Lehen urratsak bertsotarako gaien azterketa automatikoan: gaiaren erregistroaren predikzioa eta gaiaren eta bertsoaren arteko erlazio semantikoaren azterketa

Egilea: Ion Lizarazu Arbulu
Tutorea: Bertol Arrieta Kortajarena

HAP

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2014eko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Bertsolaritza eta informatikaren azterketa oinarri duen master-tesi honek bi lan hartzen ditu bere baitan. Batetik, ikasketa automatiko bidez bertsotarako gaien erregistroa aurresatea. Hala, gaia emanik, gai hori *umorezkoa* den ala ez ebazten saiatu gara. Bestetik, LSA bidezko bertsotarako emandako gaiaren eta bertsoaren arteko antzekotasun semantikoaren neurketa burutu dugu.

Abstract

This master thesis contains two works that involve both, computer scient and 'bertsolaritza'. On the one hand, we have a work where the goal is to predict the register of 'bertso' by machine learning. On the other hand, the goal is to measure the semantic relationship between the bertso and the topic given for develop the bertso.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1 Proiektuaren definizioa

Euskal Herrian betidanik izan da joera bazkari edota afari osteak kantuz alaitzeko. Kanturako ohitura horren barnean kokatzen da bertsolaritza, kantugintza inprobisatua. Lagunartekotasunean inprobisatzetik harago egin du bertsolaritzak aspaldiko urteetatik hona. Gaur egun egin ohi diren hainbat eta hainbat txapelketak, jardun zurrunekoek (Bertsolari Txapelketa Nagusia, Gipuzkoako Bertsolari Txapelketa, Bizkaiko Bertsolari Txapelketa, etab.) zein lasaiago edo alaiagokoek (Plazatik Gaztetxera, Bardoak Bertso Txapelketa, Martxoan Bertsoa, etab.), bertsolaritzaren garapenean lagundu dute. Hala ere, oinarriak betikoa izaten jarraitzen du; neurria eta doinua aukeratu, dagokion gaiari buruzko esaldiak errimatuz prestatu, eta bertsoa kantatu. Historiaren hastapenetatik gaurkotasunezko maiteminduriko begiradetarainoko errepasoa eman zion Maialen Lujanbiok bertso bitartez 'sua' gaiari:

Hura asmakizuna (7)
homo habilisena (6)
bazun intentzioa (7)
bazuen sena (5)
bi harrik elkar jota (7)
txispa bat aurrena (6)
asmakizun haundina (7)
mendeetan barrena (6)
sua da problema (6)
zenbait basorena (6)
edo jaki dena (6)
berotzen duena (6)
eta bi begiradek (7)
sortzen dutena (5)

Maialen Lujanbiok 2009ko Bertsolari Txapelketa Nagusian irabazle izan zenean kartzelako gaian 'sua'ri abestutako bertsoa dugu goian aurkeztua. Hamalauko neurrian (hamalau lerro osatu behar), txikiaren moldeko (neurri txikiek ohiko duten 7-6-7-6 egituran oinarritua) 9 puntuko bertsoa osatzeko hautua egin zuen, *Betroiarena* deituriko doinuan abestuz. Neurri horretan sartu zituen suarekiko azalpenak, lerro bakoitzeko silaba kopuru egokiekin, eta berdez koloreztatutako errimekin.

Hizkuntzaren Azterketa eta Prozesamendua masterrerako lan hau informatika eta bertsolaritza uztartzeko bidean aurkitzen da, lehendik ere esparru horretan egin diren lanei jarraiki. Bertsolaritzari informatika munduan tartea egiterako orduan, lehen pausoa bertsotarako laguntza emateko tresnen sorkuntza izan zen, hau da, bertsoaren alde teknikoa aztertu eta horren inguruko laguntza emango zuen tresneria sortzea. Bide horretan eginak dira errima-bilatzailea, Bertsotarako Arbel Digitala eta SegaPoto android mugikorretarako aplikazioa. Bertsolaritzaren arlo teknikoan laguntzeko tresnak sortzeaz gain, bestelako azterketa ere egin da, hala-nola Bertsolari Txapelketa Nagusietako finalen azterketa estatistikoa. Lan horrek Bertsolari Txapelketa Nagusiko bertsoen edo bertsokeraren bilakaera

