ez-funtsezko hitzak egileek oharkabean erabiltzen dituzte, eta gaiarekiko independenteak dira. Hartara, egileen estilo-ezaugarriak hauteman daitezke horietatik. [Stamatatos, 2009]-en arabera, ezaugarriak aukeratzerakoan, maiztasuna egonkortasun eza baino eraginkorragoa da. Hala ere, bien konbinaketak emaitza hobeak ematen ditu.

Hitz-poltsak erabiltzea soluzio sinple eta eraginkorra izan arren, ez dute testuinguruko informazioa kontuan hartzen. Hori arazo bat da, hitz batek testuinguruaren arabera

[1] *Mota-Token* = *V/N*, non *V* egileak erabilitako hiztegiaren luzera den eta *N* testuan agertutako token desberdinen kopurua.

esanahi desberdinak eduki baititzake. Horregatik, testuinguruko informazioa kontuan hartzeko n-gramak erabiltzen dira (n ondoko hitzak). Hala ere, n-gramak erabiltzeak ez du beti emaitza hobetzen, zenbait arazo sortzen baitira:

- Hitzen arteko konbinazio guztiak kontuan hartzeak problemaren tamaina nabarmen handitzen du.
- Hitzen arteko konbinazio gehienak ez dira testuan aurkitzen, eta horregatik algoritmo sailkatzaileak erabiltzea zaila da.
- Estiloaren informazioa baino gehiago edukiaren informazioa jasotzen da.

Aipatutako ezaugarriez gain, egileen estiloa definitzeko, egiten dituzten akatsez balia gaitezke: akats ortografikoak, formatuko akatsak... Ezaugarri hauek zuzentzaile ortografikoa erabiliz lor ditzakegu, adibidez. Hala ere, hizkuntza askotarako zuzentzaile ortografikoak lortzea arazoa izan daiteke.

Horrelako metodoak edozein hizkuntzatan eta edozein corpusetan aplika ditzakegu.

### 3.1.2 Karaktere-ezaugarriak

Testu bat karaktere sekuentzia baten modura ere ikus dezakegu. Horrela, karaktere alfabetikoak, zenbakiak, hizki xehe eta larriak, puntuazio ikurrak... hartuko ditugu kontuan. Informazio hau erraz atzi daiteke edozein hizkuntza natural edo corpusetik, eta oso erabilgarria izango da egileen estiloa zehazteko.

Egile-esleipenerako estiloa adierazteko metodo bat karaktere mailako n-gramak erabiltzea izango litzateke (n ondoko karaktereak). Hauek informazio ugari eskaini dezakete: informazio lexikoa, testuinguruko informazioa, puntuazio-ikur eta letra larrien erabilera, testuetako akatsak... Hala ere, batzuetan, informazio erredundantea eskainiko digute (|ze_| eta |_ze| 3-gramek, adibidez), posible baita bi n-grama edo gehiagok informazio bera eskaintzea. Hitzekin bezala, maiztasun handieneko n-grama karaktereak dira estiloa zehazteko garaian ezaugarri garrantzitsuenak. N izanik n-gramaren luzera, n handi batek, adierazpen edo errepresentazioaren tamaina handitzeaz gain, informazio lexikoaren eta testuinguruaren informazio handia eskainiko digu. Bestalde, n txiki batek (2 edo 3) azpi-hitzen edo silaben informazioa emango digu, baina ez testuinguruarena. N on bat aukeratzea hizkuntzaren araberakoa izango da, batzuetan besteetan baino hitz luzeagoak baitaude. [Stamatatos, 2009]-en lanean ikus daitekeenez, lantzen duen problemarako n

onena 4 da ingelesean. Hitz luzeagoak erabiltzen diren hizkuntzetan n handiagoa erabili beharko genuke. Oso eraginkorrak dira egile-esleipeneko problemetan eta arrakasta handia izan dute [Keselj et al., 2003]-en, [Koppel et al., 2009]-en eta [Grieve, 2007]-ren lanetan, adibidez.

### 3.1.3 Ezaugarri sintaktikoak

Testu bat adierazteko edo errepresentatzeko beste metodo landuago bat informazio sintaktikoa erabiltzea da. Egileek, ohartu gabe, antzeko patroi sintaktikoak erabili ohi dituzte. Horregatik, [Stamatatos, 2009]-en iritziz, informazio lexikoarekin alderatuta, informazio sintaktikoa egilearen hatz-markatzat har daiteke. Bestalde, estiloa zehaztean ez-funtsezko hitzen arrakastak informazio sintaktikoaren erabilgarritasuna erakusten du, normalean egitura sintaktiko askotan aurki baitaitezke. Testuetatik informazio sintaktikoa jasotzeko hizkuntzaren prozesamenduko tresna trinko eta zehatzak beharrezkoak dira. Prozesu hori hizkuntzaren araberakoa izango da, eta nahi dugun hizkuntzarako analizatzaile sintaktiko zehatz baten beharra edukiko dugu. Ezaugarri hauek, interpretatzailearen akatsak direla eta, zarata edukiko dute. Hala ere, ezaugarri sintaktikoak ezaugarri lexikoak baino konplexuagoak direnez, egile-esleipeneko lanetan ez dira asko erabiltzen.

### 3.1.4 Berariazko ezaugarriak

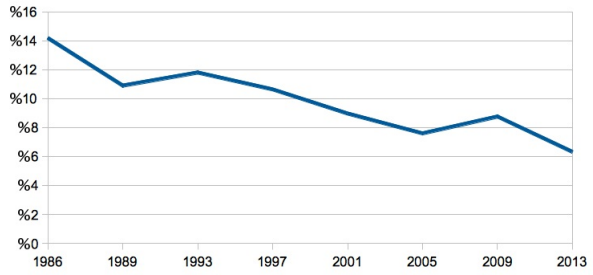
Ezaugarri lexiko, sintaktiko edo karaktere-ezaugarriak aplikazioarekiko independenteak dira, testutik atera baitaitezke hizkuntzaren prozesamenduko tresna egokien bidez. Estiloaren ñabardurak hobeki adierazteko problema edota hizkuntza konkretu bakoitzerako berariazko neurriak defini ditzakegu.

Iazko irailean, Bertsolari Txapelketa Nagusia hastearekin batera, *IXA* taldeak[2] Elhuyar aldizkarian argitaratutako artikulu[3] baten ([Agirrezabal et al., 2013]) eta urtarrilean *Bertsozale*[4]-n eginiko analisi estatistikoaren arabera, 3.1 irudian ikus daitekeen moduan, urteak pasa ahala txapelketetan euskara batuaren erabilera gero eta handiagoa da.

[2] *IXA* taldea Euskal Herriko Unibertsitateko ikerkuntza-taldea da, hizkuntzaren tratamendu automatikoan lan egiten duena.

[3] http://aldizkaria.elhuyar.org/gai-librean/bota-bertsoa-eta-guk-aztertuko-dugu/

[4] http://www.bertsozale.com/eu/bertsolari-txapelketa-nagusia-2013/albisteak/1986-2013-arteko-bertsolari-txapelketa-nagusien-analisi-estatistikoa




**3.1 Irudia:** *Euskara batuan ez dauden euskarazko hitzen ehunekoa*

Hala ere, euskara batuan erabiltzen ez diren hitzek bertsokeraren berri eman dezaketela uste dugu. Horretarako, hitz ez-estandarrez baliatuko gara. Horrelako hitzek bertsolariek erabilitako euskalkien, laburduren... informazioa eskainiko digute. Horregatik estilo-adierazle egokiak izan daitezke.

Bestalde, bertsoen ezaugarriak aipatzen baditugu, errima ezin dugu ahaztu, askorentzat bertsoen alderdi teknikoen ardatza baita. Diotenez, errimatuz ari bagara (nahiz eta errima guztiz aberatsa ez izan), bertsoa osatzen ari gara. Horrexegatik kontuan hartzeko ezaugarria izango da errima, bertsokeraren adierazle egokia izan daitekeena.

Bertsolaritza munduan, kontuan hartzeko ezaugarriak dira oinak ere. Oina puntuko[5] azken hitza da, errima duen hitza. Bertso batean oin guztiak desberdinak izan behar dira, hitz bera (forma eta esanahi berdinez) bi oinetan ipiniz gero, poto egiten baitu bertsolariak, eta hau akats itsusitzat hartzen baita. Hala ere, oinetako hitzek formaz berdinak izan arren adiera bana badute, ez da potorik egingo. Ikus dezakegunez, ezaugarri hau ere bertsokeraren adierazle izan daiteke.

## 3.2 Ikasketa-algoritmoak

Gizakiok esperientziatik ikasteko gaitasuna daukagu. Bizitzen ditugun esperientziak, ikusten ditugun adibideak... gure garunean “ereduak” sortzeko erabiltzen ditugu, eta eredu hauek arrazonatzeko erabil ditzakegu.

Ikasketa automatikoaren helburua, besteak beste, gure ikasteko gaitasuna imitatzea da. Proiektuaren helburua bertsoak egilearen arabera sailkatzea da. Horretarako, gainbegiratutako sailkatzaileak erabiliko ditugu, zehazki, gizakiok gauzak sailkatzeko dugun gaitasuna simulatzen saiatzen baitira.

[5] Puntua errimatik errimara doan bertso zatia da.

Jarraian, [Arrieta, 2010]-ren doktoretza tesian oinarrituz, proiektuan erabilitako ikasketa-algoritmoak azalduko ditugu: *Naive Bayes*, erabaki-zuhaitzak, *K-NN* eta *support vector machines* edo *sostengu-bektoreen makinak* (*SVM*). *Naive Bayes* algoritmoa erabili dugu, ikasketa-algoritmo sinpleenetako bat delako; erabaki-zuhaitzak, berriz, adibide askorekin portaera ona erakusten dutelako; *K-NN*-z baliatu gara [Luyckx and Daelemans, 2008]-ek lortu zituzten emaitza onak direla eta; *support vector machines* algoritmoa erabili dugu, atributu askorekin portaera ona erakutsi izan duelako ([Diederich et al., 2003] eta [Koppel et al., 2009]-en lanetan, adibidez), eta gaur egun HPan puri-purian dagoen ikasketa-algoritmoa delako.

### 3.2.1 *Naive Bayes*

Sailkatzaile estokastiko[6] sinpleena da, baina ikasketa automatikoaren arloan eta HPan arrakastaz erabili izan dena. Bayes sailkatzailearen instantzia konkretu honek, izan ere, ezaugarri asko eraginkortasunez konbinatzeko gaitasuna du. Gainera, ondo dabil sarrerako datuen multzoa oso handia denean.

Probabilitatearen banakako banaketan oinarritzen da. Adibide baten klasea asmatzeko, behatutako adibidearen probabilitatea maximizatzen duena aukeratzen da. Horretarako, Bayes-en teorematik eratorritako formula sinple bat erabiltzen da, non atributu guztiei dagozkien balioak emanik $(a_1, a_2, \ldots a_n)$ emaitza-atributuaren klase probableena $(V_{nb})$ aukeratzen baita:

$$V_{nb} = argmax_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

*Naive Bayes* sailkatzaileak hipotesi batekin jokatzen du: atazaren deskribapenerako erabilitako atributu edo ezaugarri bakoitza beste edozein bezain garrantzitsua dela; alegia, independenteak direla atributu guztiak. Hipotesi hau, ordea, ez da betetzen askotan, baina, hala eta guztiz ere, baldintza hori betetzen dela suposatzeak dakarren sinplifikazioak eredu dotore eta eraginkorrak eman ohi ditu.

Eskuratutako ezagutza ezin denez modu ulergarri batean adierazi, ikasketa azpisinbolikoko algoritmoa dela esaten da.

[6] Atazan inplikatutako atributuen dependentzia probabilistikoak deskribatzen dituzte eredu estokastikoek, graforen baten bidez normalean. Grafoko adabegi bakoitzak zorizko aldagai bat adierazten du eta probabilitate-banaketa bat du esleituta. Banakako banaketa hauen bidez, behatutako adibide guztien baterako banaketa kalkula daiteke.

*Naive Bayes* algoritmoaren erabileraren adibide argigarriak daude Internet sarean[7].

HPan, besteak beste, eginkizun hauetarako erabili izan da: testuinguruaren araberako zuzenketa ortografikoa, morfosintaxiaren etiketatzea, preposizio-sintagmen desanbiguazioa, desanbiguazio semantikoa eta dokumentuen sailkapena.

### 3.2.2 Erabaki-zuhaitzak

Ikasketa automatikoko eskema klasiko hau *zatitu eta irabazi* teknikan oinarritzen da, eta grafikoki adierazita, zuhaitz baten itxura hartzen du; hortik datorkio izena. Erabaki-zuhaitzak sortzeko prozesua modu errekurtsiboan azal daiteke. Lehendabizi, atributu edo ezaugarri bat aukeratu behar da erro-adabegian kokatzeko, eta bere balio posible bakoitzeko adar bat egiten da. Gero, prozesua errepika daiteke errekurtsiboki, adar bakoitzerako, baina adar bakoitzeko baldintzak bete dituzten adibideekin soilik. Adabegi-ume bakoitzeko adibide guztiek sailkapen bera dutenean amaitzen da prozesua, kasu horretan ezingo baita adabegia gehiagotan banatu; adabegi hori, beraz, hostoa izango da. Erabaki beharreko gauza bakarra, eskema honetan, zera da: une bakoitzean aukeratu beharreko atributua. Unean uneko atributuaren aukeraketak, ordea, berebiziko garrantzia dauka, behin atributu hori erabili eta gero ez baita gerora hartuko diren erabakietan atributu bera berriz erabiltzen. Bestalde, geroz eta atributu gehiago izan, orduan eta denbora gehiago beharko du ikasketa-algoritmoak.

HPko ia maila guztietan erabili izan dira erabaki-zuhaitzak. Emaitza onak lortu izan dira, esaterako, ahotsaren ezagutzan, morfosintaxiaren etiketatzean, desanbiguazio semantikoan, analisi sintaktikoan, laburpenen sorkuntzan, entitateen ezagutzan, dokumentuen sailkapenean eta itzulpen automatikoan.

### 3.2.3 *K-NN*

Distantzien kalkuluan oinarritzen den sailkatzailea da. Ideia sinplea eta intuitiboa da. Kasu berri bat sailkatzerakoan bere hurbileneko K auzokideen klaseen artean sarrien agertzen den klasea egokituko zaio. Hurbileneko auzokideak zeintzuk diren jakiteko distantziak neurtu beharko dira, eta horretarako zenbait metodo daude. Hala ere, lan honetan distantzia euklidestarrak erabiliko ditugu sailkatu nahi den instantziaren atributuek entrenamendurako instantzien atributuekiko duten distantziak neurtzeko.

[7] http://www.inf.u-szeged.hu/∼ormandi/teaching/mi2/ webgunean *naive bayes* atala aukeratuta, edota http://www.statsoft.com/textbook/stnaiveb.html webgunean.

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

K-ren balioa aukeratzerakoan, kontuan hartu beharko da K txiki bat ez ohiko puntuen edo puntu zaratatsuen aurrean oso sentikorra izango dela. Bestalde, K handi bat aukeratuz gero, ereduak klaserik handiena esleitzeko joera dauka.

*K-NN* sailkatzaileak HPko hainbat mailatan erabili arren, dokumentuen sailkapenean izan dute arrakasta gehien.

### 3.2.4 *Support Vector Machines* edo *sostengu-bektoreen makinak*

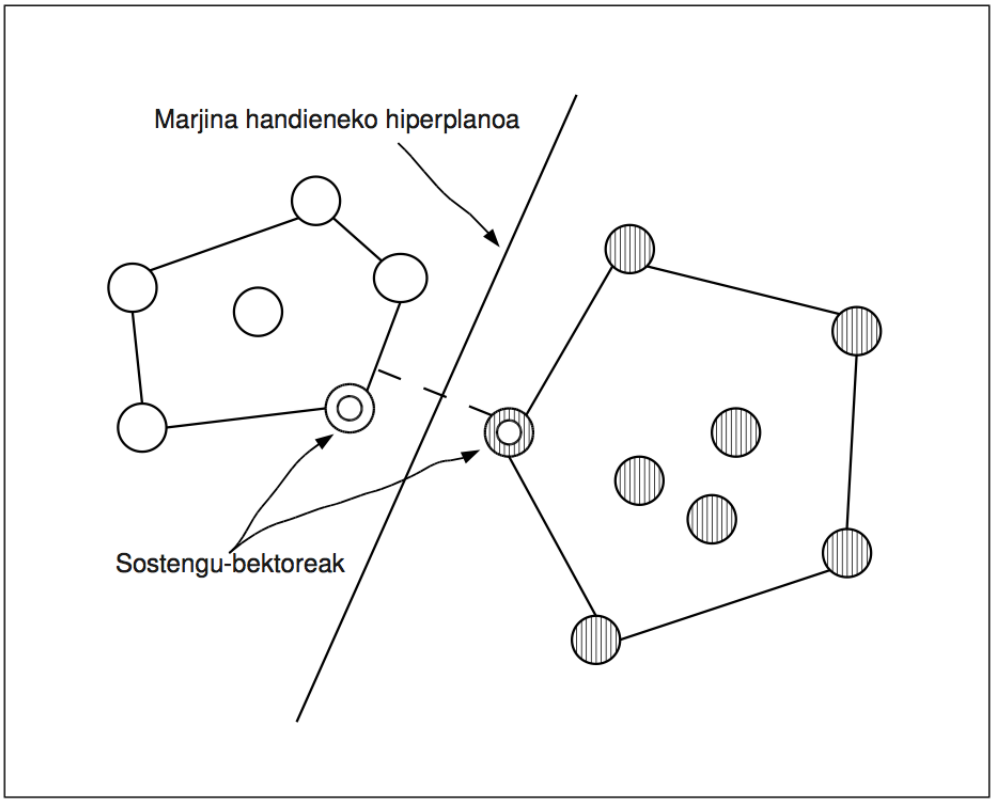
Ikasketa automatikoko algoritmo honek eredu linealek dituzten desabantailak konpontzen ditu. Izan ere, linealak ez diren datu multzoentzat soluzio bat ematen du. Bere forma sinpleenean, ordea, eredu linealetan oinarritzen da, marjina handieneko hiperplanoa[8] deitzen zaion eredu lineal berezi bat baliatzen baitu. Hainbat atazatan erabiltzen da eredu lineal hau.

Har dezagun, adibidez, bi klaseko datu multzo bat, zeina linealki banagarria den; beste modu batean esanda, bada hiperplano bat —zuzen bat, alegia— instantzien espazioan, zeinak instantzia guztiak sailkatzen dituen zuzenaren alde batera eta bestera. 3.2 irudian argiago uler dezakegu azaltzen ari garena. Kontuan izan marraz beteak dauden zirkuluak klase batekoak direla; hutsak, beste klasekoak.

Marjina handieneko hiperplanoa bi klaseen artean banaketa handiena ematen diguna da; hots, hiperplanoko alde bateko eta besteko instantziak elkarrengandik urrutien jartzen dituena. Hiperplanotik gertuen dauden instantziei *support vector* deritze (*sostengu-bektore*). Gutxienez, *sostengu-bektore* bana dago klase bakoitzeko. Marjina handieneko hiperplanoa eskuratu ondoren eta bi klaseetako sostengu-bektoreak izanik, gainerako ikasketa-instantzia guztiak baztergarriak lirateke.

Esan dugun moduan, ikasketarako datuak linealki banagarriak ez diren kasuetarako ere orokortu daiteke eskema hau, *kernel* deituriko funtzioen bitartez. Sarrerako atributuen espazioa dimentsio handiagoko espazio batean bilaka daiteke, eta hori sailkatzaile polinomikoen edo hiru geruzako neurona-sareen bidez adierazia geratzen da azkenean. Praktikan, goi-muga antzeko bat markatzen duen parametro bat kalkulatzea da gakoa, eta horretarako, esperimentuak egitea beste aukerarik ez dago.

[8] *maximum margin hyperplane*.




**3.2 Irudia:** *Marjina handieneko hiperplano bat, eta dagozkion sostengu-bektoreak*

Amaitzeko, esan beharra dago *support vector machines (SVM)* izeneko ikasketa automatikoko eskema hau ez dela batere azkarra, adibide asko dituen ikasketa corpusekin lan egiterakoan, batik bat. Gainera, ez da sinbolikoa; beraz, ezin da eskuratutako ezagutza gizakiarentzat ulergarria den adierazpide batera ekarri. Hala eta guztiz ere, emaitza onak lortzen ditu oro har, erabaki-muga konplexuak eta finak eskuratzen dituelako, eta portaera bereziki ona dauka atributu askoko atazetan, eta baita linealki banagarriak ez diren problemetan ere.

Hala ere, lan honetan eredu lineal gisa soilik erabili dugu, *Weka* paketearen inplementazioan.

Patroien identifikazioarekin zerikusia duten hainbat arlotan erabilia izan da ikasketa automatikoko eskema hau; hala nola, bioinformatikan. HParen baitan, ataza hauetan erabilia izan da arrakastaz, besteak beste: azaleko sintaxiaren analisi automatikoan eta dokumentuen sailkapenean.

## 3.3 Atributu-aukeraketa

Atributuen aukeraketa ikasketa automatikoan askotan erabiltzen den prozesua da. Bertan, atributu guztietatik azpimultzo bat aukeratzen da, ondoren ikasketa-algoritmo bat aplikatzeko. Datu-meatzaritzaren eta ikasketa automatikoaren arloan ohikoa da datu edo informazio kopuru handiak edukitzea, eta hori atributu-aukeraketarekin maneia dezakegu. Atributuek eskaintzen duten informazioa adierazgarria den jakitea da prozesu honen helburua. Atributuen aukeraketa prozesuarekin, instantziaren atributuen kopurua gutxitzen da, informazio garrantzigabea, erredundantea edo kaltegarria duten atributuak baztertuz. Hartara, atributu-aukeraketak hainbat abantaila eskaintzen dizkigu; hala nola, sailkatzaileen zehaztasuna hobetzea, konputazio denbora jaistea, kostua txikitzea... Horregatik, proiektu honetako esperimentuetan atributu-aukeraketa ere aplikatuko dugu.

Atributu-aukeraketan bi atal bereizten dira:

- **Bilaketa**. Atributuen espazioan, ezaugarri multzo egokiena bilatu behar da. Hainbat aukera dauden arren, lan honetan atributuak zein ordenetan eta zenbat atributu aztertzen ditugun hartuko dugu kontuan –*ranker* aukera–.