nolakoa izan den ezagutzeko aukera eman digu. Bestalde, bertsolaritza eta informatika uztartuta egin diren lanen artean, inguruan eginiko lan garrantzitsuenetakoa bertso-sorkuntza automatikoarena da handizaleena. BertsoBot proiektuarekin hasi zen, eta zenbait bide jorratzen jarraitzen duen ingurunea da sorkuntzarena, azken batean makinak bertsoa automatikoki sortu eta kanta dezan. Azkenik, bertsoaren arlo semantikoa aztertzen ere egin da lanik; agur-bertsoetako egitura diskurtsiboa aztertzen eta bertso-sorkuntza semantikoki koherente baterako bidean.

Gure tesi honetan, berriz, bertsoaren azterketan gehiago sakondu dugu sorkuntzan baino. Izan ere, aurkezten diren bi lanak bertsolaritzarekin lotutako azterketa lanak dira, biak ala biak bertsotarako ariketekin lotuak.

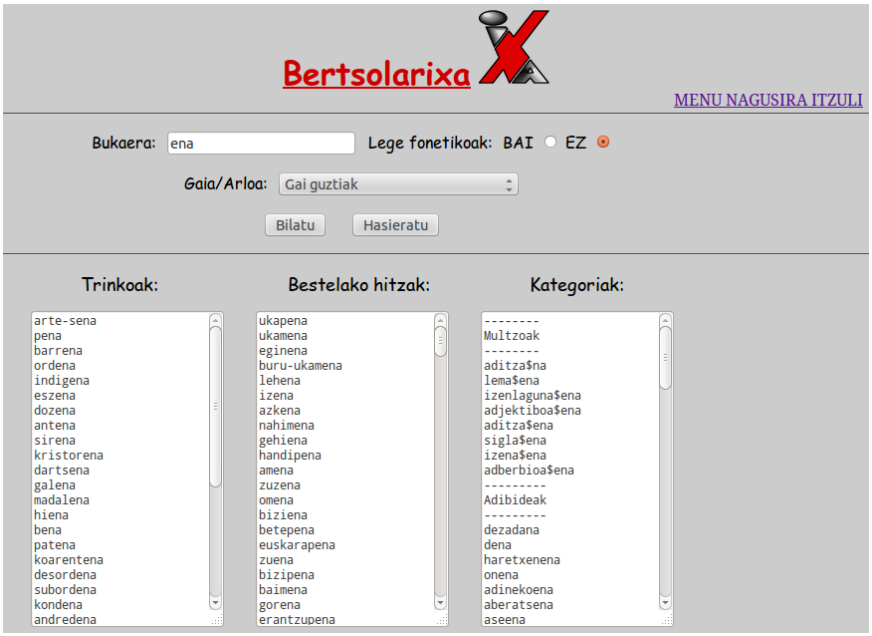
Batetik, bertsotan egiteko emandako gaia *umorezkoa* ala *dramatikoa* den automatikoki hautematean oinarritzen den lana daukagu. Ikasketa automatiko bidez eginiko lan honetarako, bertsotarako gaiez osatutako corpus etiketatu bat izan dugu esku artean. Corpus horretatik ikasi eta metodo desberdinak erabiliz gaiaren erregistroa zein den asmatzen saiatzea izan da lanaren funtsa, hau da, ematen den gaia *umorezkoa* den ala ez automatikoki ebazten saiatzea.

Bestalde, bertsoa kantatzeko emandako hainbat ariketarekin (bi motatako ariketak: hitza emanda eta gaia emanda) egin dugu bigarren lana. Gai-jartzaileak jarritako ariketa eta bertsolariaren bertsoaren arteko erlazio semantikoaren zenbatekoa lortzen saiatu gara. Erlazio semantiko hau lortzeko bidea *Latent Semantic Analysis* bidez jorratu dugu.

2 Aurrekariak

2.1 BertsolariXa

Informatika eta bertsolaritza uztartuz garaturiko lehen lana izan zen BertsolariXa (Arrieta et al., 2001). Bertsoaren oinarri den errima bilatzen laguntzeko tresna dugu hau, zeinak hitz bukaera bat eman eta horrekin errimatzen duten hitz zerrenda itzultzen duen. Hitzak itzultzeaz gain, kategoria multzo desberdinak emateko gai da sistema, modu honetan atzizkidun hitzak sortuz. Esaterako, 'ena' idazten diogunean, 'pena', 'ordena' edota 'ukapena' hitzak itzultzen dizkigu eta baita 'adjektiboa' + 'ena' edo 'izena' + 'ena' bezalako kategoria multzoak ere, dagozkien adibideekin (hobeena, alkateena).