- **Ebaluazioa**. Atributu azpimultzo horren kalitate edo doitasuna ebaluatzeko metodo ezberdinak erabiltzen dira: informazio-irabazia, elkar-informazioa, *chi-square*... Lan honetan, [Koppel et al., 2009]-ek egiten duten moduan, atributuek eskaintzen diguten informazio-irabaziaren arabera egingo da atributu-aukeraketako ebaluazioa.

Egileak esleitzeko ikerketetan erabiltzen diren ezaugarri multzoek ezaugarri bat baino gehiago konbinatu ohi dituzte. Ezaugarri lexiko eta karaktere-ezaugarriek, adibidez, ezaugarri-multzoen tamaina nabarmen handitzen dute. Kasu horietan, atributuen edo ezaugarrien aukeraketa aplika daiteke errepresentazio edo adierazpen tamaina txikitzeko.

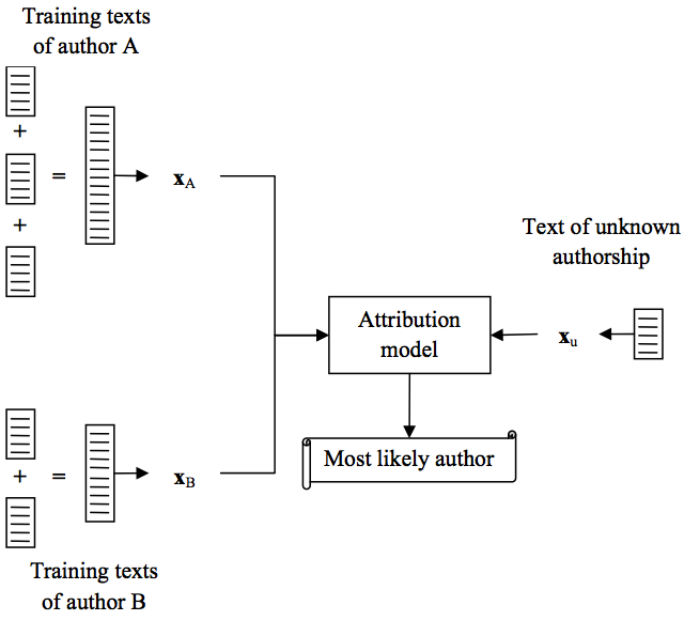
## 3.4 Esleipen-metodoak

Esleipen-metodoei dagokienez, [Stamatatos, 2009]-ek profilean oinarritutako metodoak eta instantzian oinarritutakoak aipatzen ditu. Profilean oinarritutako metodoan, helburua egile beraren testu guztiak hartu eta honen estilo orokorra definitzea izango da. Horretarako, 3.3 irudian ikus dezakegun bezala, entrenamendurako testuak egileen arabera katea-

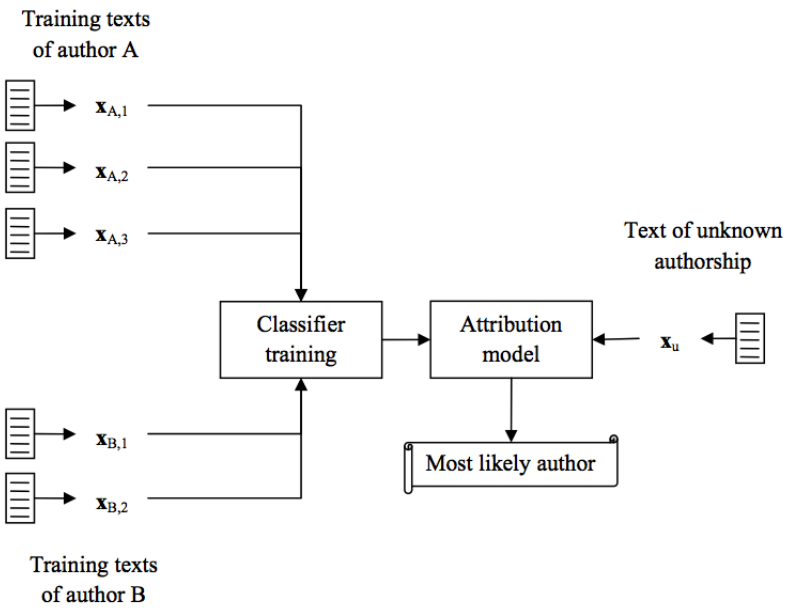
tzen dira. Horren abantaila nagusia zera da, entrenamendurako testuak motzak direnean, kateamenduak adierazpen fidagarriagoak sortzen dituela.

Instantzian oinarritutako metodoan, berriz, entrenamendurako testu bakoitzak sailkapen ereduari bakarka laguntzen dio; hau da, sailkatzaile-algoritmoa egile ezagunen instantzia multzoak erabiliz entrenatuko da esleipen-eredu bat garatzeko. Algoritmo mota hauek, modelo fidagarri bat sortzeko, klase bakoitzeko zenbait instantzia behar dituzte. Egile baten testu bakarra badugu eta testua nahiko luzea bada, luzera berdineko zenbait zatitan banatu beharko dugu. Egile bereko eta luzera ezberdineko zenbait testu dituGunean, bestalde, entrenamendurako testu-instantziaren luzera normalizatu beharko dugu.

Instantzian oinarritutako metodoek ikasketa automatikoko algoritmo potenteak erabiltzen dituzte, dimentsio altuko datuak, zaratatsuak eta sakabanatuak maneiatzeko gai direnak (3.2.4 azpiatalean aurki dezakegun *SVM*, adibidez). Kasu honetan, errazagoa da adierazpen edo errepresentazioetan estilo-ezaugarri ezberdinak konbinatzea. Profilean oinarritutako metodoetan hori zailagoa da, antzekotasunean oinarritutako metodoak erabiltzen baitituzte (3.2.1 azpiatalean aurki dezakegun Naive Bayes, adibidez), eta normalean ezaugarri multzo homogeneoak maneiatzen baitituzte. Gainera, profilean oinarritutako metodoan sinadura eta agurra moduko estilo-ezaugarri batzuk ezin dira erraz erabili, profila egilearen estiloaren propietate orokorrak adierazten saiatzen da eta.




**3.3 Irudia:** *Profilean oinarritutako metodoaren ohiko arkitektura*




**3.4 Irudia:** *Instantzian oinarritutako metodoaren ohiko arkitektura*

3.4 irudian ikus daitekeen moduan, instantzian oinarritutako metodoetan entrenamendurako fasea daukagu, instantzia bakoitzetik estilo bat ikasten duena. Profilean oinarritutako metodoan, berriz, 3.3 irudian errepara daitekeenez, kateamenduari esker entrenamendu fasea sinpleagoa da. Aztertutako lanetan ikus dezakegunez, egile-esleipenean instantzian oinarritutako metodoa da gehien erabiltzen dena.

## 3.5 Ebaluazioa

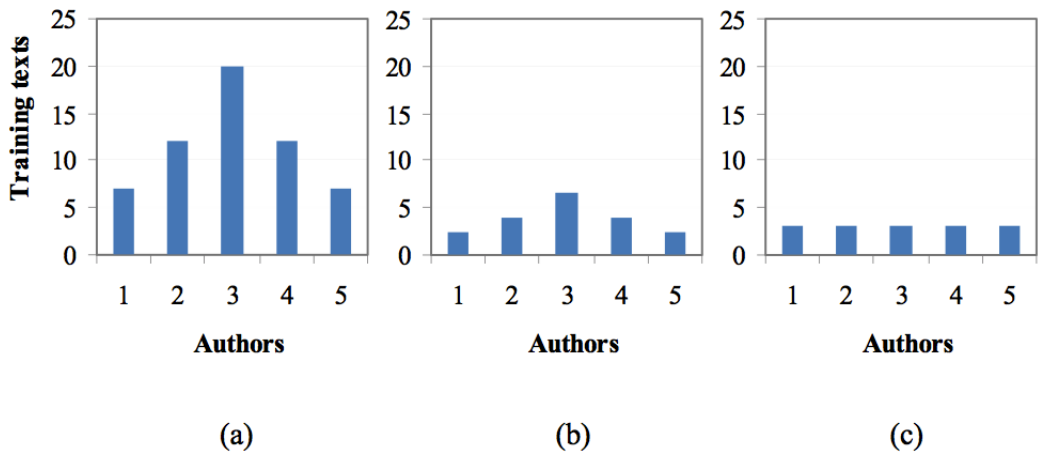
Ebaluazioaz hitz egitean, egileen esleipeneko sistemak ebaluatzerakoan aurki ditzakegun parametroek izan behar duten balioak aztertzeaz ezin gara ahaztu. Horrelako lanetan hainbat erabaki hartu behar dira; hala nola, entrenamendurako corpusen tamaina, *Develop* corpusaren tamaina, *Test* corpusaren tamaina, corpus bakoitzean egingo diren adibideen aukeraketa...

[Stamatatos, 2009]-en iritziz, egoera ideala autoreak ongi definituta edukitzea litzateke, bakoitza testu multzo batekin. Honez gain, egilea ezkutatuta duten testuen multzo bat ere eduki beharko genuke, garatutako sistema ebaluatzeko.

Entrenamendurako corpusaren tamainari dagokionez, zenbait zalantza sortzen dira. Zein da testuaren luzera egokia estilo-ezaugarriak behar bezala adierazteko? [Stamatatos, 2009]-

ek aztertutako ikasketa batzuek testu motzekin (<1000 hitz) emaitza onak lortzen dituzte. [Luyckx and Daelemans, 2008]-ek, bestalde, beren lanean 1.400 hitzeko saiakerak erabiltzen dituzte, eta [Grieve, 2007]-k 40.000 hitz inguruko testuekin egiten du lan. Eskura dugun corpus handiena erabiltzea izaten da joera nagusia, eta lortutako emaitzen arabera corpusa handitu ala ez erabakitzea (betiere, corpusa handitzeak dakarren kostua kontuan harturik).

Testuen kopurua zenbatekoa izan behar den ere ez dago argi. [Koppel et al., 2009]-ek 250 hitzetako 217-745 post bitarte erabiltzen dituzte, adibidez, eta [Yu, 2012]-k 85 saiakera. Argi dagoena da testuen banaketak orekatua izan behar duela. Entrenamendurako corpusen banaketa desorekatua denean (3.5 irudiko (a) grafikoa), garapenerako eta testerako banaketa egokiak lortzeko metodo ezberdinak daude. Aukera bat, 3.5 irudiko (b) grafikoan ikus daitekeen moduan, aurreko banaketa imitatzea izango litzateke. Metodo hau testuak gaiaren arabera sailkatzerakoan izaten da egokiago. Beste aukera, 3.5 irudiko (c) grafikoan adierazten den moduan, banaketa orekatzea izango litzateke. [Stamatatos, 2009]-en iritziz, metodo hau egile-esleipenerako egokiago izan ohi da.




**3.5 Irudia:** *5 autoreen Train, Develop eta Test testuen banaketa modu desberdinak*

Egile kopuruari dagokionez, berriz, normala den bezala, gero eta egile gehiago eduki, problemaren zailtasuna handiagoa izango da. Horrela, [Grieve, 2007]-k bere probetan lortutako emaitzetan ikus dezakegun moduan, egile kopurua handitu ahala sistemaren zehaztasuna txikiagoa da.

Bestalde, [Stamatatos, 2009]-en iritziz, testu guztiak genero berekoak, antzeko garaikoak eta gai berdintsukoak izatea egokiago litzateke. Honez gain, hizkuntza prozesatzeko

dugun gaitasuna ere hartu behar da kontuan ebaluazioan. Azkenik, esleipen-metodo baten gaitasuna zehazteko, genero ezberdinetan probatu beharko litzateke.

## 3.6 Ondorioak

Aurrekariak aztertu eta gero, gure lanerako baliagarri izango zaiguna bilduko dugu jarraian. Lehenengo lana, corpusa nolakoa izango den zehaztea da. 3.5 atalean ikusi dugun moduan, testuen luzera eta kopuruari buruz oraindik ere gauzak ez daude batere argi. Hala ere, testu guztiak genero berekoak, antzeko garaikoak eta gai berdintsukoa izan beharko dutela jakin badakigu. Gure kasuan, 1989. urtetik aurrerako Bertsolari Txapelketa Nagusiko bertsoak aukeratuko ditugu. Hala ere, gaiari dagokionez ezin dugu inongo iragazketarik egin, corpuseko bertsoen kopurua nabarmen jaitsiko bailitzateke. Bertso kopuruari dagokionez, berriz, ahalik eta bertso kopuru gehien edukitzen saiatuko gara bertso gehien dituzten bertsolariak aukeratuz. Dakigunez, zenbat eta bertsolari gutxiago aukeratu, orduan eta emaitza hobeak lortuko ditugu. Hortaz jabetzeko eta bertsolari kopuruak duen eragina ikusteko, 2, 5, 10 eta 15 bertsolarirekin egingo ditugu probak. Testuen banaketak orekatua izan behar duela ere azaldu dugu. Hortaz, aukeratutako bertsolariek bertso kopuru bera edukiko dute.

3.4 atalean ikusi dugun moduan, bi esleipen-metodo bereizten diren arren, dudarik gabe egile-esleipeneko lanetan gehien erabiltzen dena instantzian oinarritutako metodoa da, eta lan honetan ere hau erabiltzea erabaki da.

Honen ostean, ezaugarri linguistikoei buruzko erabakiak hartzeko txanda da. Aurrekarien azterketan ezaugarri lexiko eta karaktere-ezaugarriak dira gehien erabiltzen direnak. Emaitza onenak eman dituztenak, eduki-hitzak, ez-funtsezko hitzak, hitz mailako n-gramak eta karaktere-mailako n-gramak dira. Hala ere, erabili beharreko eduki-hitzen eta ez-funtsezko hitzen kopuruan ez dago adostasunik. [Stamatatos, 2009]-en ustez, 150tik 675ra bitarteko ez funtsezko hitzen kopurua erabiltzea da egokiena; [Koppel et al., 2009]-en ustez, 512; eta [Yu, 2012]-n ustez, 35. Hori dela eta, hainbat proba egingo ditugu gehien erabiltzen diren 50, 100, 150, 200, 250 eta 300 ez-funtsezko hitzekin. Gehienez 300 ez-funtsezko hitz erabiltzea erabaki dugu bi bertsolarirekin egindako probetan 347 ez-funtsezko hitz desberdin besterik ez baitituzte erabiltzen.

Eduki-hitzen kasuan [Koppel et al., 2009]-ek 1.000 erabiltzen dituzte. Kasu honetan ere, kopuruak eduki dezakeen eragina ikusteko, hainbat proba egingo ditugu 250, 500,

750, 1.000, 1.250, 1.500, 1.750 eta 2.000 eduki-hitzekin, bi bertsolarirekin osatutako corpusean 2.087 eduki-hitz desberdin besterik ez baitituzte erabiltzen.

Hitz-mailako n-gramen kasuan, n-ren balio handiak ez dira erabiltzen, n handitu ahala problemaren tamaina izugarri handitzen baita. 20.000 hitz ezberdineko corpusa edukiz gero, adibidez, 2-gramekin *20.000 x 19.999 = 400* milioi parametro ditugu, 4-gramak erabiliz $20.000^2 \times 19.999 = 1{,}6 \times 10^{17}$... Ondorioz, gure probetan n-k 2, 3 eta 4 balioak hartuko ditu.

Karaktere mailako n-grametan, berriz, 4.2.2.2 azpiatalean ikusi dugunez, n on bat aukeratzea ez da lan erraza. Horregatik, n ugari erabiliko ditugu gure probetan, 2, 3, 4, 5, 6 eta 7 balioak dituztenak, hain zuzen ere.

Bestalde, 3.1.3 azpiatalean ikusi dugunez, ezaugarri sintaktikoak ez dira horrelako lanetan erabiltzen, eta ondorioz, gure lanean ere ez ditugu erabiliko.

Berariazko ezaugarriak deiturikoei dagokienez, bertsoen izaera berezia dela eta, lan honetan berebiziko garrantzia eduki dezakete. Horregatik, hitz ez-estandarrak, errimak eta oinak ere erabiliko ditugu gure esperimentuetan. Bi bertsolarirekin egindako probetan 346 hitz ez-estandar erabiltzen dituztenez, esperimentuetan hainbat proba egingo ditugu bertsolariek gehien erabiltzen dituzten 50, 100, 150, 200, 250, 300 eta 350 hitz ez-estandarrekin.

Ikasketa-algoritmoen kasuan, 3.2 atalean aztertutakoak aukeratu ditugu, horiek baitira egile-esleipenean gehien erabiltzen diren algoritmoak.

Honez gain, [Koppel et al., 2009]-ek beren lanean atributu-aukeraketa ere erabiltzen dute. Eduki-hitzak eta karaktere mailako 3-gramak erabiltzean, adibidez, gehien erabilitako 10.000etatik informazio-irabazi gehien ematen duten 1.000ak erabiltzen dituzte. Horregatik, 3.3 atalean azaltzen den bezala, gure esperimentuetan informazio-irabazi gehien ematen diguten atributuak erabiliz ere egingo ditugu probak.

Azkenik, garatutako sailkatzaileek emandako emaitzak nola aurkeztu ere azaldu beharko da. Aurrekarietan zehaztasuna erabiltzen da, hau da, zuzen sailkatutako dokumentuen ehunekoa. Gure lanean ere, ildo horretatik jarraituz, zuzen sailkatutako bertsoen ehunekoa emango dugu emaitza gisa.

# 4. KAPITULUA

# Proiektuaren garapena

Kapitulu honetan, 3. kapituluan aztertu eta erabakitakoarekin, gure lanaren garapenaren nondik norakoa azalduko dugu.

Horrela, 4.1 atalean, proiektuan zehar erabilitako teknologia ezberdinak ikusiko ditugu. Honen ostean, 4.2 atalean, garatutako esperimentuen prestaketa azalduko dugu. Jarraian, 4.3 atalean bertsolari kopuru ezberdinak erabiliz egindako saioen emaitzak ikusiko ditugu. Azkenik, 4.4 atalean, egindako saioetatik ateratako ondorioak bilduko ditugu.

## 4.1 Teknologiaren aukeraketa

Atal honetan, corpusa prestatzeko, hortik ezaugarri linguistikoak atzitzeko, ikasketa-algoritmoak aplikatzeko, emaitzekin lan egiteko eta memoria idazteko aukeratutako teknologiak azalduko dira.

Corpusa prestatzeko eta ezaugarri linguistikoak atzitzeko, *XML* fitxategiekin lan egingo dugu. *XML* datuak irakurtzeko moduan biltegiratzeko erabiltzen den markaketa-lengoaia da. Informazioa modu egituratuan eskaintzen digu etiketak erabiliz. Lengoaia hedagarri moduan sailkatuta dago, erabiltzaileok bere elementuak sortzeko aukera baitugu. Horrela, *XML* fitxategiekin modu errazean egin daiteke lan. Horretarako, *PERL* lengoaiaren alde egin da. Baina zergatik *PERL*? Testu fitxategiak aztertu eta manipulatzeko berezi-ki sorturiko programazio-lengoaia da, oso egokia hizkuntzarekin erlazionatutako atazetarako: adierazpen erregularren bitartez testuetan bilaketak eta aldaketak egin daitezke,

fitxategiak lerroz lerro irakurtzeko gai da, karaktere eta hitz terminoak maneiatzen ditu... Honez gain, gure lanerako baliagarriak zaizkigun *XML::Simple*, *Text::Ngrams* eta *Text::Ngram* moduluak eskaintzen dizkigu, hurrenez hurren *XML* fitxategiekin, hitz mailako n-gramekin eta karaktere mailako n-gramekin lan egiteko erabili direnak. Gainera, software librea da, doakoa eta ikasteko erraza.

*Naive Bayes*, erabaki-zuhaitzak, *K-NN* eta *support vector machines* edo *sostengu-bektoreen makinak* (*SVM*) ikasketa-algoritmoekin lan egiteko *Weka*[1] softwarea erabili da. Software honek abantaila ugari eskaintzen dizkigu: lan egiteko interfazea dauka, lan konplexuak erraz egiteko aukera ematen digu, librea da, *Java*-z inplementatuta dago... Honez gain, informazio-irabazia kontuan hartuz atributuak aukeratzeko aukera ere ematen digu. Gainera, kontsola bidez ere erabili daiteke, eta horrela esperimentuak automatizatu ditzakegu. *Weka* erabiltzeko, *PERL*-ez ezaugarri linguistikoak atera eta *ARFF* fitxategiak sortuko ditugu. *ARFF* fitxategia *ASCII* testu fitxategi bat da, atributu berdinak dituzten instantzia zerrenda bat deskribatzen duena.

*Weka* erabiliz lortuko ditugun emaitzak *CSV* fitxategietan gorde ditzakegu. *CSV* fitxategiak formatu irekiko dokumentuak dira, datuak taula modura adierazteko egokiak. Ondoren, fitxategi hauek *R* lengoaiarekin maneiatuko ditugu, horrelako datuekin lan egiteko ezin hobea baita, eta taulak eta grafikoak erraz sortzeko aukera eskaintzen baitigu.

Azkenik, proiektuko lana azalduko duen memoria idazteko *LaTeX* erabiliko da. *LaTeX* testuak osatzeko sistema da, tesiak, artikulua eta liburu teknikoak idazteko sarritan erabiltzen dena.

## 4.2 Esperimentuen prestaketa

Atal honetan, corpusaren prestaketa, bertsolarien estiloa adierazteko erabiliko ditugun ezaugarri linguistikoak nola lortuko ditugun eta sailkatzailea garatzeko ikasketa-algoritmoak zeintzuk izango diren azalduko ditugu.

### 4.2.1 Corpusaren prestaketa

Proiektua garatzen hasteko lehenengo lana, 1989. urtetik aurrera egindako Bertsolari Txapelketa Nagusietako bertsoen corpusak lortzea izango da. Corpus hauek, Xenpelar Doku-

[1] Datu-meatzaritzako atazetarako erabiltzen diren ikasketa automatikoko algoritmoen multzoa da WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

mentazio Zentroak[2] gure eskura jarri zituenak, *IXA* taldeko zerbitzarian[3] aurki daitezke. Bertan, corpusekin lan egiteko hainbat aukera ditugu:

- Urteka banatutako corpusak, *corpusaXXXX.xml* izenarekin.
- Urteka banatutako corpusak analizatuta, *corpusaAnalizatutaXXXX.xml* izenarekin. Aurrekoaren berdina baina morfologikoki eta sintaktikoki analizatuta.
- Urte guztietako corpusa morfologikoki eta sintaktikoki analizatuta, *corpusaAnalizatutaOsoa.xml* izenarekin.
- Urteka banatutako corpusak morfologikoki eta sintaktikoki analizatuta, eta gainera hitzak estandarrak edo ez-estandarrak diren analizatuta, *corpusaAnalizatutaEzestXXXX.xml* izenarekin.
- *Train/Develop/Test* corpusak, *corpusaAnalizatuta[Train|Develop|Test].xml* izenarekin. Corpus osoa hiru ausazko zatitan banatuta, %80 *Train*, %10 *Test* eta %10 *Develop*.