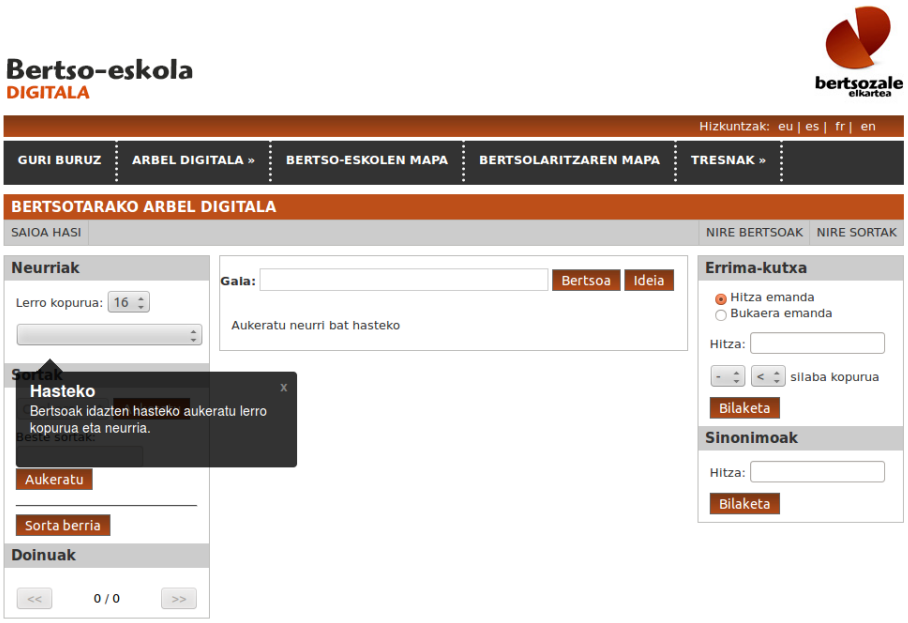


1 irudia: BertsolariXa web aplikazioaren irudia

2.2 Bertsotarako Arbel Digitala

Bertsotarako Arbel Digitala (Agirrezabal et al., 2012) (BAD) proiektua, bertsotan egiten ikasteko proiektu bezala garatu zen. Bertan lerro kopuru bat, bertsoaren neurria (zortziko txikia, hamarreko handia...) eta doinua aukeratzen dira. Momentura agertuko dira, aukeratutako lerro kopurua, eta lerro bakoitzaren alboan, sartu behar diren silaba kopurua, uneoro bistaratuz, silaba kopuruan "motz", "luze" edo "ondo" gabiltzan. Bertsoa idaztera goazenean, beti dauzkagu ideia batzuk buruan, sartu nahi ditugun hitzak etab. Horretarako aukera ematen du "ideia" botoiak, bertan taula bat daukagu, nahi duguna idazteko. Laguntza horretan sartzen dira, errima bilatzailea eta sinonimo bilatzailea. Horrez gain, bertsoa zuzendu botoian klik egin ezkero, lerro bakoitza aztertu, eta lerro bakoitzean ditugun errima akatsak eta silaba kopuru okerrak (hala bada) ezagutarazten dizkigu. Saioa

hasiko bagenu, bertsoa gorde botoian klik egin, eta (azken oharrak eman ostean) gordeta geratuko da. Ondoren, ikusi nahi baditugu, Nire Bertsoak eta Nire Sortak ataletan klik besterik ez dugu egin behar.