Dakigunez, lan honetarako bertsoak morfologikoki analizatuta egoteaz gain, hitz bat estandarra edo ez-estandarra den ere jakitea komeni zaigu. Horregatik, *corpusaAnalizatutaEzestXXXX.xml* corpusekin lan egingo dugu, 1989, 1993, 1997, 2001, 2005 eta 2009 urtekoekin hain zuzen ere, 2013koa ez baitago eskuragarri. Jarraian, corpusen egitura ikus dezakegu:

```
<Bertsoak>
  <Bertsoa trzenb="1">
    <Kategoria>BAT-BATEKOA</Kategoria>
    <Saioa>1993-12-19 DONOSTIA TXAPELKETA</Saioa>
    <Gai-jartzailea/>
    <edo-Gai-emailea>AZKUNE-LAZARO</edo-Gai-emailea>
    <edo-Aurkezlea/>
```

[2] 1991ean Bertsozale Elkarteak sortua, bertsolaritzaren ondarea bildu, antolatu eta gizarteratuz, maila guztietako ikerkuntza bultzatzeko helburuarekin.

[3] Corpusak */sc01a4/users/maguirrezaba008/corpusak/corpusakXml* karpetan ditugu.

```
<Bertsolariak>PEÑAGARIKANO-ANJEL-MARI</Bertsolariak>
<Gaia>Hitza gaia: Danbolina</Gaia>
<Neurria>BEDERATZI PUNTUKOA</Neurria>
<Doinua>GAZTALONDO HANDIAN</Doinua>
<Lana>BAKARKA GAIA-EMANDA HITZA</Lana>
<balorazioa-saioarena/>
<Bertsoaldiari-oharrak/>
<Deskriptoreak> //  </Deskriptoreak>
<Bertso-gaiburuak>70-35 DANBOLINA</Bertso-gaiburuak>
<Gai-oharrak/>
<Argitaratua-Argitalpena>Bertsolari Txapelketa Nagusia 1993
</Argitaratua-Argitalpena>
<Funtsa-Bilduma>Txapelketetako Bilduma</Funtsa-Bilduma>
<Iturria>Xenpelar Dokumentazio Zentroa</Iturria>
<Kokapena/>
<Transkribatzailea/>
<Transkripzio-oharrak/>
<Izenbururako/>
<Bertso-kopurua>1</Bertso-kopurua>
<Transkripzioa>
  <lg zenb="1" iden="0">
    <l>
      <w wid="1" lemma="zaku" ezest="0" morpha="IZE ARR BIZ- ABS
       NUMS MUGM HAS_MAI" syntch="%SINT" syntfu="@OBJ">Zakua</w>
      <w wid="2" lemma="hartu" ezest="0" morpha="ADI SIN PART
       NOTDEK" syntch="%ADIKAT" syntfu="@-JADNAG">hartu</w>
      <w wid="3" lemma="eta" ezest="0" morpha="LOT JNT EMEN AORG"
       syntch="" syntfu="@PJ">eta</w>
    </l>
    <l>
      ...
    </l>
    ...
  </lg>
```

```
      </Transkripzioa>
    </Bertsoa>
    ...
  </Bertsoak>
.....................................................................
```

Ikus dezakegun moduan, corpusetan 1989, 1993, 1997, 2001, 2005 eta 2009 urteko bertsoak bertsoka egituratzen dira. Bertso bakoitzean, bertsoari buruzko informazio ugari aurki dezakegu: kategoria, saioa, gai-jartzailea, bertsolaria, gaia, neurria... Horien artean transkripzioa daukagu, bertso bakarraz edo hainbat bertsoz osatua egon daitekeena. Bertsoak lerroz osatuak daude, eta lerroak analizatutako hitzez.

Corpus horiek lortu ondoren, gure beharretara egokitu beharko ditugu. Aipatutako urte guztietako corpusekin bakarra sortzea izango da lehenengo lana. Ondoren, corpus horretatik abiatuz, bertsoak bertsolarika sailkatuko ditugu, gure lanerako egokiagoa izango baita. Bestalde, gure probak hainbat bertsolari kopuru erabiliz egingo ditugunez, nahi ditugun bertsolarien bertsoak bakarrik beharko ditugu. Horrela, ondorengo egitura duen corpusa osatuko dugu:

```
.....................................................................
  <?xml version='1.0' encoding='ISO-8859-1'?>
  <Bertsoak>
    <Bertsolaria id="1">
      <Izena>ITURRIAGA-UNAI</Izena>
      <Transkripzioa>
        <lg>
          <l>
            <w lemma="batzuk" morpha="DET DZG NMGP DES MG HAS_MAI"
             syntch="%SINT" syntfu="@ADLG" wid="1">Batzurentzat</w>
            <w lemma="gogor" morpha="ADJ ARR IZAUR- ABS NUMS MUGM"
             syntch="" syntfu="@OBJ" wid="2">gogorra</w>
          </l>
          ...
        </lg>
        ...
```

```
      </Transkripzioa>
    </Bertsolaria>
    ...
  </Bertsoak>
```

Ikus dezakegunez eta arestian aipatu bezala, gure corpusa bertsolarika egongo da sailkatuta. Bertsolari bakoitzeko, izena eta transkripzioa bakarrik dira erabiliko ditugun eremuak. Transkripzioan, dagokion bertsolariaren bertso guztiak aurkituko ditugu. Bertsoak, lehen bezala, lerroz egongo dira osatuak, eta lerroak analizatutako hitzez.

Sortu berri dugun corpus honetatik abiatuz, hurrengo pausoa *Train/Develop/Test* banaketa ausaz osatzea izango da. *Train* corpusa bertsoen %80arekin egongo da osatuta, eta honekin bertsolarien bertsokera edo estiloa ikasiko dugu. *Develop* corpusa bertsoen %10arekin osatuko da, eta honekin, ahalik eta sailkatzaile onena garatzen saiatuko gara emaitzarik onenak bilduz. Azkenik, *Test* corpusa ere bertsoen %10arekin egongo da osatuta, eta honekin, sailkatzaile onenak garatuta ditugunean, azken probak egingo ditugu sistemaren amaierako zehaztasuna zein den ikusteko.

### 4.2.2 Ezaugarri linguistikoak

Egileen esleipen-lanetan egilearen esleipena egiten saiatzeko hainbat ezaugarri linguistiko erabili ohi dira. Helburua, ezaugarri horiek erabiliz, bertsolari bakoitzaren bertsokera adieraztea izango da, eta horrekin bertso bat errepresentatzea. Horretarako, lan honetan, hiru multzotan sailkatu ditugu ezaugarri hauek: bertsoen ezaugarri lexikoak, karaktere-ezaugarriak eta berariazko ezaugarriak deiturikoak aztertu ditugu.

#### 4.2.2.1 Ezaugarri lexikoak

Lan honetan, bertsoak token segida gisa ikus ditzakegu. Horietako bakoitza hitz, zenbaki edo puntuazio ikur bati egokituko zaio.

Esan bezala, corpusean token horietako bakoitza morfologikoki aztertuta dago. Horrela, lematizatzailea erabiliz, ez-funtsezko hitzak eta eduki-hitzak aztertzeko beharrezkoak ditugun lemak dauzkagu. Honez gain, euskarazko ez-funtsezko hitzen eta eduki-hitzen zerrendak ere baditugu. Lehenengoa, euskarazko adizlagun, lokailu, determinatzaile, PRT

(partikulak), BST (bestelakoak) eta izenordainez osatua egongo da. Bigarrena, aldiz, testu bati esanahia ematen dioten hitzez dago osatua; hala nola, izen, adjektibo, adberbio, aditz... Horrela, bertsoetan hitz hauek aurkitzeko, bertsoetako hitz bakoitza aipatutako zerrendetan azaltzen den ikusi beharko dugu. Aztertzen ari garen hitza dagokion zerrendan azaltzen ez bada, hitzaren lemarekin gauza bera egingo dugu. Hau dena abiapuntutzat hartuz, bertsolarien estiloa adierazi nahian, *hash* taulak erabiliz bertsolariek erabiltzen dituzten ez-funtsezko hitzak eta eduki-hitzak kontatzen dira. Ondoren, gehien erabilitakoak aukeratzen dira, eta azkenik, bertso bakoitzean hitz kopuruko erabilitako ez-funtsezko hitz eta eduki-hitzak kalkulatzen dira. Kalkulatutako balio hauekin, bertso bakoitza zenbaki errealez eraikitako bektorearekin errepresentatuko dugu. Lan honetan, zenbait esperimentu egingo dira bertsolariek gehien erabiltzen dituzten 50, 100, 150, 200, 250 eta 300 ez-funtsezko hitzekin, eta gehien erabiltzen dituzten 250, 500, 750, 1.000, 1.250, 1.500, 1.750 eta 2.000 eduki-hitzekin.

Ezaugarri hauez gain, hitzen n-gramak ere aztertuko ditugu. Horretarako, ez dugu inolako bertsoen analisirik beharko. N-grama bakoitza ditugun n-grama guztiekiko zenbatetan errepikatzen den kalkulatuko dugu. Kalkulatutako balio hauekin, bertso bakoitza zenbaki errealez eraikitako bektorearekin errepresentatuko dugu. Lan honetan, n-k 2, 3 eta 4 balioak hartuz egingo ditugu probak.

#### 4.2.2.2 Karaktere-ezaugarriak

Karaktere mailako ezaugarriei dagokienez, lan honetan karaktereen n-gramak erabiliko dira. Hitz mailako n-gramekin egin dugun moduan, n-grama bakoitza ditugun n-grama guztiekiko zenbatetan errepikatzen den kalkulatuko dugu. Kalkulatutako balio hauekin, bertso bakoitza zenbaki errealez eraikitako bektorearekin errepresentatuko dugu. Esperimentuetan n-k 2, 3, 4, 5, 6 eta 7 balioak hartuko ditu.

#### 4.2.2.3 Berariazko ezaugarriak

Arestian aipatu bezala, bertsoak token segida gisa ikus ditzakegu, eta horietako bakoitza hitz, zenbaki edo puntuazio ikur bati egokituko zaio.

Dakigunez, corpusean token horietako bakoitza hitz estandarra edo ez-estandarra den markatuta dago. Hitza ez-estandarra bada, ondorengo adibidean ikusiko dugun moduan, bigarren lema bat edukiko du, hitz ez-estandar horri dagokiona.

```
.....................................................................
  <w wid="15" lemma="adina" ezest="1" lemmaezest="aina" morpha="ADB
   ARR ZERO AORG" syntch="%SINT" syntfu="@ADLG">ainako</w>
.....................................................................
```

Helburua bertsolariek erabilitako hitz ez-estandarrak zeintzuk diren jakitea da. Horretarako, hasiera batean, bertsoetan azaldutako hitz ez-estandarren lemak kontatzen ziren. Hala ere, aditz ez-estandarren berezitasuna dela eta, kasu horretan lema kontatu beharrean aditza bera kontatzen genuen. “Ainako” hitz ez-estandarraren kasuan, adibidez, “aina” kontatzen genuen, eta “leike”-ren kasuan, aditza denez, “*edin” beharrean “leike” bera. Baina emaitzak nahi bezain onak izan ez zirenez, hainbat proba egin ziren. Horrela, emaitza onenak hitza ez-estandarra den kasuetan hitza bera kontatuz lortzen dira, bigarren lema kontuan hartu gabe. Azkenik, bertso bakoitzean hitz kopuruko hitz ez-estandarrak zenbatetan agertu diren kalkulatuko dugu. Kalkulatutako balio hauekin, bertso bakoitza zenbaki errealez eraikitako bektorearekin errepresentatuko dugu.

Hitz ez-estandarrez gain, bertsolarien estiloa adierazi nahian, errimak ere kontuan hartu ditugu. Horretarako, bertsoen errima jakin behar izan dugu. Oin batetik errima lortzeko, *foma*[4]-n idatzitako transduktore bat erabili da, *IXA* taldeak gure esku jarri duena. Honek, oin bat emanda zenbait errima itzultzen ditu. “Patroi” hitza emanda adibidez, “aPTKBDGRoi”, “oi” eta “BDGRoi” errimak itzuliko dizkigu. Dakigun moduan, “eBDGRa” errimak eba/eda/ega/era hartzen ditu barnean, *bodegero* legearen arabera ontzat jotakoak; horren arabera, “arreba”, “hobe da”, “ernega” eta “tankera” hitzek, adibidez, errimatu egiten dute. Berdintsu gertatzen da “ePTKa”, “eNMa” eta antzeko errimekin. Lan honetarako, errimarik luzeena hartuko dugu kontuan, aberatsena hori izango baita. Bertso gehienetan oina lerro bikoitietako azken hitza den arren, hori ez da beti horrela. Horregatik, gure corpusean bertso bateko oina zein den jakiteko, bertso bakoitzeko lerro guztietako azken hitzen errimak aterako ditugu, eta bertsoaren errima, ateratako errimetatik gehien errepikatzen dena izango da. Errimekin bertsoa errepresentatzeko, zenbaki bitarrez osatutako bektore bat erabiliko dugu. Posizio bakoitza errima bati egokituko zaio. Bertso batean erabiltzen den errimak 1 balioa hartuko du eta gainontzekoek 0.

Errima lortuz gero, oinak ere lor ditzakegu, errimatzen duten hitzak izango baitira. Horrela, bertso batean erabilitako oinak zeintzuk diren jakin ditzakegu, eta informazio hori gure esperimentuetarako erabili. Oinekin bertsoa errepresentatzeko, zenbaki osoez

[4] *Foma* konpilatzailea, programazio lengoaia eta C liburutegia da, egoera finituko automatak eta erabilera desberdinetako transduktoreak eraikitzeko erabiltzen dena.

osatutako bektore bat erabiliko dugu. Posizio bakoitza oin bati egokituko zaio, oinaren agerpen kopuruaren balioarekin.

### 4.2.3 Ikasketa-algoritmoak, instantziak eta ikasi beharrekoa

Erabiliko ditugun sailkatzaile automatikoek, ikasteko abiapuntutzat *Train* corpusetik ateratako 4.2.2.1 atalean azaldutako estilo-adierazleak izango dituzte. Ondoren, *Train*-etik ikasitakoarekin, *Develop* eta *Test* corpusetako bertsoei dagokien bertsolariak esleitzen saiatuko da. 3. kapituluan erabaki dugunez, *Weka* paketeko ondorengo ikasketa-algoritmoen inplementazioak erabiliko ditugu: *Naive Bayes*, erabaki-zuhaitzak *J48 algoritmoa*, *K-NN* eta *Support Vector Machine* (*SMO*).

Lan honetan, bertso bakoitza instantzia bat izango da, aipatutako ezaugarri linguistikoen kopuruekin eraikitako zenbakien bektore batez osatuta egongo dena, eta ikasi beharrekoa bertsoaren egilea izango da.

## 4.3 Egindako saioak

Bertsolari-esleipen eredu on bat sortzeko asmoarekin, hainbat bertsolari kopururekin, hainbat ezaugarri linguistikorekin eta hainbat ikasketa-algoritmorekin egindako saioak aurki daitezke jarraian.

### 4.3.1 2 bertsolari (I)

Bi bertsolarirekin egingo dugun lehen esperimentu honetan lortutako emaitzak hurrengo esperimentuetarako oinarri izango dira; hau da, lehen proba hauetan garatutako sailkatzaile onenak izango dira gainontzeko esperimentuetan erabiliko ditugunak.

Lehenengo lana, bertso gehien dituzten bi bertsolarirekin corpusa sortu eta *Train/Develop/Test* banaketa osatzea izango da. Bertsolariak hauek izango dira:

- Andoni Egaña, 127 bertsorekin.
- Unai Iturriaga, 110 bertsorekin.

Bi bertsolari hauen 110 bertsoekin corpus berri bat sortuko dugu, bertsoak bertsolarien arabera sailkatuz:

- Andoni Egaña: 110 bertso eta 4.974 token.
- Unai Iturriaga: 110 bertso eta 5.070 token.

Ondoren, *Train/Develop/Test* banaketa egingo da ausaz:

- Andoni Egaña: *Train*-erako 88 bertso (3.991 token), *Develop*-erako 11 bertso (436 token) eta *Test*-erako 11 bertso (547 token).
- Unai Iturriaga: *Train*-erako 88 bertso (4.016 token), *Develop*-erako 11 bertso (501 token) eta *Test*-erako 11 bertso (553 token).

Corpusa prestatu eta gero, lehenengo esperimentuak ezaugarri lexikoekin burutuko ditugu. Ondoren, karaktere-ezaugarriak erabiliz egingo ditugu probak. Jarraian, berariazko ezaugarriekin lortutako emaitzak ikusiko ditugu. Proba hauetan atributu-aukeraketa ere aplikatuko da, 3. kapituluan azaldutako bilaketa-prozesuan atributu kopuru desberdinak kontuan hartuz. Honen ostean, emaitza onenak eman dituzten ezaugarri linguistikoak konbinatuko ditugu, bakarka lortutako emaitzak hobetu nahian. Aipatutako esperimentu hauek, *Train* corpusarekin entrenatuz eta *Develop* corpusarekin probatuz egingo dira. Lortutako emaitzak, zuzen sailkatutako bertsoen ehunekoak izango dira. Azkenik, lortutako baldintza onenekin *Test* corpusean probak egin eta azken emaitzak lortuko ditugu.

#### 4.3.1.1 Ezaugarri lexikoak

4.1, 4.2, 4.3, 4.4 eta 4.5 tauletan dauzkagu ezaugarri lexikoak erabiliz eta arestian aipatutako ikasketa-algoritmoekin lortutako emaitzak. Emaitzak hobetu nahian, eduki-hitzekin eta ez-funtsezko hitzekin atributu-aukeraketa aplikatzen dugu bilaketa-prozesuan atributu kopuru desberdinak kontuan hartuz. Hitz mailako n-gramek ematen dituzten emaitza kaxkarrak direla eta, ezaugarri linguistiko honi ez diogu atributu-aukeraketarik aplikatu.

Emaitzei erreparatuta, agerikoa da eduki-hitzak direla *SMO* algoritmoarekin bertsokera adierazteko ezaugarri linguistiko egokienak, zenbait probetan %95.45 eta %90.91ko emaitzak lortzen baititugu. Ikus dezakegunez, atributu-aukeraketarik gabe 500 eta 1.000 eduki-hitz erabilita lortzen ditugu emaitzarik onenak. Atributu-aukeraketarekin, berriz, 500 eta 1.000 eduki-hitzekin eta bilaketa-prozesuan atributu guztiak kontuan hartuz atributu-aukeraketarik gabe lortutako emaitza berak ematen dizkigu. Honez gain, atributu-aukeraketako bilaketa-prozesuan atributuen %50a kontuan hartuz, 1.000 eta 1.250 eduki-hitzek %90.91ko zehaztasuna eskaintzen digute, eta 1.500 eduki-hitzek %95.45ekoa.

| Ezaugarri lexikoak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 250 eduki-hitz | 59.09 | 54.55 | 50 | 63.64 | 63.64 | 81.82 |
| | 500 eduki-hitz | 54.55 | 50 | 54.55 | 72.73 | 81.82 | **95.45** |
| | 750 eduki-hitz | 45.45 | 45.45 | 45.45 | 63.64 | 81.82 | 86.36 |
| | 1.000 eduki-hitz | 45.45 | 50 | 54.55 | 63.64 | 81.82 | **90.91** |
| | 1.250 eduki-hitz | 50 | 50 | 40.91 | 63.64 | 81.82 | 86.36 |
| | 1.500 eduki-hitz | 59.09 | 63.64 | 63.64 | 63.64 | 81.82 | 86.36 |
| | 1.750 eduki-hitz | 45.45 | 50 | 45.45 | 77.27 | 81.82 | 86.36 |
| | 2.000 eduki-hitz | 63.64 | 59.09 | 59.09 | 77.27 | 81.82 | 86.36 |
| | 250 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 59.09 | 54.55 | 50 | 72.73 | 63.64 | 81.82 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 54.55 | 50 | 54.55 | 72.73 | 81.82 | **95.45** |
| | 750 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 45.45 | 45.45 | 45.45 | 72.73 | 81.82 | 86.36 |
| | 1.000 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 45.45 | 50 | 54.55 | 72.73 | 81.82 | **90.91** |
| | 1.250 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 50 | 50 | 40.91 | 77.27 | 81.82 | 86.36 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 59.09 | 63.64 | 63.64 | 72.73 | 81.82 | 86.36 |
| | 1.750 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 45.45 | 50 | 45.45 | 72.73 | 81.82 | 86.36 |
| | 2.000 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 59.09 | 59.09 | 63.64 | 86.36 | 81.82 | 86.36 |
| | 250 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 50 | 40.91 | 50 | 72.73 | 68.18 | 86.36 |

**4.1 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri lexikoak *erabiliz lortutako emaitzak (I)*

| Ezaugarri lexikoak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 50 | 50 | 50 | 72.73 | 72.73 | **90.91** |
| | 750 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 45.45 | 45.45 | 50 | 77.27 | 72.73 | 86.36 |
| | 1.000 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 54.55 | 50 | 45.45 | 68.18 | 72.73 | 86.36 |
| | 1.250 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 50 | 45.45 | 54.55 | 68.18 | 72.73 | 81.82 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 81.82 | 72.73 | 45.45 | 68.18 | 72.73 | 81.82 |
| | 1.750 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 50 | 59.09 | 54.55 | 72.73 | 72.73 | 81.82 |
| | 2.000 eduki-hitz eta bilaketa-prozesuan atributuen %75 | 59.09 | 63.64 | 54.55 | 86.36 | 72.73 | 72.73 |
| | 250 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 63.64 | 81.82 | 68.18 | 81.82 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 54.55 | 50 | 45.45 | 77.27 | 72.73 | 81.82 |
| | 750 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 54.55 | 54.55 | 54.55 | 68.18 | 77.27 | 86.38 |
| | 1.000 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 50 | 50 | 50 | 77.27 | 77.27 | **90.91** |
| | 1.250 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 50 | 50 | 54.55 | 68.18 | 77.27 | **90.91** |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 59.09 | 63.64 | 68.18 | 77.27 | **95.45** |

**4.2 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri lexikoak *erabiliz lortutako emaitzak (II)*

| Ezaugarri lexikoak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 1.750 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 54.55 | 59.09 | 72.73 | 77.27 | 86.36 |
| | 2.000 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 54.55 | 50 | 77.27 | 77.27 | 86.36 |
| | 250 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 63.64 | 63.64 | 68.18 | 81.82 | 63.64 | 77.27 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 54.55 | 63.64 | 63.64 | 81.82 | 72.73 | 72.73 |
| | 750 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 72.73 | 68.18 | 68.18 | 86.36 | 68.18 | 72.73 |
| | 1.000 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 50 | 22.73 | 50 | 86.36 | 72.73 | 77.27 |
| | 1.250 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 31.82 | 45.45 | 40.19 | 72.73 | 68.18 | 77.27 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 50 | 50 | 59.09 | 72.73 | 68.18 | 81.82 |
| | 1.750 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 54.55 | 59.09 | 63.64 | 68.18 | 68.18 | 81.82 |
| | 2.000 eduki-hitz eta bilaketa-prozesuan atributuen %25 | 59.09 | 54.55 | 63.64 | 68.18 | 63.64 | 72.73 |
| Ez-funtsezko hitzak | 50 ez-funtsezko hitz | 68.18 | 63.64 | 63.64 | **86.36** | 68.18 | 54.55 |
| | 100 ez-funtsezko hitz | 54.55 | 50 | 54.55 | 77.27 | 72.73 | 72.73 |
| | 150 ez-funtsezko hitz | 45.45 | 59.09 | 59.09 | 77.27 | 72.73 | 77.27 |
| | 200 ez-funtsezko hitz | 40.91 | 50 | 54.55 | 77.27 | 72.73 | 72.73 |
| | 250 ez-funtsezko hitz | 54.55 | 59.09 | 68.18 | 77.27 | 72.73 | 72.73 |
| | 300 ez-funtsezko hitz | 68.18 | 68.18 | 77.27 | 77.27 | 72.73 | 72.73 |

**4.3 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri lexikoak *erabiliz lortutako emaitzak (III)*

| Ezaugarri lexikoak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Ez-funtsezko hitzak | 50 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 68.18 | 63.64 | 63.64 | **86.36** | 68.18 | 59.09 |
| | 100 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 54.55 | 50 | 54.55 | 77.27 | 72.73 | 72.73 |
| | 150 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 45.45 | 59.09 | 59.09 | 77.27 | 72.73 | 77.27 |
| | 200 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 40.91 | 50 | 54.55 | 77.27 | 72.73 | 72.73 |
| | 250 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 68.18 | 59.09 | 63.64 | 77.27 | 72.73 | 72.73 |
| | 300 ez-funtsezko hitz eta bilaketa-prozesuan atributu guztiak | 68.18 | 68.18 | 77.27 | 77.27 | 72.73 | 72.73 |
| | 50 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 63.64 | 68.18 | 68.18 | 81.82 | 72.73 | 68.18 |
| | 100 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 63.64 | 59.09 | 59.09 | 72.73 | 68.18 | 68.18 |
| | 150 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 50 | 59.09 | 54.55 | 77.27 | 77.27 | 77.27 |
| | 200 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 45.45 | 54.55 | 50 | 77.27 | 72.73 | 72.73 |
| | 250 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 54.55 | 54.55 | 54.55 | 77.27 | 72.73 | 72.73 |
| | 300 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %75 | 72.73 | 72.73 | 68.18 | 81.82 | 68.18 | 59.09 |
| | 50 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 59.09 | 63.64 | 72.73 | 68.18 | 63.64 |

**4.4 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri lexikoak *erabiliz lortutako emaitzak (IV)*

| Ezaugarri lexikoak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Ez-funtsezko hitzak | 100 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 63.64 | 59.09 | 72.73 | 68.18 | 77.27 |
| | 150 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 45.45 | 50 | 59.09 | 81.82 | 68.18 | 68.18 |
| | 200 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 59.09 | 63.64 | 63.64 | 81.82 | 68.18 | 72.73 |
| | 250 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 68.18 | 81.82 | 68.18 | 77.27 |
| | 300 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %50 | 54.55 | 59.09 | 54.55 | 77.27 | 77.27 | 68.18 |
| | 50 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 54.55 | 54.55 | 59.09 | 72.73 | 68.18 | 59.09 |
| | 100 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 68.18 | 63.64 | 72.73 | 72.73 | 68.18 | 72.73 |
| | 150 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 63.64 | 68.18 | 68.18 | 72.73 | 68.18 | 77.27 |
| | 200 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 63.64 | 54.55 | 63.64 | **86.36** | 77.27 | 63.64 |
| | 250 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 68.18 | 59.09 | 63.64 | **86.36** | 77.27 | 63.64 |
| | 300 ez-funtsezko hitz eta bilaketa-prozesuan atributuen %25 | 72.73 | 72.73 | 68.18 | 81.82 | 68.18 | 59.09 |
| Hitz mailako n-gramak | n=2 | 50 | 50 | 50 | 50 | 63.64 | 68.18 |
| | n=3 | 50 | 50 | 50 | 50 | 54.55 | 45.45 |
| | n=4 | 50 | 50 | 50 | 50 | 50 | 50 |

**4.5 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri lexikoak *erabiliz lortutako emaitzak (V)*

Ez-funtsezko hitzek, bestalde, emaitza onenak *J48* algoritmoarekin eskaintzen dizkigute, zenbaitetan bertsoen %86.36 zuzen sailkatzera iritsiz. Kasu honetan, atributu-aukeraketarik gabe 50 ez-funtsezko hitz erabiliz lortzen dugu aipatutako emaitza. Atributu-aukeraketarekin, aldiz, bilaketa-prozesuan atributuen %25a kontuan hartuz, eta 200 edo 250 ez-funtsezko-hitz erabiliz iristen gara %86.36ko zehaztasunera.

Hitz mailako n-gramek, aldiz, ez dizkigute nahi bezain emaitza onak eskaintzen. Ezaugarri linguistiko honekin emaitzarik onena %68.18koa da, *SMO* ikasketa-algoritmoa erabiliz eta n-ren balioa 2 izanik.

Honez gain, aipatzekoa da aplikatutako bi ezaugarri lexikoetan (eduki-hitzetan eta ez-funtsezko hitzetan, hain zuzen ere) atributu-aukeraketarekin kasu askotan emaitzak hobetu arren, atributu-aukeraketarik gabe lortutako emaitzarik onenak hobetzera ez dela iritsi.

#### 4.3.1.2 Karaktere-ezaugarriak

4.6, 4.7 eta 4.8 tauletan dauzkagu karaktere mailako n-gramekin eta arestian aipatutako ikasketa-algoritmoekin lortutako emaitzak. Emaitzak hobetu nahian, atributu-aukeraketa aplikatu dugu bilaketa-prozesuan 100, 500, 1.000, 1.500 eta 2.000 atributu, eta atributu guztiak kontuan hartuz.

Emaitzei erreparatuta, *SMO* eta *Naive Bayes* dira algoritmorik onenak. N-ren balioari dagokionez, emaitza onenak 4, 5 eta 6 balioekin lortu dira. Ikus dezakegun moduan, lortutako emaitzarik onenak %86.36koak dira, aipatutako n-ren balioekin atributu-aukeraketarik aplikatu gabe, eta atributu-aukeraketarekin bilaketa-prozesuan atributu guztiak kontuan hartuz.

Bestalde, taulak gainetik begiratuta, oro har badirudi atributu-aukeraketak kasu honetan emaitzak txartzen dituela.

#### 4.3.1.3 Berariazko ezaugarriak

4.9, 4.10 eta 4.11 tauletan dauzkagu berariazko ezaugarriekin eta arestian aipatutako ikasketa-algoritmoekin lortutako emaitzak. Emaitzak hobetu nahian, atributu-aukeraketa aplikatu dugu bilaketa-prozesuan atributu kopuru desberdinak kontuan hartuz.

Aipatutako taulei erreparatuz gero, nabarmena da berariazko ezaugarrietan hitz ez-estandarrak direla emaitza onenak ematen dituztenak, errima eta oinek oso emaitza kax-

| Karaktere-ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Karaktere mailako n-gramak | n=2 | 63.64 | 54.55 | 54.55 | 68.18 | 59.09 | 72.73 |
| | n=3 | 54.55 | 59.09 | 63.64 | 68.18 | 63.64 | 72.73 |
| | n=4 | 59.09 | 68.18 | 59.09 | 45.45 | 81.82 | **86.36** |
| | n=5 | 54.55 | 72.73 | 59.09 | 59.09 | 81.82 | **86.36** |
| | n=6 | 54.55 | 50 | 59.09 | 50 | **86.36** | **86.36** |
| | n=7 | 50 | 50 | 50 | 54.55 | 81.82 | 77.27 |
| | n=2 eta bilaketa-prozesuan atributu guztiak | 63.64 | 54.55 | 54.55 | 68.18 | 59.09 | 72.73 |
| | n=3 eta bilaketa-prozesuan atributu guztiak | 54.55 | 59.09 | 63.64 | 63.64 | 59.09 | 72.73 |
| | n=4 eta bilaketa-prozesuan atributu guztiak | 59.09 | 68.18 | 59.09 | 50 | 81.82 | **86.36** |
| | n=5 eta bilaketa-prozesuan atributu guztiak | 54.55 | 72.73 | 59.09 | 63.64 | 81.82 | **86.36** |
| | n=6 eta bilaketa-prozesuan atributu guztiak | 54.55 | 54.55 | 59.09 | 50 | **86.36** | **86.36** |
| | n=7 eta bilaketa-prozesuan atributu guztiak | 50 | 50 | 50 | 54.55 | 81.82 | 77.27 |
| | n=2 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 68.18 | 81.82 | 68.18 | 68.18 | 68.18 | 63.64 |
| | n=3 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 50 | 50 | 50 | 50 | 77.27 | 68.18 |
| | n=4 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 50 | 50 | 50 | 59.09 | 63.64 | 63.64 |
| | n=5 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 54.55 | 50 | 50 | 59.09 | 54.55 | 54.55 |

**4.6 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* karaktere-ezaugarriak *erabiliz lortutako emaitzak (I)*

| Karaktere-ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Karaktere mailako n-gramak | n=6 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | n=7 eta bilaketa-prozesuan 100 atributu kontuan hartuz | 45.45 | 45.45 | 45.45 | 50 | 45.45 | 45.45 |
| | n=3 eta bilaketa-prozesuan 500 atributu kontuan hartuz | 54.55 | 54.55 | 54.55 | 72.73 | 68.18 | 77.27 |
| | n=4 eta bilaketa-prozesuan 500 atributu kontuan hartuz | 54.55 | 72.73 | 59.09 | 59.08 | 68.18 | 68.18 |
| | n=5 eta bilaketa-prozesuan 500 atributu kontuan hartuz | 50 | 50 | 50 | 59.09 | 54.55 | 54.55 |
| | n=6 eta bilaketa-prozesuan 500 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | n=7 eta bilaketa-prozesuan 500 atributu kontuan hartuz | 45.45 | 45.45 | 45.45 | 50 | 45.45 | 50 |
| | n=3 eta bilaketa-prozesuan 1.000 atributu kontuan hartuz | 54.55 | 59.09 | 59.09 | 68.18 | 63.64 | 63.64 |
| | n=4 eta bilaketa-prozesuan 1.000 atributu kontuan hartuz | 68.18 | 63.64 | 68.18 | 59.09 | 68.18 | 63.64 |
| | n=5 eta bilaketa-prozesuan 1.000 atributu kontuan hartuz | 45.45 | 59.09 | 54.55 | 59.09 | 54.55 | 72.73 |
| | n=6 eta bilaketa-prozesuan 1.000 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | n=7 eta bilaketa-prozesuan 1.000 atributu kontuan hartuz | 45.45 | 50 | 45.45 | 50 | 45.45 | 50 |
| | n=3 eta bilaketa-prozesuan 1.500 atributu kontuan hartuz | 68.18 | 54.55 | 63.64 | 72.73 | 72.73 | 77.27 |

**4.7 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* karaktere-ezaugarriak *erabiliz lortutako emaitzak (II)*

| Karaktere-ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Karaktere mailako n-gramak | n=4 eta bilaketa-prozesuan 1.500 atributu kontuan hartuz | 50 | 50 | 54.55 | 68.18 | 68.18 | 68.18 |
| | n=5 eta bilaketa-prozesuan 1.500 atributu kontuan hartuz | 59.09 | 59.09 | 59.09 | 59.09 | 63.64 | 68.18 |
| | n=6 eta bilaketa-prozesuan 1.500 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | n=7 eta bilaketa-prozesuan 1.500 atributu kontuan hartuz | 50 | 50 | 50 | 50 | 45.45 | 50 |
| | n=3 eta bilaketa-prozesuan 2.000 atributu kontuan hartuz | 63.64 | 59.09 | 63.64 | 72.73 | 72.73 | 72.73 |
| | n=4 eta bilaketa-prozesuan 2.000 atributu kontuan hartuz | 54.55 | 59.09 | 50 | 68.18 | 68.18 | 72.73 |
| | n=5 eta bilaketa-prozesuan 2.000 atributu kontuan hartuz | 45.45 | 45.45 | 50 | 59.09 | 54.55 | 59.09 |
| | n=6 eta bilaketa-prozesuan 2.000 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | n=7 eta bilaketa-prozesuan 2.000 atributu kontuan hartuz | 50 | 50 | 50 | 50 | 45.45 | 50 |

**4.8 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* karaktere-ezaugarriak *erabiliz lortutako emaitzak (III)*

| Berariazko ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Hitz ez-estandarrak | 50 hitz ez-estandar | 68.18 | 63.64 | 72.73 | 63.64 | 81.82 | 81.82 |
| | 100 hitz ez-estandar | 72.73 | 68.18 | 68.18 | 63.64 | 81.82 | **86.36** |
| | 150 hitz ez-estandar | 54.55 | 50 | 54.55 | 63.64 | 81.82 | **86.36** |
| | 200 hitz ez-estandar | 54.55 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 250 hitz ez-estandar | 50 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 300 hitz ez-estandar | 50 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 350 hitz ez-estandar | 54.55 | 54.55 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 50 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 68.18 | 63.64 | 72.73 | 63.64 | 81.82 | 81.82 |
| | 100 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 72.73 | 68.18 | 68.18 | 63.64 | 81.82 | **86.36** |
| | 150 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 50 | 50 | 50 | 63.64 | 81.82 | **86.36** |
| | 200 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 54.55 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 250 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 50 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 300 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 50 | 50 | 50 | 63.64 | 81.82 | 81.82 |
| | 350 hitz ez-estandar eta bilaketa-prozesuan atributu guztiak | 54.55 | 54.55 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 50 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 68.18 |
| | 100 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 63.64 | 50 | 50 | 63.64 | 63.64 | 68.18 |
| | 150 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 50 | 50 | 50 | 63.64 | 63.64 | 72.73 |

**4.9 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* berariazko ezaugarriak *deiturikoak erabiliz lortutako emaitzak (I)*

| Berariazko ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Hitz ez-estandarrak | 200 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 50 | 50 | 50 | 63.64 | 63.64 | 72.73 |
| | 250 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 50 | 50 | 50 | 63.64 | 63.64 | 68.18 |
| | 300 hitz ez-estandar eta bilaketa-prozesuan atributuen %25 | 50 | 50 | 50 | 63.64 | 63.64 | 63.64 |
| | 50 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 68.18 |
| | 100 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 77.27 |
| | 150 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 81.82 |
| | 200 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 77.27 |
| | 250 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 59.09 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 300 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 63.64 | 59.09 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 350 hitz ez-estandar eta bilaketa-prozesuan atributuen %50 | 59.09 | 54.55 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 50 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 77.27 |
| | 100 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 81.82 |
| | 150 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 77.27 |

**4.10 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* berariazko ezaugarriak *deiturikoak erabiliz lortutako emaitzak (II)*

| Berariazko ezaugarriak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Hitz ez-estandarrak | 200 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 59.09 | 59.09 | 54.55 | 63.64 | 81.82 | 77.27 |
| | 250 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 54.55 | 54.55 | 50 | 63.64 | 81.82 | 81.82 |
| | 300 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 54.55 | 54.55 | 54.55 | 63.64 | 81.82 | 77.27 |
| | 350 hitz ez-estandar eta bilaketa-prozesuan atributuen %75 | 54.55 | 54.55 | 50 | 63.64 | 81.82 | 81.82 |
| Errimak | - | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | Bilaketa-prozesuan atributu guztiak kontuan hartuz | 50 | 50 | 50 | 54.55 | 54.55 | 54.55 |
| | Bilaketa-prozesuan 25 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 40.91 | 40.91 |
| | Bilaketa-prozesuan 50 atributu kontuan hartuz | 50 | 50 | 50 | 54.55 | 40.91 | 40.91 |
| | Bilaketa-prozesuan 100 atributu kontuan hartuz | 54.55 | 54.55 | 54.55 | 54.55 | 54.55 | 54.55 |
| Oinak | - | 50 | 50 | 50 | 50 | 45.45 | 59.09 |
| | Bilaketa-prozesuan atributu guztiak kontuan hartuz | 50 | 50 | 50 | 50 | 45.45 | 59.09 |
| | Bilaketa-prozesuan 200 atributu kontuan hartuz | 50 | 50 | 50 | 50 | 40.91 | 45.45 |
| | Bilaketa-prozesuan 400 atributu kontuan hartuz | 45.45 | 45.45 | 45.45 | 50 | 50 | 50 |
| | Bilaketa-prozesuan 600 atributu kontuan hartuz | 50 | 50 | 50 | 50 | 54.55 | 63.64 |

**4.11 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* berariazko ezaugarriak *deiturikoak erabiliz lortutako emaitzak (III)*

karrak ematen baitituzte. 100 eta 150 hitz ez-estandarrekin, *SMO* algoritmoa erabiliz, %86.36ko zehaztasuna lortzen dugu, bai atributu-aukeraketarik gabe, eta baita atributu-aukeraketarekin bilaketa-prozesuan atributu guztiak kontuan hartuz ere.

Hitz ez-estandarren eta errimen kasuan, nahiz eta batzuetan atributu-aukeraketa emaitzak hobetu, ez da gai izan atributu-aukeraketarik gabe lorturiko emaitzarik onena hobetzeko. Oinen kasuan, bestalde, bai.

Algoritmoei dagokienez, taulei begiratuta, karaktere-ezaugarriekin gertatzen den modura *Naive Bayes* eta *SMO* dira emaitza onenak ematen dituztenak.

#### 4.3.1.4 Ezaugarrien konbinaketak

Atributu ezberdinen konbinaketa egokiek dokumentu baten egilea hautematen lagun dezaketela argi uzten dute [Koppel et al., 2009]-ek eta [Grieve, 2007]-k. Hori dela eta, 4.3.1.1, 4.3.1.2 eta 4.3.1.3 ataletan egindako probetan lortutako emaitza onenak kontuan hartuz, horiek hobetu nahian, ezaugarri linguistiko onenak konbinatuko ditugu binaka eta hirunaka. Horrela, ondorengoak dira elkar konbinatzeko aukeratu ditugun ezaugarriak (bakoitzaren alboan eskaini dizkiguten emaitza onenak ditugu):

- **Eduki-hitzak**
  - 500 eduki-hitz: %95.45.
  - 1.000 eduki-hitz: %90.91.
- **Ez-funtsezko hitzak**
  - 50 ez-funtsezko hitz: %86.36.
- **Karaktere mailako n-gramak**
  - n=4: %86.36.
  - n=5: %86.36.
  - n=6: %86.36.
- **Hitz ez-estandarrak**
  - 100 hitz ez-estandar: %86.36.
  - 150 hitz ez-estandar: %86.36.

Eman dituzten emaitza kaxkarrak direla eta, gainontzeko ezaugarriak baztertu egin ditugu.

Gauzak horrela, 4.12, 4.13, 4.14 eta 4.15 tauletan ditugu egindako saioen emaitzak.