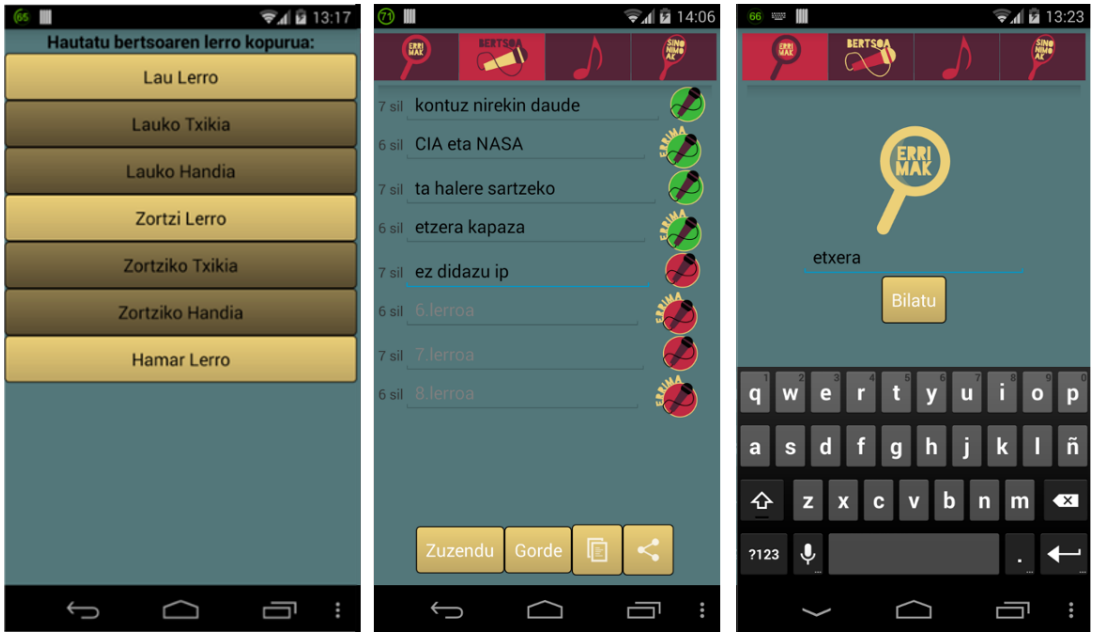


2 irudia: Bertsotarako Arbel Digitala web aplikazioaren irudia

2.3 SegaPoto

Bertsotarako Arbel Digitala (Agirrezabal et al., 2012) proiektuaren mugikorretarako hedapen gisa garatu zen. Android sistema-eragileetarako egindako aplikazioa da, kontuan hartuta Android bertsio hedatuena 2.3.x zela. Aplikazioak era lokalean funtzionalitate batzuk dauzka, baina eragiketa garrantzitsuenak egiteko sarerako konexioa beharrezkoa da. Modu honetakoak dira errima-bilatzailea (erabiltzaileak aukeratutako hitz-bukaera edo hitzarekin errimatzen duten hitzak bilatzen ditu), sinonimo-bilatzailea (erabiltzaileak pasatako hitzaren sinonimo direnak itzultzen ditu) eta bertso-zuzentzailea (bertsoa osatzean, bertsoko oinek elkarrekin errimatzen duten ala ez eta lerro bakoitzeko silaba kopurua egokia den ala ez egiaztatzen du). Sare-konexio gabe, lerro kopuru bakoitzeko neurriak aukeratu daitezke, eta honen bertsoa idazteko txantiloia ikus daiteke. Era lokalean egiten den eragiketa garrantzitsuena idazte prozesuan gauzatzen da, silaba-kontatzailea. Erabiltzailea bertsoa idazten ari denean, idatzitako letra bakoitzeko, sistemak esaldia silaba-kontatzaileari pasatzen dio, eta honek esaldiaren silaba kopuru posibleak ematen dizkio. Honela, silaba kopurua egokia bada, irudia aldatuko da, egokitasuna adieraziz. Sisteman bi erabiltzaile mota daude: erregistratu gabeko erabiltzailea eta erabiltzaile erregistratua. Erregistratu gabeko erabiltzaileak bertsoa zuzendu, errimatzen duten hitzak bilatu eta sinonimoak bilatzeko aukera izango du, betiere sarerako konexioa baldin badauka. Erregistratuta-

ko erabiltzaileak, erregistratu gabearen eragiketak egiteko aukera dauka, eta horiez gain, bertsoa gordetzeko eta aurrez gordetako bertsoak ikusteko aukerak dauzka.