Aipatutako taulak aztertuz gero, ezaugarri linguistikoak konbinatzeak maiz aurretik lortutako emaitzak hobetu dituela ikus dezakegu. Ezaugarri linguistiko bakarrarekin lortutako zehaztasunik handiena %95.45ekoa zen, eduki-hitzekin lortua. Eduki-hitzak ez-funtsezko hitzekin edo karaktere mailako n-gramekin konbinatzean aipatutako zehaztasun horretara iristen gara, eta eduki-hitzak hitz ez-estandarrekin konbinatuz, zehaztasun hori

| Ezaugarrien konbinaketak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 59.09 | 59.09 | 59.09 | 68.18 | 72.73 | **95.45** |
| | 1.000 eduki-hitz eta 50 hitz ez-funtsezko | 63.64 | 54.55 | 54.55 | 68.18 | 77.27 | 86.36 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=4 | 59.09 | 59.09 | 63.64 | 68.18 | 81.82 | 86.36 |
| | 500 eduki-hitz eta n=5 | 50 | 45.45 | 50 | 77.27 | 81.82 | **95.45** |
| | 500 eduki-hitz eta n=6 | 50 | 50 | 50 | 77.27 | 81.82 | 90.91 |
| | 1.000 eduki-hitz eta n=4 | 54.55 | 59.09 | 59.09 | 68.18 | 81.82 | 86.36 |
| | 1.000 eduki-hitz eta n=5 | 50 | 59.09 | 45.45 | 77.27 | 81.82 | **95.45** |
| | 1.000 eduki-hitz eta n=6 | 50 | 50 | 50 | 72.73 | 81.82 | 90.91 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 59.09 | 63.64 | 54.55 | 63.64 | 81.82 | **100** |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 68.18 | 63.64 | 54.55 | 63.64 | 81.82 | **100** |
| | 1.000 eduki-hitz eta 100 hitz ez-estandar | 68.18 | 59.09 | 54.55 | 72.73 | 81.82 | 90.91 |
| | 1.000 eduki-hitz eta 150 hitz ez-estandar | 59.09 | 54.55 | 68.18 | 68.18 | 81.82 | 86.36 |
| Ez-funtsezko hitzak + Karaktere n-gramak | 50 hitz ez-funtsezko eta n=4 | 63.64 | 63.64 | 59.09 | 68.18 | 81.82 | 81.82 |
| | 50 hitz ez-funtsezko eta n=5 | 59.09 | 59.09 | 54.55 | 68.18 | 81.82 | 81.82 |
| | 50 hitz ez-funtsezko eta n=6 | 50 | 54.55 | 54.55 | 72.73 | 86.36 | 86.36 |
| Ez-funtsezko hitzak + Hitz ez-estandarrak | 50 hitz ez-funtsezko eta 100 hitz ez-estandar | 50 | 54.55 | 54.55 | 72.73 | 72.73 | 72.73 |

**4.12 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri linguistikoak konbinatuz *lortutako emaitzak (I)*

| Ezaugarrien konbinaketak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Ez-funtsezko hitzak + Hitz ez-estandarrak | 50 hitz ez-funtsezko eta 150 hitz ez-estandar | 50 | 40.91 | 54.55 | 63.64 | 72.73 | 72.73 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | **100** | **100** | **100** | **100** | **100** | **100** |
| | 150 hitz ez-estandar eta n=4 | **100** | **100** | **100** | **100** | **100** | **100** |
| | 100 hitz ez-estandar eta n=5 | **100** | **100** | **100** | 90.91 | **100** | **100** |
| | 150 hitz ez-estandar eta n=5 | **100** | **100** | **100** | 90.91 | **100** | **100** |
| | 100 hitz ez-estandar eta n=6 | **100** | **100** | **100** | 63.64 | **100** | **100** |
| | 150 hitz ez-estandar eta n=6 | **100** | **100** | **100** | 63.64 | **100** | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Karaktere n-gramak | 500 eduki-hitz, 50 ez-funtsezko eta n=4 | 59.09 | 63.64 | 59.09 | 68.18 | 81.82 | 81.82 |
| | 500 eduki-hitz, 50 ez-funtsezko eta n=5 | 50 | 45.45 | 45.45 | 90.91 | 81.82 | 81.82 |
| | 500 eduki-hitz, 50 ez-funtsezko eta n=6 | 54.55 | 50 | 54.55 | 59.09 | 81.82 | 86.36 |
| | 1.000 eduki-hitz, 50 ez-funtsezko eta n=4 | 54.55 | 68.18 | 59.09 | 68.18 | 81.82 | 81.82 |
| | 1.000 eduki-hitz, 50 ez-funtsezko eta n=5 | 50 | 50 | 50 | 90.91 | 81.82 | 81.82 |
| | 1.000 eduki-hitz, 50 ez-funtsezko eta n=6 | 50 | 59.09 | 54.55 | 59.09 | 81.82 | 90.91 |

**4.13 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri linguistikoak konbinatuz *lortutako emaitzak (II)*

| Ezaugarrien konbinaketak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 100 ez-estandar | 59.09 | 54.55 | 59.09 | 63.64 | 77.27 | 90.91 |
| | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 68.18 | 63.64 | 77.27 | 72.73 | 77.27 | **95.45** |
| | 1.000 eduki-hitz, 50 ez-funtsezko eta 100 ez-estandar | 54.55 | 59.09 | 54.55 | 68.18 | 77.27 | 90.91 |
| | 1.000 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 68.18 | 554.55 | 72.73 | 63.64 | 77.27 | 90.91 |
| Ez-funtsezko hitzak + Karaketre n-gramak + Hitz ez-estandarrak | 50 hitz ez-funtsezko, n=4 eta 100 ez-estandar | 63.64 | 63.64 | 63.64 | 63.64 | 81.82 | 77.27 |
| | 50 hitz ez-funtsezko, n=4 eta 150 ez-estandar | 63.64 | 63.64 | 54.55 | 63.64 | 81.82 | 81.82 |
| | 50 hitz ez-funtsezko, n=5 eta 100 ez-estandar | 50 | 59.09 | 50 | 72.73 | 86.36 | 81.82 |
| | 50 hitz ez-funtsezko, n=5 eta 150 ez-estandar | 54.55 | 59.09 | 50 | 72.73 | 86.36 | 86.36 |
| | 50 hitz ez-funtsezko, n=6 eta 100 ez-estandar | 50 | 45.45 | 50 | 50 | 86.36 | 81.82 |
| | 50 hitz ez-funtsezko, n=6 eta 150 ez-estandar | 54.55 | 50 | 45.45 | 50 | 86.36 | 81.82 |
| Eduki-hitzak + Karaktere n-gramak + Hitz ez-estandarrak | 500 eduki-hitz, n=4 eta 100 ez-estandar | 59.09 | 59.09 | 54.55 | 54.55 | 81.82 | 90.91 |
| | 500 eduki-hitz, n=4 eta 150 ez-estandar | 63.64 | 59.09 | 59.09 | 54.55 | 81.82 | 90.91 |
| | 500 eduki-hitz, n=5 eta 100 ez-estandar | 50 | 50 | 50 | 81.82 | 81.82 | 90.91 |

**4.14 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri linguistikoak konbinatuz *lortutako emaitzak (III)*

| Ezaugarrien konbinaketak | Bestelako ezaugarriak | Ikasketa-algoritmoak | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak + Karaktere n-gramak + Hitz ez-estandarrak | 500 eduki-hitz, n=5 eta 150 ez-estandar | 50 | 50 | 50 | 81.82 | 81.82 | 86.36 |
| | 500 eduki-hitz, n=6 eta 100 ez-estandar | 50 | 50 | 50 | 59.09 | 81.82 | 90.91 |
| | 500 eduki-hitz, n=6 eta 150 ez-estandar | 50 | 50 | 50 | 77.27 | 81.82 | 86.36 |
| | 1.000 eduki-hitz, n=4 eta 100 ez-estandar | 54.55 | 63.64 | 54.55 | 54.55 | 81.82 | 90.91 |
| | 1.000 eduki-hitz, n=4 eta 150 ez-estandar | 54.55 | 59.09 | 50 | 54.55 | 81.82 | 90.91 |
| | 1.000 eduki-hitz, n=5 eta 100 ez-estandar | 50 | 54.55 | 54.55 | 81.82 | 81.82 | 86.36 |
| | 1.000 eduki-hitz, n=5 eta 150 ez-estandar | 50 | 50 | 50 | 81.82 | 81.82 | 86.36 |
| | 1.000 eduki-hitz, n=6 eta 100 ez-estandar | 50 | 50 | 50 | 59.09 | 81.82 | 90.91 |
| | 1.000 eduki-hitz, n=6 eta 150 ez-estandar | 50 | 50 | 50 | 77.27 | 81.82 | 81.82 |

**4.15 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz eta* ezaugarri linguistikoak konbinatuz *lortutako emaitzak (IV)*

hobetzea lortzen dugu %100ko emaitzarekin. Hala ere, aipagarriena hitz ez-estandarrak karaktere mailako n-gramekin konbinatzean lortutako emaitzak dira, ia kasu guztietan bertsoen %100ak zuzen sailkatzen baitira. Bestalde, aipatzekoa da oro har, ezaugarri linguistikoak hirunaka konbinatuz ez ditugula binaka konbinatuz bezain emaitza onak lortu. Emaitza onena %95.45ekoa izan da, 500 eduki-hitz, 50 ez-funtsezko hitz eta 150 hitz ez-estandar konbinatuz. Honez gain, nabarmena da *SMO* dela emaitza onenak eskaintzen dituen ikasketa-algoritmoa.

#### 4.3.1.5 Aukeratutako atributuak

4.3.1.1, 4.3.1.2, 4.3.1.3 eta 4.3.1.4 azpiataletan garatutako sailkatzaileetatik emaitza onenak eman dituztenak aukeratuko ditugu, eta horiek izango dira gainontzeko esperimentuetan erabiliko ditugunak. Horrela, jarraian ditugu aukeratutako sailkatzaileak atalka, eta bakoitzean eskaini diguten emaitza onenarekin:

- **Eduki-hitzak**
  - 500 eduki-hitz: %95.45.
  - 500 eduki-hitz eta atributu-aukeraketako bilaketa-prozesuan atributu guztiak kontuan hartuta: %95.45.
  - 1.500 eduki-hitz eta atributu-aukeraketako bilaketa-prozesuan atributuen %50a kontuan hartuta: %95.45.
- **Eduki-hitzak + Ez-funtsezko hitzak**
  - 500 eduki-hitz eta 50 ez-funtsezko hitz: %95.45.
- **Eduki-hitzak + Karaktere mailako n-gramak**
  - 500 eduki-hitz eta n=5: %95.45.
  - 1.000 eduki-hitz eta n=5: %95.45.
- **Eduki-hitzak + Hitz ez-estandarrak**
  - 500 eduki-hitz eta 100 hitz ez-estandar: %100.
  - 500 eduki-hitz eta 150 hitz ez-estandar: %100.
- **Hitz ez-estandarrak + Karaktere mailako n-gramak**

  - 100 hitz ez-estandar eta n=4: %100.
  - 150 hitz ez-estandar eta n=4: %100.
  - 100 hitz ez-estandar eta n=5: %100.
  - 150 hitz ez-estandar eta n=5: %100.
  - 100 hitz ez-estandar eta n=6: %100.
  - 150 hitz ez-estandar eta n=6: %100.

- **Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak**
  - 500 eduki-hitz, 50 ez-funtsezko hitz eta 150 hitz ez-estandar: %95.45.

#### 4.3.1.6 Azken probak *Test* corpusean

4.3.1.5 azpiataletan aukeratu ditugun atributuak erabiliz *Test* corpusean egingo ditugu probak. Horrela, lortutako azken emaitzekin, garatutako sailkatzaileak ebaluatuko ditugu.

4.16 taulan ikus daitekeenez, eta normala den bezala, oro har sailkatzaileen garapenean lortutako emaitzak azken proba hauetan txartu egiten dira. Hala ere, garapenean bezala, instantzia edo bertso guztiak zuzen sailkatu eta %100ko zehaztasuna lortu dugu hitz ez-estandarrak eta karaktere mailako n-gramak konbinatuz.

| **Ezaugarriak** | **Bestelako ezaugarriak** | **Ikasketa-algoritmoak** | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz | 50 | 54.54 | 54.54 | 59.09 | 68.18 | 63.64 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 50 | 54.54 | 54.54 | 50 | 68.18 | 63.64 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 54.55 | 54.54 | 59.09 | 59.09 | 63.64 | 77.27 |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 54.55 | 59.09 | 68.18 | 63.64 | 77.27 | 77.27 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=5 | 50 | 50 | 50 | 45.45 | 54.55 | 63.64 |
| | 1.000 eduki-hitz eta n=5 | 50 | 50 | 50 | 45.45 | 54.55 | 68.18 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 68.18 | 59.09 | 63.64 | 63.64 | 63.64 | 63.64 |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 50 | 54.55 | 54.55 | 63.64 | 63.64 | 59.09 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | 81.82 | 81.82 | 86.36 | 59.09 | 86.36 | **100** |
| | 150 hitz ez-estandar eta n=4 | 81.82 | 81.82 | 86.36 | 59.09 | 86.36 | **100** |
| | 100 hitz ez-estandar eta n=5 | 59.09 | 59.09 | 59.09 | 50 | 86.36 | **100** |
| | 150 hitz ez-estandar eta n=5 | 59.09 | 59.09 | 59.09 | 50 | 86.36 | **100** |
| | 100 hitz ez-estandar eta n=6 | 63.64 | 59.09 | 59.09 | 50 | 77.27 | **100** |
| | 150 hitz ez-estandar eta n=6 | 63.64 | 59.09 | 59.09 | 50 | 77.27 | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 50 | 50 | 50 | 63.64 | 77.27 | 72.73 |

**4.16 Taula:** *Egaña eta Iturriagaren bertsoak erabiliz Test corpusean egindako probak*

### 4.3.2 2 bertsolari (II)

4.3.1 atalean, Andoni Egaña eta Unai Iturriaga ziren gure sailkatzaileak garatzeko aukeratuta bertsolariak, bertso kopuru gehien zituztenak baitziren. Andoni Egaña zarauztarra da, eta horregatik, erdialdeko euskara edo gipuzkera da bere euskalkia. Unai Iturriaga berriz, durangarra da, eta ondorioz, mendebaldeko euskaraz edo bizkaieraz mintzo da. 4.3.1 atalean egindako esperimentuetan hori nabarmena da, hitz ez-estandarrek emaitza bikainak eman baitituzte. Garatutako sailkatzaile onenetan adibidez, hitz ez-estandarrak karaktere mailako n-gramekin konbinatzen dira.

Gauzak horrela, antzeko hizkera duten bi bertsolari aukeratu ditugu garatutako sailkatzaileak ebaluatzeko:

- Andoni Egaña, 127 bertsorekin.
- Maialen Lujanbio, 95 bertsorekin.

Maialen Lujanbio hernaniarra da, eta Andoni Egañarekin gertatzen den moduan, erdialdeko euskara edo gipuzkera da bere euskalkia. Ondorioz, gure probetarako bertsolari egokiak dira.

Dakigun bezala, bi bertsolari hauen 95 bertsoekin corpus berri bat sortuko beharko dugu, bertsoak bertsolarien arabera sailkatuz:

- Andoni Egaña: 95 bertso eta 4.483 token.
- Maialen Lujanbio: 95 bertso eta 4.315 token.

Ondoren, *Train/Develop/Test* banaketa egingo da ausaz:

- Andoni Egaña: *Train*-erako 75 bertso (3.639 token), *Develop*-erako 10 bertso (393 token) eta *Test*-erako 10 bertso (451 token).
- Maialen Lujanbio: *Train*-erako 75 bertso (3.396 token), *Develop*-erako 10 bertso (473 token) eta *Test*-erako 10 bertso (446 token).

Corpusa prestatu eta gero, 4.3.1.5 azpiatalean aukeratutako sailkatzaile onenekin egindako esperimentuen emaitzak 4.17 taulan ikus ditzakegu .

| **Ezaugarriak** | **Bestelako ezaugarriak** | **Ikasketa-algoritmoak** | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz | 50 | 55 | 50 | 65 | 80 | 70 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 50 | 55 | 50 | 60 | 80 | 70 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 45 | 55 | 50 | 55 | 65 | 55 |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 60 | 70 | 55 | 70 | 65 | 75 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=5 | 50 | 50 | 50 | 50 | 70 | 65 |
| | 1.000 eduki-hitz eta n=5 | 50 | 50 | 50 | 50 | 70 | 55 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 55 | 55 | 50 | 65 | 80 | 70 |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 55 | 55 | 50 | 65 | 75 | 65 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | 70 | 70 | 70 | 50 | 95 | 95 |
| | 150 hitz ez-estandar eta n=4 | 70 | 70 | 70 | 50 | 95 | 95 |
| | 100 hitz ez-estandar eta n=5 | 70 | 75 | 70 | 50 | 95 | **100** |
| | 150 hitz ez-estandar eta n=5 | 70 | 75 | 70 | 50 | 95 | **100** |
| | 100 hitz ez-estandar eta n=6 | 65 | 65 | 65 | 50 | 90 | **100** |
| | 150 hitz ez-estandar eta n=6 | 65 | 65 | 65 | 50 | 90 | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 60 | 55 | 55 | 65 | 70 | 65 |

**4.17 Taula:** *Egaña eta Lujanbioren bertsoak erabiliz Test corpusean egindako probak*

Emaitzei erreparatuta, ez dirudi euskalkiak sailkatzaileen arrakastaren ardatz direnik. Antzeko hizkera duten bi bertsolariren bertsoak erabilita ere, bertsoen %100a zuzen sailkatzea lortu dugu *SMO* algoritmoarekin hitz ez-estandarrak eta karaktere mailako n-gramak konbinatuz.

### 4.3.3 5 bertsolari

Ondorengo esperimentuan, garatutako sailkatzaileak bost bertsolariren bertsoekin jarriko ditugu proban. Lehenengo lana, bertso gehien dituzten bost bertsolarirekin corpusa sortu eta *Train/Develop/Test* banaketa osatzea izango da. Bertsolariak hauek izango dira:

1. Andoni Egaña, 127 bertsorekin.
2. Unai Iturriaga, 110 bertsorekin.
3. Maialen Lujanbio, 95 bertsorekin.
4. Jon Maia, 91 bertsorekin.
5. Aitor Mendiluze, 89 bertsorekin.

Bost bertsolari hauen 89 bertsoekin corpus berri bat sortuko dugu, bertsoak bertsolarien arabera sailkatuz:

1. Andoni Egaña: 89 bertso eta 4.187 token.
2. Unai Iturriaga: 89 bertso eta 4.129 token.
3. Maialen Lujanbio: 89 bertso eta 4.018 token.
4. Jon Maia: 89 bertso eta 4.286 token.
5. Aitor Mendiluze: 89 bertso eta 3.528 token.

Ondoren, *Train/Develop/Test* banaketa egingo da ausaz:

1. Andoni Egaña: *Train*-erako 71 bertso (3.292 token), *Develop*-erako 9 bertso (394 token) eta *Test*-erako 9 bertso (501 token).

2. Unai Iturriaga: *Train*-erako 71 bertso (3.264 token), *Develop*-erako 9 bertso (408 token) eta *Test*-erako 9 bertso (457 token).

3. Maialen Lujanbio: *Train*-erako 71 bertso (3.135 token), *Develop*-erako 9 bertso (398 token) eta *Test*-erako 9 bertso (485 token).

4. Jon Maia: *Train*-erako 71 bertso (3.354 token), *Develop*-erako 9 bertso (485 token) eta *Test*-erako 9 bertso (447 token).

5. Aitor Mendiluze: *Train*-erako 71 bertso (2.751 token), *Develop*-erako 9 bertso (398 token) eta *Test*-erako 9 bertso (379 token).

Corpusa prestatu eta gero, 4.18 taulan biltzen dira lortutako emaitzak.

Emaitzei erreparatuta, oro har bi bertsolarirekin lortutako emaitzak txartzen diren arren, *SMO* algoritmoarekin hitz ez-estandarrak eta karaktere mailako n-gramak konbinatuz bertsoen %100a zuzen sailkatzea lortu dugula ikus dezakegu.

### 4.3.4 10 bertsolari

Ondorengo esperimentuan, garatutako sailkatzaileak hamar bertsolariren bertsoekin jarriko ditugu proban. Lehenengo lana, bertso gehien dituzten hamar bertsolarirekin corpusa sortu eta *Train/Develop/Test* banaketa osatzea izango da. Bertsolariak hauek izango dira:

1. Andoni Egaña, 127 bertsorekin.

2. Unai Iturriaga, 110 bertsorekin.

3. Maialen Lujanbio, 95 bertsorekin.

4. Jon Maia, 91 bertsorekin.

5. Aitor Mendiluze, 89 bertsorekin.

6. Igor Elortza, 79 bertsorekin.

7. Sustrai Colina, 77 bertsorekin.

8. Amets Arzallus, 72 bertsorekin.

9. Aitor Sarriegi, 72 bertsorekin

| **Ezaugarriak** | **Bestelako ezaugarriak** | **Ikasketa-algoritmoak** | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz | 22.22 | 24.44 | 22.22 | 22.22 | 37.78 | 44.44 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 20 | 26.67 | 22.22 | 26.67 | 37.78 | 44.44 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 22.22 | 13.33 | 13.33 | 22.22 | 26.67 | 40 |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 24.44 | 22.22 | 26.67 | 24.44 | 40 | 48.89 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=5 | 20 | 20 | 20 | 26.67 | 46.67 | 55.56 |
| | 1.000 eduki-hitz eta n=5 | 20 | 22.22 | 22.22 | 20 | 44.44 | 51.11 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 20 | 20 | 20 | 20 | 42.22 | 46.67 |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 22.22 | 20 | 20 | 17.78 | 44.44 | 48.89 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | 66.67 | 64.44 | 64.44 | 60 | 80 | **100** |
| | 150 hitz ez-estandar eta n=4 | 66.67 | 64.44 | 64.44 | 60 | 80 | **100** |
| | 100 hitz ez-estandar eta n=5 | 64.44 | 64.44 | 64.44 | 40 | 84.44 | **100** |
| | 150 hitz ez-estandar eta n=5 | 64.44 | 64.44 | 64.44 | 40 | 84.44 | **100** |
| | 100 hitz ez-estandar eta n=6 | 51.11 | 42.22 | 46.67 | 40 | 84.44 | **100** |
| | 150 hitz ez-estandar eta n=6 | 65 | 65 | 65 | 50 | 90 | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 24.44 | 24.44 | 24.44 | 22.22 | 42.22 | 51.11 |

**4.18 Taula:** *5 bertsolarirekin Test corpusean egindako probak*

10. Mikel Mendizabal, 64 bertsorekin.

Hamar bertsolari hauen 64 bertsoekin corpus berri bat sortuko dugu, bertsoak bertsolarien arabera sailkatuz:

1. Andoni Egaña: 64 bertso eta 3.085 token.
2. Unai Iturriaga: 64 bertso eta 3.108 token.
3. Maialen Lujanbio: 64 bertso eta 3.077 token.
4. Jon Maia: 64 bertso eta 3.095 token.
5. Aitor Mendiluze: 64 bertso eta 2.449 token.
6. Igor Elortza: 64 bertso eta 2.542 token.
7. Sustrai Colina: 64 bertso eta 2.462 token.
8. Amets Arzallus: 64 bertso eta 2.750 token.
9. Aitor Sarriegi: 64 bertso eta 2.495 token.
10. Mikel Mendizabal: 64 bertso eta 2.357 token.

Ondoren, *Train/Develop/Test* banaketa egingo da ausaz:

1. Andoni Egaña: *Train*-erako 52 bertso (2.406 token), *Develop*-erako 6 bertso (316 token) eta *Test*-erako 6 bertso (363 token).
2. Unai Iturriaga: *Train*-erako 52 bertso (2.525 token), *Develop*-erako 6 bertso (265 token) eta *Test*-erako 6 bertso (318 token).
3. Maialen Lujanbio: *Train*-erako 52 bertso (2.540 token), *Develop*-erako 6 bertso (263 token) eta *Test*-erako 6 bertso (274 token).
4. Jon Maia: *Train*-erako 52 bertso (2.532 token), *Develop*-erako 6 bertso (294 token) eta *Test*-erako 6 bertso (269 token).
5. Aitor Mendiluze: *Train*-erako 52 bertso (1.984 token), *Develop*-erako 6 bertso (234 token) eta *Test*-erako 6 bertso (231 token).

6. Igor Elortza: *Train*-erako 52 bertso (2.097 token), *Develop*-erako 6 bertso (203 token) eta *Test*-erako 6 bertso (242 token).
7. Sustrai Colina: *Train*-erako 52 bertso (1.936 token), *Develop*-erako 6 bertso (224 token) eta *Test*-erako 6 bertso (302 token).
8. Amets Arzallus: *Train*-erako 52 bertso (2.302 token), *Develop*-erako 6 bertso (257 token) eta *Test*-erako 6 bertso (191 token).
9. Aitor Sarriegi: *Train*-erako 52 bertso (2.066 token), *Develop*-erako 6 bertso (241 token) eta *Test*-erako 6 bertso (188 token).
10. Mikel Mendizabal: *Train*-erako 52 bertso (1.878 token), *Develop*-erako 6 bertso (233 token) eta *Test*-erako 6 bertso (246 token).

Corpusa prestatu eta gero, 4.19 taulan biltzen dira lortutako emaitzak.

Aipatutako taulak aztertuta, oro har bost bertsolarirekin lortutako emaitzak txartzen diren arren, *SMO* algoritmoarekin hitz ez-estandarrak eta karaktere mailako n-gramak konbinatuz bertsoen %100a zuzen sailkatzea lortu dugula ikus dezakegu.

### 4.3.5 15 bertsolari

Ondorengo esperimentuan, garatutako sailkatzaileak hamabost bertsolariren bertsoekin jarriko ditugu proban. Lehenengo lana, bertso gehien dituzten hamabost bertsolarirekin corpusa sortu eta *Train/Develop/Test* banaketa osatzea izango da. Bertsolariak hauek izango dira:

1. Andoni Egaña, 127 bertsorekin.
2. Unai Iturriaga, 110 bertsorekin.
3. Maialen Lujanbio, 95 bertsorekin.
4. Jon Maia, 91 bertsorekin.
5. Aitor Mendiluze, 89 bertsorekin.
6. Igor Elortza, 79 bertsorekin.
7. Sustrai Colina, 77 bertsorekin.

| **Ezaugarriak** | **Bestelako ezaugarriak** | **Ikasketa-algoritmoak** | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz | 15 | 13.33 | 13.33 | 16.67 | 20 | 30 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 15 | 13.33 | 13.33 | 16.67 | 20 | 30 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 10 | 10 | 10 | 15 | 21.67 | 26.67 |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 18.33 | 20 | 18.33 | 21.66 | 23.33 | 33.33 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=5 | 10 | 11.67 | 10 | 15 | 30 | 36.67 |
| | 1.000 eduki-hitz eta n=5 | 10 | 13.33 | 10 | 21.67 | 28.33 | 40 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 11.67 | 11.67 | 11.67 | 16.67 | 21.67 | 31.67 |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 15 | 16.67 | 20 | 16.67 | 21.67 | 33.33 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | 70 | 68.33 | 70 | 16.67 | 71.67 | **100** |
| | 150 hitz ez-estandar eta n=4 | 70 | 68.33 | 70 | 16.67 | 71.67 | **100** |
| | 100 hitz ez-estandar eta n=5 | 41.67 | 43.33 | 43.33 | 50 | 53.33 | **100** |
| | 150 hitz ez-estandar eta n=5 | 41.67 | 43.33 | 43.33 | 50 | 53.33 | **100** |
| | 100 hitz ez-estandar eta n=6 | 33.33 | 38.33 | 36.67 | 13.33 | 50 | **100** |
| | 150 hitz ez-estandar eta n=6 | 33.33 | 38.33 | 36.67 | 13.33 | 50 | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 20 | 20 | 20 | 30 | 25 | 35 |

**4.19 Taula:** *10 bertsolarirekin Test corpusean egindako probak*

8. Amets Arzallus, 72 bertsorekin.
9. Aitor Sarriegi, 72 bertsorekin
10. Mikel Mendizabal, 64 bertsorekin.
11. Jesus Mari Irazu, 63 bertsorekin.
12. Jokin Sorozabal, 59 bertsorekin.
13. Sebastian Lizaso, 56 bertsorekin.
14. Jon Enbeita, 55 bertsorekin.
15. Xabier Euzkitze, 52 bertsorekin.

Hamabost bertsolari hauen 52 bertsoekin corpus berri bat sortuko dugu, bertsoak bertsolarien arabera sailkatuz:

1. Andoni Egaña: 52 bertso eta 2.489 token.
2. Unai Iturriaga: 52 bertso eta 2.475 token.
3. Maialen Lujanbio: 52 bertso eta 2.334 token.
4. Jon Maia: 52 bertso eta 2.547 token.
5. Aitor Mendiluze: 52 bertso eta 2.049 token.
6. Igor Elortza: 52 bertso eta 2.016 token.
7. Sustrai Colina: 52 bertso eta 2.218 token.
8. Amets Arzallus: 52 bertso eta 2.251 token.
9. Aitor Sarriegi: 52 bertso eta 2.161 token.
10. Mikel Mendizabal: 52 bertso eta 1.868 token.
11. Jesus Mari Irazu: 52 bertso eta 2.303 token.
12. Jokin Sorozabal: 52 bertso eta 1.841 token.
13. Sebastian Lizaso: 52 bertso eta 2.094 token.

14. Jon Enbeita: 52 bertso eta 2.176 token.

15. Xabier Euzkitze: 52 bertso eta 1.954 token.

Ondoren, *Train/Develop/Test* banaketa egingo da ausaz:

1. Andoni Egaña: *Train*-erako 42 bertso (2.010 token), *Develop*-erako 5 bertso (237 token) eta *Test*-erako 5 bertso (242 token).

2. Unai Iturriaga: *Train*-erako 42 bertso (1.989 token), *Develop*-erako 5 bertso (226 token) eta *Test*-erako 5 bertso (260 token).

3. Maialen Lujanbio: *Train*-erako 42 bertso (1.818 token), *Develop*-erako 5 bertso (280 token) eta *Test*-erako 5 bertso (236 token).

4. Jon Maia: *Train*-erako 42 bertso (2.002 token), *Develop*-erako 5 bertso (266 token) eta *Test*-erako 5 bertso (279 token).

5. Aitor Mendiluze: *Train*-erako 42 bertso (1.685 token), *Develop*-erako 5 bertso (159 token) eta *Test*-erako 5 bertso (205 token).

6. Igor Elortza: *Train*-erako 42 bertso (1.655 token), *Develop*-erako 5 bertso (116 token) eta *Test*-erako 5 bertso (245 token).

7. Sustrai Colina: *Train*-erako 42 bertso (1.829 token), *Develop*-erako 5 bertso (210 token) eta *Test*-erako 5 bertso (179 token).

8. Amets Arzallus: *Train*-erako 42 bertso (1.831 token), *Develop*-erako 5 bertso (203 token) eta *Test*-erako 5 bertso (217 token).

9. Aitor Sarriegi: *Train*-erako 42 bertso (1.792 token), *Develop*-erako 4 bertso (177 token) eta *Test*-erako 5 bertso (192 token).

10. Mikel Mendizabal: *Train*-erako 42 bertso (1.552 token), *Develop*-erako 5 bertso (166 token) eta *Test*-erako 5 bertso (150 token).

11. Jesus Mari Irazu: *Train*-erako 42 bertso (1.884 token), *Develop*-erako 5 bertso (189 token) eta *Test*-erako 5 bertso (230 token).

12. Jokin Sorozabal: *Train*-erako 42 bertso (1.510 token), *Develop*-erako 5 bertso (158 token) eta *Test*-erako 5 bertso (173 token).

13. Sebastian Lizaso: *Train*-erako 42 bertso (1.677 token), *Develop*-erako 5 bertso (212 token) eta *Test*-erako 5 bertso (205 token).

14. Jon Enbeita: *Train*-erako 42 bertso (1.841 token), *Develop*-erako 5 bertso (190 token) eta *Test*-erako 5 bertso (145 token).

15. Xabier Euzkitze: *Train*-erako 42 bertso (1.559 token), *Develop*-erako 5 bertso (227 token) eta *Test*-erako 5 bertso (168 token).

Corpusa prestatu eta gero, 4.20 taulan biltzen dira lortutako emaitzak.

Aipatutako tauletako emaitzei erreparatuta, oro har hamar bertsolarirekin lortutako emaitzak txartzen diren arren, *SMO* algoritmoarekin hitz ez-estandarrak eta karaktere mailako n-gramak konbinatuz bertsoen %100a zuzen sailkatzea lortu dugula ikus dezakegu.

## 4.4 Ondorioak

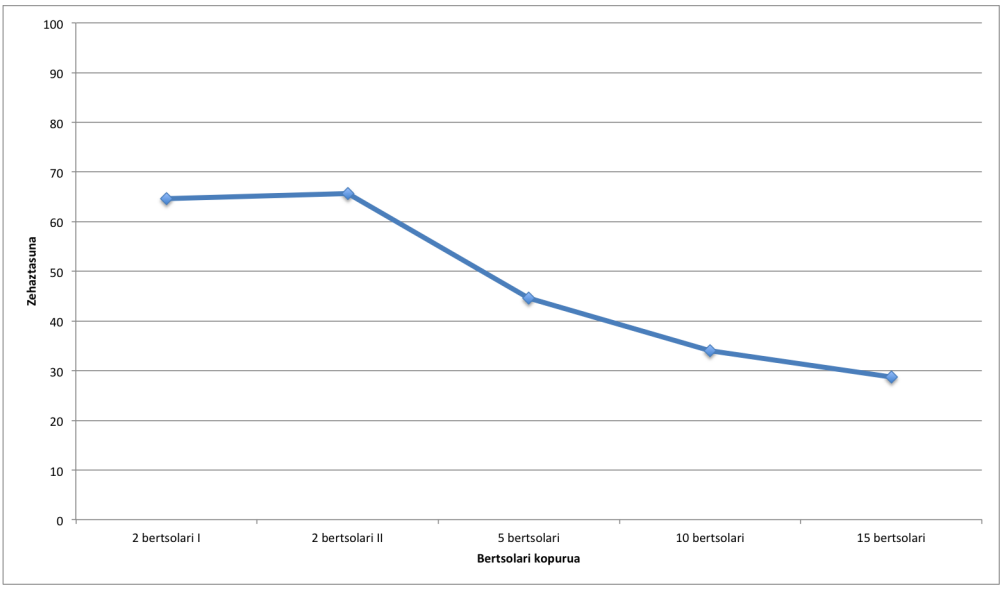
4.3 atalean zenbait bertsolari kopururekin hainbat saio egin ditugu. Egindako saioetako emaitzak aztertuz ondorio hauek atera ditzakegu.

Lehenik eta behin, aipatzekoa da hitz ez-estandarrak karaktere mailako n-gramekin konbinatzeak izan duen arrakasta. Ezaugarri linguistiko horiek erabiliz, *SMO* algoritmoarekin bertso guztiak zuzen sailkatzea lortu ditugu bi, bost, hamar eta hamabost bertsolarirekin egindako probetan.

Bertsolari kopuruaren arabera eginiko esperimentuak alderatuz gero, bertsolari kopurua handitu ahala, oro har emaitzak txartzen direla nabarmena da. 4.1 grafikoan argi antzeman dezakegu bertsolari kopuruak problemaren konplexutasunarekin duen erlazio zuzena. Hori neurtzeko, 4.3 ataleko saio bakoitzean *Test* corpusaren gainean egindako esperimentu guztien batezbestekoa kalkulatu da.

| **Ezaugarriak** | **Bestelako ezaugarriak** | **Ikasketa-algoritmoak** | | | | | |
|---|---|---|---|---|---|---|---|
| | | *9-NN* | *10-NN* | *11-NN* | *J48* | *Naive Bayes* | *SMO* |
| Eduki-hitzak | 500 eduki-hitz | 13.3 | 13.3 | 12 | 12 | 12 | 13.33 |
| | 500 eduki-hitz eta bilaketa-prozesuan atributu guztiak | 13.33 | 13.33 | 12 | 13.33 | 12 | 13.33 |
| | 1.500 eduki-hitz eta bilaketa-prozesuan atributuen %50 | 6.67 | 6.67 | 8 | 12 | 13.33 | 18.67 |
| Eduki-hitzak + Ez-funtsezko hitzak | 500 eduki-hitz eta 50 hitz ez-funtsezko | 9.33 | 9.33 | 9.33 | 17.33 | 9.33 | 20 |
| Eduki-hitzak + Karaktere n-gramak | 500 eduki-hitz eta n=5 | 6.67 | 6.67 | 6.67 | 6.67 | 10.67 | 26.67 |
| | 1.000 eduki-hitz eta n=5 | 6.67 | 6.67 | 6.67 | 13.33 | 9.33 | 28 |
| Eduki-hitzak + Hitz ez-estandarrak | 500 eduki-hitz eta 100 hitz ez-estandar | 9.33 | 8 | 10.67 | 10.67 | 13.33 | 18.67 |
| | 500 eduki-hitz eta 150 hitz ez-estandar | 8 | 10.67 | 12 | 12 | 13.33 | 21.33 |
| Hitz ez-estandarrak + Karaktere n-gramak | 100 hitz ez-estandar eta n=4 | 62.67 | 61.33 | 61.33 | 24 | 58.67 | **100** |
| | 150 hitz ez-estandar eta n=4 | 62.67 | 61.33 | 62.67 | 24 | 58.67 | **100** |
| | 100 hitz ez-estandar eta n=5 | 44 | 46.67 | 46.67 | 24 | 49.33 | **100** |
| | 150 hitz ez-estandar eta n=5 | 44 | 46.67 | 46.67 | 24 | 49.33 | **100** |
| | 100 hitz ez-estandar eta n=6 | 37.33 | 37.33 | 37.33 | 33.33 | 52 | **100** |
| | 150 hitz ez-estandar eta n=6 | 37.33 | 37.33 | 37.33 | 33.33 | 52 | **100** |
| Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak | 500 eduki-hitz, 50 ez-funtsezko eta 150 ez-estandar | 5.33 | 5.33 | 6.67 | 12 | 9.33 | 16 |

**4.20 Taula:** *15 bertsolarirekin Test corpusean egindako probak*




**4.1 Irudia:** *Bertsolari kopuruak sistemaren zehaztasunean duen eragina*

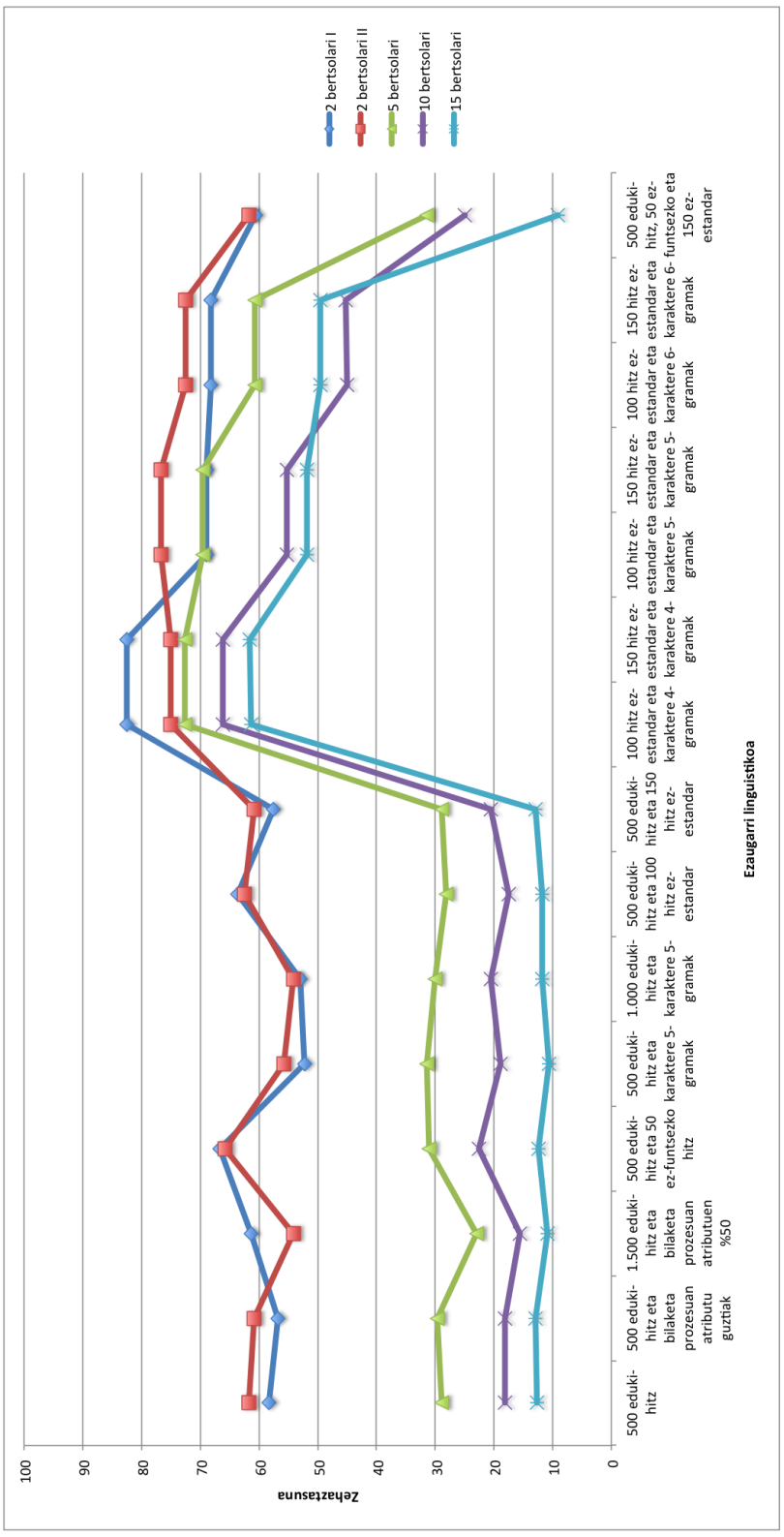
Egindako saioetako taulei erreparatuz emaitza onenak ematen dituzten ezaugarri lexikoak zeintzuk diren jakitea zaila ez den arren, bertsolari kopuru bakoitzeko izandako arrakasta 4.2 grafikoan biltzen da.

4.2 grafikoan 4.3 ataleko saio bakoitzean *Test* corpusaren gainean ezaugarri linguistiko bakoitzeko egindako esperimentu guztien batezbestekoa irudikatu da. Atal honen hasieran aipatu bezala, hitz ez-estandarrak karaktere mailako n-gramekin konbinatzeak duen arrakasta agerikoa da grafiko horretan. Bestalde, 4.1 grafikoan ondorioztatu dugun moduan, bertsolari kopurua handitzeak problemaren konplexutasuna ere handitzen duela garbi ikus daiteke.

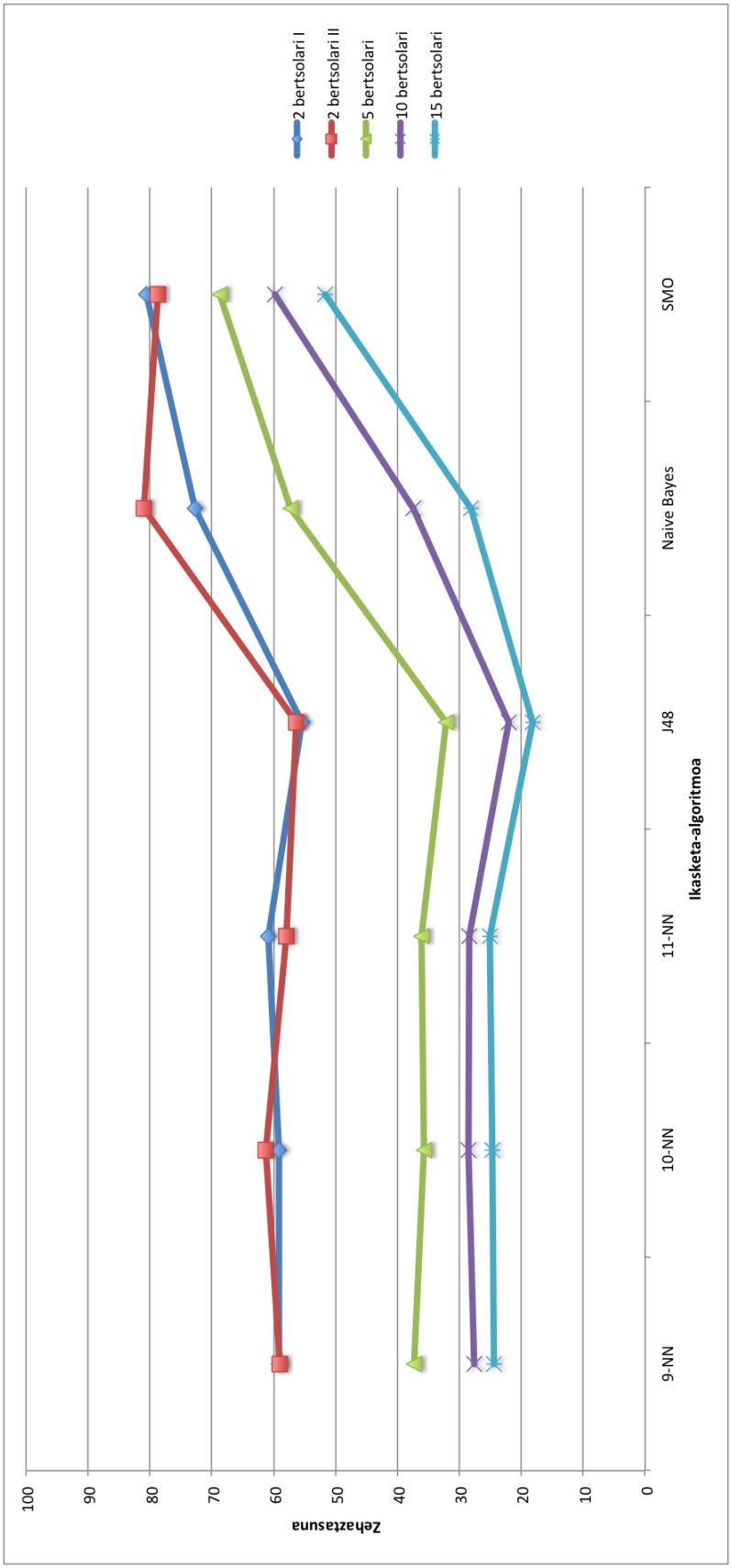
Honez gain, erabili ditugun ikasketa-algoritmoak ere aldera ditzakegu. Horrela, 4.3 ataleko saio bakoitzean *Test* corpusaren gainean ikasketa-algoritmo bakoitzeko egindako esperimentu guztien batezbestekoa kalkulatu, eta emaitzak 4.3 grafikoan irudikatu dira.

Grafiko horretan ikus daitekeen moduan, bi bertsolarirekin egindako bigarren esperimentuan izan ezik, *SMO* da nabarmen ikasketa-algoritmorik arrakastatsuena. Bestalde, *J48* da emaitza txarrenak ematen dituena. Kasu honetan ere, agerikoa da bertsolari kopuruak emaitzetan duen eragina.