3 irudia: SegaPoto android aplikazioaren irudi batzuk

2.4 BertsoBot proiektua

BertsoBot proiektuaren (Agirrezabal, 2012) helburua automatikoki bertsoak sortu eta abestuko dituen robot baterantz bidea egitea da. Hizkuntzaren prozesamenduko teknikak erabilita, poesia-sorkuntza automatikoan lehen urratsak eman dira. Hau erdiesteko, corpusen prozesamenduan oinarritutako bilaketak erabili dira, bai bilaketa arruntak eta baita bilaketa semantiko aurreratuak ere, horretarako IXA taldean garatutako hainbat tresna erabiliaz.

Bertso-sorkuntza automatikoan eginiko lanetako batean (Agirrezabal et al., 2013a), bertso corpusetik POS (Part-Of-Speech) sekuentziak erauzi eta hauen probabilitateak kalkulatzen dira. Sorkuntza prozesu horretarako hiru esperimentu desberdin prestatu ziren: corpuseko estrofan oinarrituta, batetik, hitzak ordezkatzen dira dagozkien POS etiketa eta atzizkiekin, bestetik, izen eta adjektiboak ordezkatuz, berdin jokatutako beste hitz batzuekin eta, azkenik, izenak bakarrik ordeztu semantikoki erlazionatutako beste batzuengatik. Amaitzeko, *Turing Test* gisako ebaluazio bat eginez ebaluatu ziren estrategiok, eta hirugarren bideak hobekuntza esanguratsuak lortu zituen.

2.5 Bota bertsoa eta guk aztertuko dugu

EHUko Informatika Fakultateko IXA taldean bertsoa eta informatika uztartuta egin diren lanen artean da 'Bota bertsoa eta guk aztertuko dugu' (Agirrezabal et al., 2013b). Lehen

pausoak egin badituzte ere, urrats sendoagoak egin aurretik bertsoak xehe-xehe aztertzeko saiakera egin dute, horien azterketa sakonak gerora sorkuntza hobea ekar dezakeelakoan. Azterketa horiek egiteko, Xenpelar Dokumentazio Zentroak bildu eta sailkatutako corpusa izan da oinarri. Erabilitako corpusak 1986tik 2009ra arte egindako txapelketa nagusietako bertsoak hartzen ditu; corpus hori 2.600 bertsoalditan sailkatutako 6.887 bertsok osatzen dute, eta urtean urtean datu-basean gordeta dauden bertsoaldiak (eta, ondorioz, bertsoak) geroz eta gehiago dira. Hainbat mailatan egin dute azterketa, betiere bertsoaren ezaugarri nagusiak kontuan izanda, hau da, errimak, neurriak, doinuak, hitzak, kategoria morfosintaktikoak eta euskara batuaren erabilera.

2.6 Agur-bertsoetako egitura diskurtsiboaren xerka

Bat-bateko bertso ekoizpenean hauek antolatzeko eskuarki baliatzen den egiturarik ote den antzematen saiatu ziren lan honetan (Osinalde Agirre, 2013). Agur-bertsoak aztergai hartuta bi urratsetan bereizi zen egitekoa: a) Alorreko adituen laguntzaz agurren azterketa ahal bezain zabala egin. Agur-bertsoetan ohiko ezaugarriak erauzi, bildu eta bateratuz. b) Bigarren pausoa, ezaugarri edo kategoria horien arabera testu-txatalak kategorizatzea. Osatutako ikerketaren helburua ere bikoitza izanik: batetik, euskaraz osatutako bat-bateko agur-bertsoen balizko egitura narratiboa erauzi nahi izan zen; bestetik, berriz, balizko egitura hori ikasketa-algoritmoen bidez zein adituen irizpenei jarraiki osatutako erregela linguistikoak baliatuta eskura ote litekeen ere egiaztatu nahi izan zen, orobat, bi metodologien irismena balioetsita. Bi prozedura horiek *Machine Learning (ML)* edo ikasketa automatikoaz eta erregela linguistikoek bideratutakoak ziren. Emaitzen berri jakiteko agur-bertsoetako estrofa bakoitzerako etiketa posibleak ezagutu behar dira, hala nola, mezua, lekua, publikoa, saioa, norbera eta betelana. Agur-bertsoa osatzeko kantu-molde horiek erabiltzen direla ondorioztatuta, ikasketa automatiko bidez hobekien etiketatu zirenak lekua publikoa eta betelana izan ziren, hiruak % 80tik gorako *F-Measure* emaitzekin. Mezua saioa eta norbera etiketak, aldiz % 60ko langaren bueltan geratu ziren.

2.7 Robot bertso-sortzailean koherentzia testuala

Bertso-sorkuntza automatikoan aurrerapausoak emateko bidean egindako lana (Astigarraga et al., 2014) da honakoa. Semantikoki koherenteak diren bertsoak sortzeko saiakerak azaltzen dira lanean. Oinarrian, makinari lau oin ematen zaizkio, eta honek testu-corpus batetik hitz horrekin bukatzen diren esaldiak lortzen ditu. Esaldi horiek, emandako oina azken hitz moduan izateaz gain, 13 silaba izan behar dituzte (zortziko txikiko neurrian estrofa bakoitzak duen neurria). Ondoren, lorturiko lau esaldi multzoetako bakoitzetik (oin bakoitzeko multzo bat) esaldi bat aukeratzeko hiru modutan egiten da esperimentua: ausaz (*Random*), *Standard Vector Space Model (VSM)* eta *Latent Semantic Analysis (LSA)*. Lorturiko emaitzak ebaluatzeko Bertsozale Elkartera jo eta 5 epaileri pasa zaizkio sorturiko bertsoak, koherentzia epaitzeko. Lan honen ondoren argi geratu zen erabilitako hiru moduetatik onena *LSA* dela; izan ere, epaileek koherentzia altuko puntuazioak esleitu zizkien metodo honi esker sortutako bertsoei.

3 Bertsotarako gaien erregistroaren identifikazioa ikasketa automatikoa erabiliz

Bertsotarako gai bat eman eta zein erregistrotakoa den (*umorezkoa* ala *dramatikoa*) ikasi nahi izan da lan honetan, ikasketa automatikoko tekniken bidez. Horretarako, 1986-2013 bitarteko Bertsolari Txapelketa nagusiko gaiak erabili dira. Zehatz esanda, eta umorea nola kontu oso subjektiboa den, gai bat *umorezkoa* den ala ez ebazten saiatu gara.

3.1 Esperimentuen prestakuntza

Emaitzak aztertu baino lehen, probak egin ahal izateko beharrezko prestakuntzak azaltzen hasiko gara.

3.1.1 Corpusa

Ezertan hasi aurretik, corpusaren prestakuntzan lan egin behar izan dugu. Horretarako, 1986az geroztiko Bertsolari Txapelketa Nagusiko gaiekin (Xenpelar Dokumentazio Zentroak bildutakoekin) osatutako XML fitxategi bat izan dugu abiapuntu. XML fitxategia ikasketa automatikorako erabilgarri izateko prozesuan, gai gisa aurkezten ziren hainbat kendu egin behar izan ditugu corpusa gure betebeharretara egokitzeko: gaiaren ordez puntua ematen zeneko kasua, oinak ematen zirenekoa, errepikatutako gaiak (kartzelakoak) eta agurra eskatuz idatzitakoak. Aztertu beharreko gaiak bakarrik izanda, etiketatze prozesuari ekin zaio. Python programazio lengoaiaz baliatuz, XML fitxategia etiketatzeko programatxo bat egin eta etiketatzaile bati pasa zaio, gaiak etiketa ditzan (umorezko/ez-umorezko). Etiketatzea amaituta, testu soila egoki ez dela-eta (gaia bere horretan), Zatiak tresna aplikatu zaio gai bakoitzari. Zatiak tresnak testu baten analisi morfosintaktikoa lortzen du, eta beste zenbait informazioz gain, hitz bakoitzak duen kategoria zein den esaten du automatikoki. Hitz guztiak erabakigarri izango ez direlakoan, soilik izenak, aditzak, adjektiboak, aditz trinkoak eta hitanoaren erabilera esplizitua egiten duten hitzak hartu dira. *Content word* edo eduki hitzak deituak hartu dira aintzat, gaiaren esanahia horiek gordetzen dutelakoan. Hitz bakoitzaren lema hartu da, zehatz esanda. Ikus 1. taulan XML fitxategiko adibide bat.