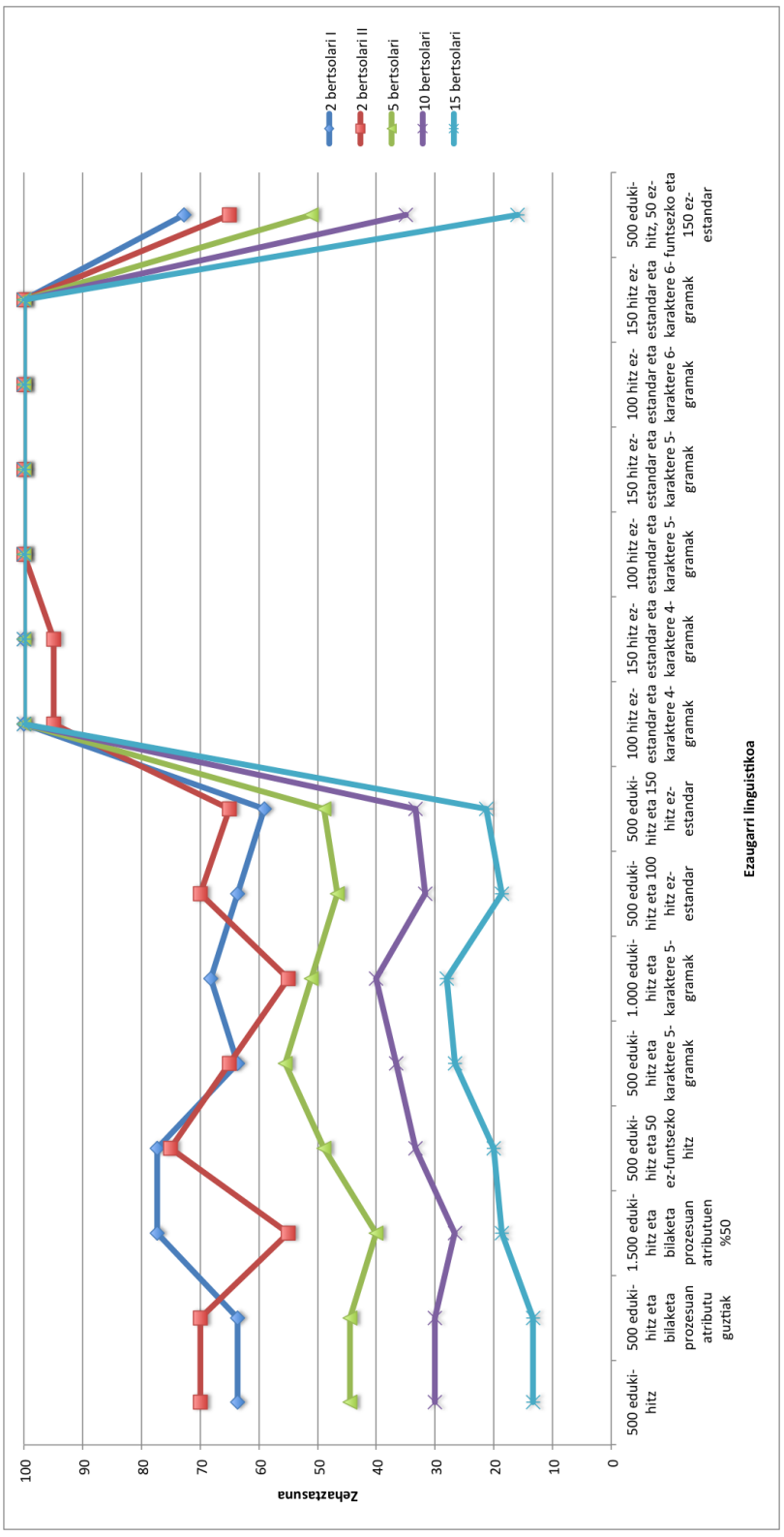
Horrela, oro har emaitza onenak ematen dituen ikasketa-algoritmoa *SMO* dela kontuan hartuz, ezaugarri linguistiko bakoitzeko algoritmo horrekin lortutako emaitzak 4.4 grafikoan irudikatu ditugu.




**4.2 Irudia:** *Egindako saio bakoitzean ezaugarri linguistikoen arrakasta*




**4.3 Irudia:** *Egindako saio bakoitzean erabilitako ikasketa-algoritmoen arrakasta*




**4.4 Irudia:** *Egindako saio bakoitzean SMO ikasketa-algoritmoarekin lortutako emaitzak*

Aipatutako azken grafikoan ikus dezakegunez, 100 edo 150 hitz ez-estandar karaktere mailako 4-grama, 5-grama eta 6-gramekin konbinatzeak emaitza bikainak ematen ditu. Hala ere, kontuan hartu behar da grafikoetan azken probetako emaitzak daudela, eta hauek, 4.3.1 azpiatalean egindako hasierako saioetan garatutako sailkatzaile guztien artean onenak direla. Horregatik, hitz ez-estandar eta karaktere mailako n-gramez gain, eduki-hitzek emandako emaitzak ere aipatzekoak dira.

Gauzak horrela, hitz ez-estandarrak eta karaktere mailako n-gramak elkar konbinatuz eta *SMO* erabiliz, bertso bat emanda bertsolari probableena zein den ebazten duen sistema bat garatu dugu, eta ikus daitekeen moduan, emaitza oso onak ematen dituena. Hala ere, garatutako sailkatzaileen arrakastaren arrazoiak, 5. atalean sakonago azalduko dugun moduan, instantzia-kopurua edota bertso-sortak izan daitezke. Corpusean aurki ditzakegun bertso kopurua baliteke nahikoa ez izatea. Honez gain, *Train* eta *Test* corpusetako bertsoetan sorta bereko bertsoak edukitzea ere izan daiteke garatutako sailkatzaileen arrakastaren arrazoia.

# 5. KAPITULUA

# Ondorioak eta etorkizuneko lanak

Proiektu honetan egile-esleipeneko ikerketa bat garatu da, zehazki bertsolariei dagokiena. Horretarako, bertsokera adierazi nahian, hainbat ezaugarri linguistiko banaka erabiltzeaz gain, horien konbinaketak ere egin dira, horrelako lanetan sarri emaitza onak ematen baituzte. Bertsoetatik ezaugarri hauek erauzita, zenbakizko bektore bidez bertsoak errepresentatu, eta hainbat ikasketa-algoritmo baliatu dira bertsoei egile zuzena esleituko dien sailkatzailea garatzeko.

Gauzak horrela, kapitulu honetan proiektuan garatutakoaren laburpena, bildutako ondorioak, lortutako ekarpenak eta etorkizuneko lanetan garatu ditzakegun zenbait ideia aurkezten dira.

## 5.1 Ondorioak

Proiektu honetan egile-esleipenaren arloko ikerketa bat egin da. Alor honetan azken hamarkadan egin diren aurrerapenak aztertu eta gero, hauek euskarara eta, zehazki, bertsoen mundura ekarri nahi izan dira. Horretarako, molde honetako lanetan ohikoak diren ezaugarri linguistikoez gain, estiloaren ñabardurak hobeki adierazteko asmoz, bertsolari baten bertsokera modu finagoan adieraz lezaketen ezaugarri bereziak –berariazko ezaugarriak– ere kontuan hartu dira. Horrela, ezaugarri horiekin guztiekin eta hainbat ikasketa-algoritmorekin saio ugari egin ditugu.

Lehenengo, corpusean bertso gehien dituzten bi bertsolarien bertsoez baliatuz, aha-

lik eta sailkatzaile fidagarrienak garatu ditugu. Honen ostean, sailkatzaile hauek ebaluatu ditugu hainbat bertsolari kopururekin. Horrela, garatutako bertsolari-esleipen ereduen kalitatea bermatu dugu, emaitza onak lortuz.

Atributuei dagokionez, bertsokera adierazteko egokienak ondorengoak izan dira, atalka:

- **Eduki-hitzak**
  - 500 eduki-hitz.
  - 500 eduki-hitz eta atributu-aukeraketako bilaketa prozesuan atributu guztiak kontuan hartuta.
  - 1.500 eduki-hitz eta atributu-aukeraketako bilaketa prozesuan atributuen %50a kontuan hartuta.
- **Eduki-hitzak + Ez-funtsezko hitzak**
  - 500 eduki-hitz eta 50 ez-funtsezko hitz.
- **Eduki-hitzak + Karaktere mailako n-gramak**
  - 500 eduki-hitz eta n=5.
  - 1.000 eduki-hitz eta n=5.
- **Eduki-hitzak + Hitz ez-estandarrak**
  - 500 eduki-hitz eta 100 hitz ez-estandar.
  - 500 eduki-hitz eta 150 hitz ez-estandar.
- **Hitz ez-estandarrak + Karaktere mailako n-gramak**
  - 100 hitz ez-estandar eta n=4.
  - 150 hitz ez-estandar eta n=4.
  - 100 hitz ez-estandar eta n=5.
  - 150 hitz ez-estandar eta n=5.
  - 100 hitz ez-estandar eta n=6.
  - 150 hitz ez-estandar eta n=6.
- **Eduki-hitzak + Ez-funtsezko hitzak + Hitz ez-estandarrak**

– 500 eduki-hitz, 50 ez-funtsezko hitz eta 150 hitz ez-estandar.

Ikus dezakegunez, gehien erabili direnak eduki-hitzak izan dira. Baina 3. kapituluan aipatu dugun moduan, edukia edo esanahia duten hitz hauek ezegonkorrak dira eta beste hitzekin ordezka daitezke, horrela, estiloa adierazi baino gehiago edukiaren informazioa eskainiz. Orduan zergatik eman dizkigute emaitza onak? Baliteke, arazoa bertso-sortetan egotea; hau da, bertsolariek jarraian gai bereko bertso bat baino gehiago botatzen dituztenez, bertso horietan hitz berdinak erabili eta bertso hauek ausaz corpus ezberdinetan banatzea (*Train*, *Develop* eta *Test* corpusetan, alegia) izan daiteke eduki-hitzen arrakastaren arrazoia. Hau da, bertso-sortak banatzen ditugunez, eta bertso-sorta bakoitzeko bertso guztiak gai berekoak direnez, gerta daiteke *Train* corpuseko bertso batean agertzen den hitzen bat edo gehiago, *Develop* eta *Test* corpusean ere agertzea. Adibidez, "hondartza" hitza agertzea *Train* corpusean, eta baita *Develop* edo *Test* corpusean ere. Halako agerpen errepikatuek argiegi eragin dezakete sailkatzailean.

Hala ere, guztien gainetik gailendu den sailkatzailea ezbairik gabe hitz ez-estandarrak karaktere mailako n-gramekin konbinatuz garatutakoa izan da. *SMO* algoritmoa erabiliz, emaitza harrigarriak eskaintzen dizkigu, egindako ia proba guztietan bertso guztiak zuzen sailkatuz. [Keselj et al., 2003]-en lana aztertuz, bagenekien karaktere mailako n-gramek emaitza oso onak ematen zituztela ingelesa, greziera edo txinera moduko hizkuntzetan, adibidez. Ondorioz, euskaraz ere emaitza onak eman zitzaketela ondoriozta genezakeen. 3. kapituluan ikusi dugun moduan, nahiz eta urteen poderioz Bertsolari Txapelketa Nagusietako bertsoetan euskara batuan ez dauden hitzen ehunekoa gutxitu, euskalkien aniztasun eta aberastasuna dela eta, bertsokera adierazteko ezaugarri onak zirela ere pentsa genezakeen. Gauzak horrela, banaka bertsokera adierazle onak izanda ere, bien arteko konbinaketarekin garatutako esleipen-ereduak eskainitako emaitza bikainak aurreikustea ezinezkoa da, eta ondorioz, bertsoei bertsolari egokia esleituko dien sailkatzaile fidagarria sortu dugula dirudi.

Emaitza hauekin harrituta, akatsen bat burutu dugun egiaztatu nahi izan dugu, eta proba ugari egin ditugu. Besteak beste, *Test* corpuseko bertsoak *Train* corpusean aurki ditzakegun begiratu dugu, baina ausazko *Train/Develop/Test* banaketa zuzena dela ikusi dugu. Akatsik ez badago, zein izan daiteke sailkatzaile hauen arrakastaren arrazoia? Batetik, instantzia gutxi ditugu. 10 bertsolarirekin adibidez, *Test* corpusean bertsolari bakoitzaren 6 bertso besterik ez ditugu, eta 15 bertsolarirekin bertsolari bakoitzaren 5 bertso. Horrela, baliteke sistemaren gaitasuna zehazteko erabili ditugun bertso kopurua nahikoa ez izatea. Bestetik, eduki-hitzekin aipatu dugun moduan, arrazoia bertso-sortak banatzea izan

daiteke, eta *Train* eta *Test* corpusetako bertsoetan sorta bereko bertsoak edukitzea. *Train* eta *Test* corpusetako egile bereko bertsoetan antzeko hitz ez-estandar eta karaktere mailako n-gramak erabiltzea ekar dezake horrek. Beraz, komenigarria litzateke lan honetan bertso-sortak banatzeak izan duen eragina aztertzea.

Emaitza onak ematen dituzten ezaugarriez gain, errimek eta oinek emandako emaitza kaxkarrak ere hausnartzekoak dira. Dudarik gabe, bertsoen ezaugarririk garrantzitsuenetakoak dira, eta hori dela eta, harrigarria da bertsokera adierazteko duten gaitasun eza. Oinak eta errimak bertsolarien artean "demokratizatuta" daudela adieraz lezake horrek.

Egile-esleipen lan gehienetan, ez-funtsezko hitzek egilearen estiloa adierazteko gaitasuna nabarmentzen da. Gure lanean, nahiz eta emaitzak txarrak ere ez izan, ez dira espero genuen bezain onak izatera iritsi. Arrazoiak hainbat izan daitezke. Baliteke bertsoetan ez-funtsezko hitzen presentzia mugatuagoa edota uniformeagoa izatea, prosako testuekin alderatuta. Gainera, aipagarria da, oro har ikasketa-algoritmorik txarrena izan arren, *J48* izan dela ezaugarri hauekin emaitza onenak eman dituen algoritmoa.

Ezaugarri linguistikoak alde batera utziz, egile-esleipenean oraindik ere galdera ugari erantzunik gabe daude. Zein da testuaren luzera egokia estiloa behar bezala adierazteko? Nola banatu daitezke egilea, generoa eta gaiaren informazioa ematen dizkiguten estiloaren ezaugarriak? Ondorioz, arlo honetan dauden erredundantzia eta metodologia irregulartasunak direla eta, egile-esleipenen ikasketak kritika mordoa jasan ditu. Hala ere, azken hamarkadan aurrerapen handiak egin dira, edozein lekutan aurki ditzakegun testuak identifikatzeko gai izatera iritsi arte. Gainera, lortzen diren emaitzak zehaztasun handikoak izan ohi dira. Ikasketa automatikoan, informazioaren berreskuratzean eta hizkuntzaren prozesamenduan eginiko garapenak izan dira aurrerapen horien eragile. Bestalde, aipatu beharra dago Interneti esker gero eta testu gehiago ditugula eskuragarri, eta ondorioz, teknologia honen erabilera gero eta handiagoa dela. Horrela, arlo honetan sakonduz, pixkanaka emaitzak hobetzeko aukera ikusten da.

## 5.2 Etorkizuneko lanak

Jarraian, etorkizuneko lanak izan daitezkeen ideia batzuk planteatuko dira, proiektu honen inguruan ikertzen jarraitzeko asmoarekin:

- Lan honetan eduki-hitzek eta hitz ez-estandarrak karaktere mailako n-gramekin

konbinatuz izan duten arrakasta bermatzeko, probak errepikatu baina *Train/Develop/Test* banaketan bertso-sortak zatitu gabe.

- Sailkatzailea berregin, egile ezezaguna duten bertso-sortak identifikatzen saiatzeko; hala nola, "Hamalau heriotzenak"[1] bertso-sortaren egilea zein den esatera garamatzan hipotesi bat formulatzeko, adibidez.

- Bertsoak alde batera utzi, eta egile-esleipena euskarazko beste zenbait arlotara hedatu, hala nola, eleberrietara edo artikuluetara.

[1] Bertso-sortarik famatu bezain ezagunenetakoa. Gaur egun ere oraindik, ez dakigu XIX. mendeko bertso-sorta honen egilea zein izan zen.

# Eranskinak

# A. ERANSKINA

## Jarraipen eta kontrola

Eranskin honetan proiektu honetarako buruturiko jarraipen txostena aurkezten da.

### A.1 Helburuak eta betekizunak

Proiektuaren helburuak ez du inongo aldaketarik jasan, eta egileen esleipenean egin diren aurrerapenak aztertu eta bertsoen mundura ekartzea izaten jarraitu du.

Betekizunei dagokienez, egile-esleipena bertsolaritza mundura ekarri da, eta gaur egungo edozein bertsoren egile probableena zein den esango digun sailkatzailea garatu dugu. Honez gain, bertsokera adierazteko atributu egokienak hitz ez-estandarren eta karaktere mailako n-gramen arteko konbinaketak direla aurkitu dugu. Emaitza onenak ematen dituen ikasketa-algoritmoa *SMO* dela ere argi geratu da. Azkenik, hainbat egile kopururekin egindako saioekin garatutako sistemaren gaitasuna frogatu dugu.

### A.2 Emangarriak

Proiektuko emangarrietan ez da inolako aldaketarik eman, eta bere horretan mantendu dira.

## A.3 Lanaren deskonposaketa egitura (LDE)

LDEan ere ez da inolako aldaketarik eman. Agian, egindako saioetan lan pakete berri bat gehitu dezakegu D eranskinean egindako esperimentuekin.

## A.4 Proiektuaren atazak

A.1 atalean azaldu dugun moduan, helburu eta betekizunetan gorabeherarik izan ez denez eta LDE diagrama bere horretan mantendu denez, proiektuaren garapenerako atazen zehaztapenak ez du inolako aldaketarik jasan.

Hala ere, plangintza egitean ataza bakoitzari esleitutako denbora estimazioek proiektua garatzerakoan dedikatutakoarekin desbideraketak izan dituzte. Hori dela eta, A.1 taulan atazen hasierako denbora-estimazioa, dedikazioa eta estimaziotik dedikaziora izandako desbideraketak laburtzen dira.

| **Ataza** | **Estimatutako denbora (orduak)** | **Dedikatutako denbora (orduak)** | **Desbideraketa (orduak)** |
|---|---|---|---|
| Plangintza | 25 | 20 | -5 |
| Aurrekarien azterketa | 120 | 105 | -15 |
| Produktuaren garapena | 150 | 180 | 30 |
| Esperimentuen prestaketa | 40 | 50 | 10 |
| Egindako saioak | 110 | 130 | 20 |
| Memoria | 80 | 90 | 10 |
| Aurkezpena | 10 | 10 | 0 |
| Jarraipen eta kontrola | 30 | 30 | 0 |
| **Guztira** | **415** | **435** | **20** |

**A.1 Taula:** *Atazen denbora estimazioa, dedikazioa eta desbideraketa*

Taulan ikus daitekeen moduan, desbideraketa batzuk positiboak (dedikazia estimazioa baino handiagoa izan denean) eta beste batzuk negatiboak (estimazioa dedikazioa baino

handiagoa izan denean) izan dira. Horregatik, guztira kalkulatu dugun desbideraketa, desbideraketa "totala" dela esan dezakegu. Aldiz, desbideraketen balio absolutua batuz gero, 60 orduko desbideraketa "akumulatu" dugula esan genezake.

## A.5 Mugarriak

Mugarriei dagokienez, barne mugarrietan atzerapen bat izan dugu. Esperimentuen prestaketei eta egindako saioei dedikatutako denbora estimatutako baino gehiago izan denez, proba guztiak ezin izan ziren planifikatutako egunerako amaitu. Hori dela eta, probei dagokion barne mugarri hau maiatzaren 30etik ekainaren 15era atzeratu zen.

## A.6 Gantt-diagrama

A.4 atalean ikusitako ataza eta azpiatazek egutegian izan duten kokapenaren adierazpen grafikoa ikus dezakegu A.1 irudiko Gantt-diagramaren bitartez.

Gantt-diagrama erreala estimatutakoarekin alderatuta, bistakoa da proiektuaren garapenean estimatutako denbora baino gehiago dedikatzeak ataza eta azpiataza horientzat egutegiaren kokapenean izandako eragina.

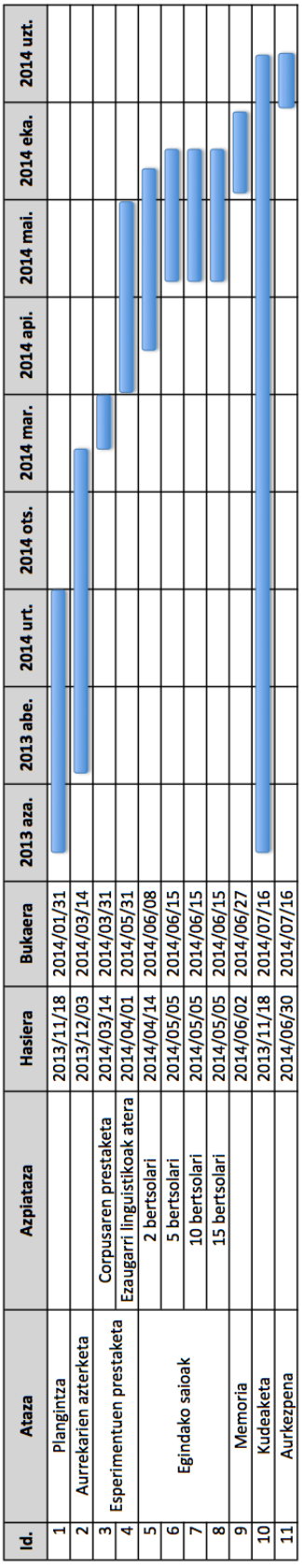
## A.7 Kronograma

A.5 atalean mugarrietan izandako aldaketarekin eta A.6 atalean aurkeztutako Gantt-diagramarekin, atazek elkar izan duten eragina erakutsi nahian A.2 irudiko kronograma erreala eraiki da.

Bertan ikus dezakegunez, desbideraketak direla eta, maiatzeko eta ekaineko lan-karga estimatutako baino handiagoa izan da.

## A.8 Lan-metodologia

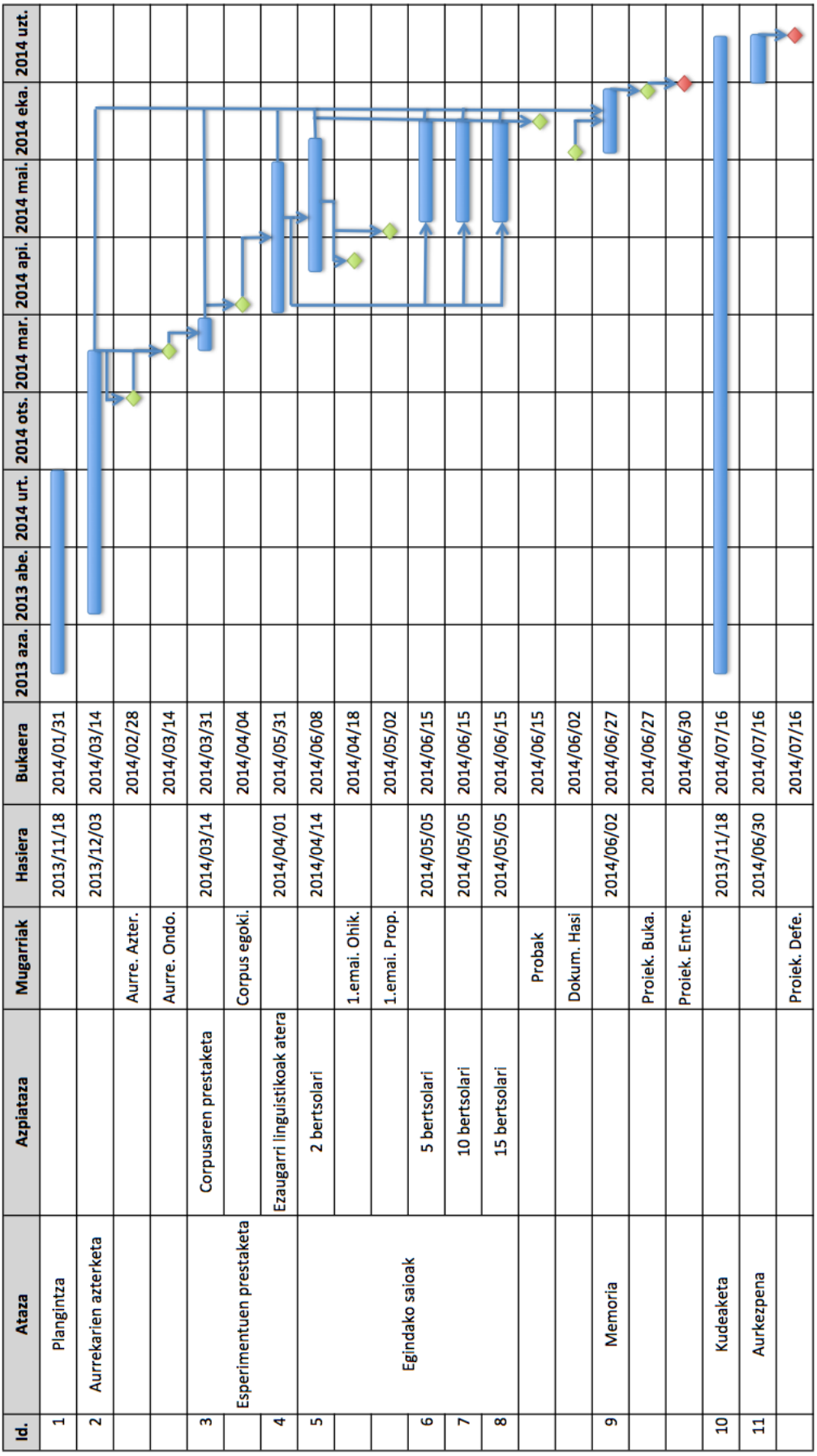
Planifikatutako moduan, bilerak barne mugarri bat bukatu orduko burutu dira. Hala ere, bilera horiez gain beste zenbait bilera ere egin direla aipatu beharra dago, proiektuari jarraipena egin eta azaldutako dudak hausnartu eta argitzeko.



| Id. | Ataza | Azpiataza | Hasiera | Bukaera |
|---|---|---|---|---|
| 1 | Plangintza | | 2013/11/18 | 2014/01/31 |
| 2 | Aurrekarien azterketa | | 2013/12/03 | 2014/03/14 |
| 3 | Esperimentuen prestaketa | Corpusaren prestaketa | 2014/03/14 | 2014/03/31 |
| 4 | | Ezaugarri linguistikoak atera | 2014/04/01 | 2014/05/31 |
| 5 | Egindako saioak | 2 bertsolari | 2014/04/14 | 2014/06/08 |
| 6 | | 5 bertsolari | 2014/05/05 | 2014/06/15 |
| 7 | | 10 bertsolari | 2014/05/05 | 2014/06/15 |
| 8 | | 15 bertsolari | 2014/05/05 | 2014/06/15 |
| 9 | Memoria | | 2014/06/02 | 2014/06/27 |
| 10 | Kudeaketa | | 2013/11/18 | 2014/07/16 |
| 11 | Aurkezpena | | 2014/06/30 | 2014/07/16 |



**A.1 Irudia:** *Gantt-diagrama erreala*




**A.2 Irudia:** *Kronograma erreala*

Planifikatutako epe eta ordutegiei dagokienez, A.5 atalean aipatutako barne mugarriaren atzeraketa da aipatzeko bakarra.

## A.9 Arriskuen jarraipena

Plangintzan proiektuan zehar atazen bat planifikatutako eperako burutzeko ezintasuna aurreikusi zen. Proiektuaren garapenari estimatutakoa baino denbora gehiago dedikatzeak, probei zegokien barne mugarri atzeratzea eragin zuen. Atzerapen hori planifikatutako kontingentzia-planari esker burutu ahal izan genuen, diseinatutako planifikazio malguari esker lan-programa egokitzeko aukera izan baikenuen.

# B. ERANSKINA

## Corpusa egokitzeko programen gida

Lan honetarako 1989. urtetik aurrera egindako Bertsolari Txapelketa Nagusietako bertsoak erabili ditugu. Dakigunez, bertso hauen corpusak *IXA* taldeko zerbitzarian daude eskuragarri. 4. kapituluan aipatu bezala, bertan corpusekin lan egiteko hainbat aukera ditugu. Gure kasuan, *corpusaAnalizatutaEzestXXXX.xml* corpusekin egingo dugu lan, 1989, 1993, 1997, 2001, 2005 eta 2009 urtekoekin hain zuzen ere, 2013koa ez baitago eskuragarri.

Corpus horietan, bertsoak bertsoka sailkatuta daude, bertso bakoitza dagokion informazioarekin (kategoria, bertsolaria, gaia, neurria...). Informazio horiez gain, transkripzioa daukagu, bertso bakarraz edo hainbat bertsoz osatua egon daitekeena. Bertsoak lerroz osatuak daude, eta lerroak analizatutako hitzez.

Proiektu honetarako, lehenengo, behar ditugun urteetako corpusak kateatuz *corpusaAnalizatutaEzestOsoa.xml* izeneko corpus bakarra osatuko dugu. Ondoren, corpus horretatik abiatuz, bertsoak bertsolarika sailkatuko ditugu, gure lanerako egokiagoa izango baita. Horretarako, `corpus_sortu.pl` programa erabiliko dugu, eta gure corpus berria zenbat bertsolariren bertsoez osatuta egongo den zehaztu beharko dugu. Aukeratuko diren bertsolariak, bertso gehien dituztenak izango dira.

```
$ perl corpus_sortu.pl [bertsolari kopurua]
```

`corpus_sortu.pl` programa exekutatuz, bertsoak bertsolarika sailkatutako *Corpus.xml* fitxategia sortzeaz gain, aukeratutako bertsolariek *corpusaAnalizatutaEzestOsoa.xml* corpusean dituzten bertsoen kopuruak azalduko dira, eta baita sortutako corpus egokituan

edukiko dituzten bertso eta tokenen kopurua ere. Bost bertsolarirekin adibidez, ondorengo irteera edukiko genuke:

```
EGAÑA-ANDONI: 127
ITURRIAGA-UNAI: 110
LUJANBIO-MAIALEN: 95
MAIA-JON: 91
MENDILUZE-AITOR: 89
-----------------------------------------
EGAÑA-ANDONI: 4187 token / 89 bertso
ITURRIAGA-UNAI: 4129 token / 89 bertso
LUJANBIO-MAIALEN: 4018 token / 89 bertso
MAIA-JON: 4286 token / 89 bertso
MENDILUZE-AITOR: 3528 token / 89 bertso
```

Kontuan hartu behar da, 3. kapituluan ikusi dugun moduan egile-esleipeneko lanetan corpusen banaketa orekatua izan ohi denez, bertsolari guztiek bertso-kopuru bera edukiko dutela.

Aipatutako *Corpus.xml* fitxategia sortu ondoren, hurrengo pausoa *Train/Develop/Test* banaketa ausaz osatzea izango da. *Train* corpusa bertsoen %80arekin egongo da osatuta, eta honekin bertsolarien bertsokera edo estiloa ikasiko dugu. *Develop* corpusa bertsoen %10arekin osatuko da, eta honekin, ahalik eta sailkatzaile onena garatzen saiatuko gara emaitzarik onenak bilduz. Azkenik, Test corpusa ere bertsoen %10arekin egongo da osatuta, eta honekin, sailkatzaile onenak garatuta ditugunean, azken probak egingo ditugu sistemaren amaierako zehaztasuna zein den ikusteko. Banaketa hau egiteko, `train-develop-test.pl` programa daukagu. Sarrera gisa bertsolari kopurua eta bertsolari hauek *Corpus.xml* fitxategian dituzten bertsoen kopurua (aurreko adibidean 89 izango litzateke) eman beharko dizkiogu.

```
$ perl train-develop-test.pl [bertsolari kopurua] [bertsoen kopurua]
```

Bost bertsolariren adibidearekin jarraituz, *Train.xml*, *Develop.xml* eta *Test.xml* fitxategiak sortzeaz gain, fitxategi bakoitzeko aurreko bertsolarien bertso eta token kopuruak azaltzen dituen ondorengo irteera edukiko genuke, bertsolarien aurreko ordena jarraituz:

```
TEST / DEVELOP / TRAIN    (token-bertso)
-----------------------------------------------
396 - 9 / 443 - 9 / 3348 - 71
488 - 9 / 401 - 9 / 3240 - 71
383 - 9 / 496 - 9 / 3139 - 71
433 - 9 / 478 - 9 / 3375 - 71
278 - 9 / 339 - 9 / 2911 - 71
```

# C. ERANSKINA

## Ezaugarri linguistikoak lortzeko programen gida

Jarraian, *Train.xml*, *Develop.xml* eta *Test.xml* fitxategietatik ezaugarri lexikoak, karaktere-ezaugarriak eta berariazko ezaugarri deiturikoak lortzeko programak nola exekutatu azalduko dugu. Ondoren, konbinatutako ezaugarri linguistikoak nola lortu azalduko dugu. Fitxategi horiek, exekutatuko ditugun programen direktorio berean egon beharko dute. Programa hauekin, aipatutako fitxategi bakoitzerako *ARFF* fitxategiak sortuko ditugu, ondoren *Weka* erabiliz zenbait ikasketa-algoritmo lan egiteko, eta bertsoei bertsolariak esleituko dizkien sailkatzaileak garatzeko.

### C.1 Ezaugarri lexikoak

Eduki-hitzak lortzeko, programekin batera aurki dezakegun *content-words.txt* izeneko euskarazko eduki-hitzen zerrenda katalogo berean eduki beharko dugu.

```
$ perl ContentWord.pl [bertsolari kopurua] [eduki-hitzen kopurua]
```

Ez-funtsezko hitzak lortzeko ere, programekin batera aurki dezakegun *funtzio-hitzakOna.txt* izeneko euskarazko ez-funtsezko hitzen zerrenda katalogo berean eduki beharko dugu.

```
$ perl FunctionWord.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua]
```

Hitz mailako n-gramak lortzeko, bertsolari kopuruaz gain n-ren balioa ere zehaztu beharko dugu.

```
$ perl Ngramak-hitzak.pl [bertsolari kopurua] [n-ren balioa]
```

## C.2 Karaktere-ezaugarriak

Karaktere mailako n-gramak lortzeko, hitz mailako n-gramekin egiten dugun moduan, bertsolari kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl Ngramak.pl [bertsolari kopurua] [n-ren balioa]
```

## C.3 Berariazko ezaugarriak

Hitz ez-estandarrak lortzeko, bertsolari kopurua eta hitz ez-estandarren kopurua zehaztuko ditugu.

```
$ perl HitzEzEstandarrak.pl [bertsolari kopurua]
[hitz ez-estandarren kopurua]
```

Errimak eta oinak lortzeko, bertsolari kopurua zehazteaz gain, programaren katalogo berean programekin batera aurki dezakegun errimak eskuratzeko *errimaPatroia.fst* automata edukitzea beharrezko da.

```
$ perl Errimak.pl [bertsolari kopurua]
$ perl Oinak.pl [bertsolari kopurua]
```

## C.4 Ezaugarriak konbinatuta

Jarraian, emaitza onenak eman dituzten ezaugarri linguistikoak konbinatu eta lortzeko programak nola exekutatu azaltzen da.

Eduki-hitzak eta ez-funtsezko hitzak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua eta ez-funtsezko hitzen kopurua zehaztu beharko ditugu.

```
$ perl FW-CW.pl [bertsolari kopurua] [ez-funtsezko hitzen kopurua]
[eduki-hitzen kopurua]
```

Eduki-hitzak eta hitz ez-estandarrak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua eta hitz ez-estandarren kopurua zehaztu beharko ditugu.

```
$ perl CW-HitzEzEstandarrak.pl [bertsolari kopurua]
[eduki-hitzen kopurua] [hitz ez-estandarren kopurua]
```

Eduki-hitzak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl CW-Ngramak.pl [bertsolari kopurua] [eduki-hitzen kopurua]
[n-ren balioa]
```

Ez-funtsezko hitzak eta hitz ez-estandarrak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, ez-funtsezko hitzen kopurua eta hitz ez-estandarren kopurua zehaztu beharko ditugu.

```
$ perl FW-HitzEzEstandarrak.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua] [hitz ez-estandarren kopurua]
```

Ez-funtsezko hitzak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, ez-funtsezko hitzen kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl FW-Ngramak.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua] [n-ren balioa]
```

Hitz ez-estandarrak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, hitz ez-estandarren kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl HitzEzEstandarrak-Ngramak.pl [bertsolari kopurua]
[hitz ez-estandarren kopurua] [n-ren balioa]
```

Eduki-hitzak, ez-funtsezko hitzak eta hitz ez-estandarrak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua, ez-funtsezko hitzen kopurua eta hitz ez-estandarren kopurua zehaztu beharko ditugu.

```
$ perl FW-CW-HitzEzEstandarrak.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua] [eduki-hitzen kopurua]
[hitz ez-estandarren kopurua]
```

Eduki-hitzak, ez-funtsezko hitzak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua, ez-funtsezko hitzen kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl FW-CW-Ngramak.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua] [eduki-hitzen kopurua] [n-ren balioa]
```

Eduki-hitzak, hitz ez-estandarrak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, eduki-hitzen kopurua, hitz ez-estandarren kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl CW-Ngramak-HitzEzEstandarrak.pl [bertsolari kopurua]
[eduki-hitzen kopurua] [n-ren balioa] [hitz ez-estandarren kopurua]
```

Ez-funtsezko hitzak, hitz ez-estandarrak eta karaktere mailako n-gramak konbinatzen dituen programa exekutatzeko, bertsolari kopurua, ez-funtsezko hitzen kopurua, hitz ez-estandarren kopurua eta n-ren balioa zehaztu beharko ditugu.

```
$ perl CW-Ngramak-HitzEzEstandarrak.pl [bertsolari kopurua]
[ez-funtsezko hitzen kopurua] [n-ren balioa]
[hitz ez-estandarren kopurua]
```

# D. ERANSKINA

## Beste zenbait saio

Proiektuan egindako saioetako emaitzei erreparatuta, 100 edo 150 hitz ez-estandar karaktere mailako 4-grama eta 5-gramekin konbinatzeak *SMO* ikasketa-algoritmoarekin emaitza oso onak ematen dituela argi ikus daiteke. Hori izan da garatzea lortu dugun bertsolari-esleipen eredu egokiena. Garatutako sistema honen kalitatea bermatzeko, D.1, D.2 eta D.3 tauletan bildu ditugun beste zenbait proba egin ditugu aipatutako ezaugarri linguistikoen konbinaketak eta *SMO* ikasketa-algoritmoa erabiliz. Emaitzak, proiektuko saio guztietan bezala, zuzen sailkatutako bertsoen ehunekoak izango dira.

| Bertso-lari kopurua | 100 hitz ez-estandar eta 4-gramak | 150 hitz ez-estandar eta 4-gramak | 100 hitz ez-estandar eta 5-gramak | 150 hitz ez-estandar eta 5-gramak |
|---|---|---|---|---|
| 2 | 100 | 100 | 100 | 100 |
| 3 | 100 | 100 | 100 | 100 |
| 4 | 100 | 100 | 100 | 100 |
| 5 | 100 | 100 | 100 | 100 |
| 6 | 100 | 100 | 100 | 100 |
| 7 | 100 | 100 | 100 | 100 |
| 8 | 100 | 100 | 100 | 100 |
| 9 | 100 | 100 | 100 | 100 |
| 10 | 100 | 100 | 100 | 100 |
| 11 | 100 | 100 | 100 | 100 |
| 12 | 100 | 100 | 100 | 100 |

**D.1 Taula:** *Garatutako sailkatzaile onenekin egindako probak hainbat bertsolari kopururekin (I)*

| **Bertsolari kopurua** | **100 hitz ez-estandar eta 4-gramak** | **150 hitz ez-estandar eta 4-gramak** | **100 hitz ez-estandar eta 5-gramak** | **150 hitz ez-estandar eta 5-gramak** |
|---|---|---|---|---|
| 13 | 100 | 100 | 100 | 100 |
| 14 | 100 | 100 | 100 | 100 |
| 15 | 100 | 100 | 100 | 100 |
| 16 | 100 | 100 | 100 | 100 |
| 17 | 100 | 100 | 100 | 100 |
| 18 | 100 | 100 | 100 | 100 |
| 19 | 100 | 100 | 100 | 100 |
| 20 | 100 | 100 | 100 | 100 |
| 21 | 100 | 100 | 100 | 100 |
| 22 | 100 | 100 | 100 | 100 |
| 23 | 100 | 100 | 100 | 100 |
| 24 | 100 | 100 | 100 | 100 |
| 25 | 99 | 99 | 99 | 99 |
| 26 | 100 | 100 | 100 | 100 |
| 27 | 100 | 100 | 100 | 100 |
| 28 | 100 | 100 | 100 | 100 |
| 29 | 100 | 100 | 100 | 100 |
| 30 | 100 | 100 | 100 | 100 |
| 31 | 100 | 100 | 100 | 100 |
| 32 | 100 | 100 | 100 | 100 |
| 33 | 100 | 100 | 100 | 100 |
| 34 | 100 | 100 | 100 | 100 |
| 35 | 100 | 100 | 100 | 100 |
| 36 | 100 | 100 | 100 | 100 |
| 37 | 100 | 100 | 100 | 100 |
| 38 | 100 | 100 | 100 | 100 |
| 39 | 100 | 100 | 100 | 100 |
| 40 | 100 | 100 | 100 | 100 |
| 41 | 100 | 100 | 100 | 100 |
| 42 | 100 | 100 | 100 | 100 |
| 43 | 100 | 100 | 100 | 100 |
| 44 | 100 | 100 | 100 | 100 |
| 45 | 100 | 100 | 100 | 100 |
| 46 | 100 | 100 | 100 | 100 |
| 47 | 100 | 100 | 100 | 100 |
| 48 | 100 | 100 | 100 | 100 |
| 49 | 100 | 100 | 100 | 100 |

**D.2 Taula:** *Garatutako sailkatzaile onenekin egindako probak hainbat bertsolari kopururekin (II)*

| Bertso-lari kopurua | 100 hitz ez-estandar eta 4-gramak | 150 hitz ez-estandar eta 4-gramak | 100 hitz ez-estandar eta 5-gramak | 150 hitz ez-estandar eta 5-gramak |
|---|---|---|---|---|
| 50 | 100 | 100 | 100 | 100 |
| 51 | 100 | 100 | 100 | 100 |
| 52 | 99.04 | 99.04 | 100 | 100 |
| 53 | 100 | 100 | 100 | 100 |
| 54 | 100 | 100 | 100 | 100 |
| 55 | 100 | 100 | 100 | 100 |
| 56 | 100 | 100 | 100 | 100 |
| 57 | 99.12 | 99.12 | 100 | 100 |
| 58 | 100 | 100 | 100 | 100 |
| 59 | 100 | 100 | 100 | 100 |
| 60 | 100 | 100 | 100 | 100 |

**D.3 Taula:** *Garatutako sailkatzaile onenekin egindako probak hainbat bertsolari kopururekin (III)*

Corpusean 110 bertsolariren bertsoak ditugunez, hasiera batean asmoa probak 110 bertsolariraino egitea zen arren, ordenagailuaren memoriak jartzen dizkigun mugak direla eta, ezin izan ditugu gainontzeko esperimentuak burutu. Zenbat eta bertsolari gehiago eduki, guztira ditugun bertsoen kopurua orduan eta handiagoa da. Horrela, bertsolari kopurua handitu ahala, 4-gramen eta 5-gramen kopurua nabarmen handitzen da, eta horrek gero eta memoria gehiago eskatzen digu.

Taulak aztertuta, nahiz eta bertsolari kopurua handitu, gure sistemaren gaitasunean eraginik ez duela argi ikusten da. Hala ere, 5. kapituluan egindako hausnarketak kontuan hartzekoak dira, eta agian, bertso-sortak banatzea izango da garatutako sailkatzaileen arrakastaren eragilea. Arrazoia hori den ikusteko, probak errepikatu beharko genituzke, baina oraingoan *Train/Develop/Test* banaketan bertso-sortak zatitu gabe.

# Bibliografia

[wik, ] Euskarazko wikipedia. url: http://eu.wikipedia.org/.

[ixa, ] Ixa ikerketa taldea. url: http://ixa.si.ehu.es.

[Agirrezabal et al., 2013] Agirrezabal, M., Arrieta, B., Astigarraga, A., and Hulden, M. (2013). Bota bertsoa, eta guk aztertuko dugu: azken urteetako Bertsolari Txapelketa Nagusien analisia.

[Arrieta, 2010] Arrieta, B. (2010). *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean*. PhD thesis.

[Astigarraga et al., 2009] Astigarraga, A., Gojenola, K., Sarasola, K., and Soroa, A. (2009). *TAPE Testu-analisirako PERL erremintak*.

[Diederich et al., 2003] Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship Attribution with Support Vector Machines.

[Grieve, 2007] Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques.

[Jockers and Witten, 2010] Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution.

[Keselj et al., 2003] Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based Author Profiles for Authorship Attribution.

[Koppel et al., 2009] Koppel, M., Schler, J., and Argamon, S. (2009). Computational Methods in Authorship Attribution.

[Luyckx and Daelemans, 2008] Luyckx, K. and Daelemans, W. (2008). Authorship Attribution and Verification with Many Authors and Limited Data.

[Stamatatos, 2009] Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods.

[Yu, 2012] Yu, B. (2012). Function Words for Chinese Authorship Attribution.