Bestalde, mota honetako lanak egiteko, hiru fitxategi behar izaten dira. Batetik, *Train* eta *Develop* fitxategiak behar dira ikasketa prozesurako, eta bestetik, *Test* fitxategiaren beharra dago amaierako probak egin ahal izateko. Honela egin dugu hiru fitxategien banaketa: 771 gaiz osaturiko *Train* fitxategia, eta 105 gaiz osaturiko *Develop* eta *Test* fitxategiak (ikus 2. taula).

```
<Bertsoaldi>
    <Gaia>
      Gaur, ehun urte bete dituen aitonari kanta iezazkiozu hiru bertso.
    </Gaia>
    <Lana>
      BAKARKA GAIA-EMANDA
    </Lana>
    <Neurria>
      BEDERATZI PUNTUKOA
    </Neurria>
    <Mota>
      DRAMATIKOA
    </Mota>
    <ZatiakWeka>
      urte bete aitona kantatu bertso
    </ZatiakWeka>
</Bertsoaldi>
```

1 taula: Etiketazio lana amaitu osteko XML fitxategiaren zati bat

	TRAIN	(%)	**DEV**	(%)	**TEST**	(%)	**GUZT**	(%)
UMORE	357	0,4630	46	0,4381	40	0,3810	443	0,4516
DRAMA	414	0,5370	59	0,5619	65	0,6190	538	0,5484
GUZTIRA	771		105		105		981	

2 taula: *Train, Develop* eta *Test* corpusen portzentajeak

3.1.2 Ebaluazioa

Ebaluazioari dagokionez, ikasketa automatikoan ebaluatzeko neurri hauek erabili ohi izan dira: doitasuna (*Precision*) eta estaldura (*Recall*). Hala, Weka-k osagai bat zuzen etiketatu duela esango dugu, baldin eskuz etiketatutako osagaiaren kategoria bera badu. Neurri hauek guztiak 3. taula gisako kontingentzia-taula batean oinarrituz kalkulatzen dira, bi klaseko emaitzak (*dramatikoa* eta *umorezkoa*) ditugunean.

	Zuzena=Dramatikoa	**Zuzena=Umorezkoa**
Esleitua=Dramatikoa	a	b
Esleitua=Umorezkoa	c	d

3 taula: Estatistikak kalkulatzeko kontingentzia-taula

3. taulako “a” zenbakiak ’*dramatikoa*’ klasekoak diren eta ’*dramatikoa*’ klasea esleitu zaien elementuen kopurua adierazten du, *True Positive* edo egiazko positiboa; “b” zenbakiak ’*umorezkoa*’ klasekoak diren baina ’*dramatikoa*’ klasea esleitu zaien elementuen

kopurua, *False Negative* edo negatibo faltsua; "c" zenbakiak '*dramatikoa*' klasekoak diren baina '*umorezkoa*' klasea esleitu zaien elementuen kopurua, *False Positive* edo positibo faltsua; eta "d" zenbakiak, berriz, '*umorezkoa*' klasekoak diren eta '*umorezkoa*' klasea esleitu zaien elementuen kopurua, *True Negative* edo egiazko negatiboa.

Azaldutako zenbakiak erabiliz kalkulatzen dira doitasuna eta estaldura, ondorengo lerroetan 'dramatiko' gisa etiketatutakoekin kalkuluak nola egingo liratekeen azalduko dugu. Doitasuna 'dramatiko' gisa esleitutakoen artean zuzen aurresan direnen ehunekoa da. Hau da, 100 gai etiketatu badira dramatiko gisa, eta horietatik 80 etiketatu badira zuzen, % 80 izango da doitasuna (80 / (80 + 20)). Estaldura corpusaren arabera 'dramatiko' direnen artean zuzen aurresandakoen ehunekoa da. Hau da, 100 gai dramatiko badaude corpusean, eta 80 etiketatu badira modu egokian dramatiko gisa, % 80 izango da estaldura (80 / (80 + 20)). Hemen 3. taulan oinarrituz egingo liratekeen kalkuluak:

$$Doitasuna = a/(a + b)$$

$$Estaldura = a/(a + c)$$

Ikasketa automatikoaren emaitzak ebaluatzeko doitasuna edo estaldura erabili ordez, bi neurriak konbinatuz lortzen den *F-Measure* neurria erabiliko dugu. Hau da *F-Measure*-ren kalkulua:

$$F - Measure = 2 * Doitasuna * Estaldura/Doitasuna + Estaldura$$

Azaldu duguna 'dramatiko' gisa etiketatutakoei dagokien kalkulua da, beraz, berdina egin beharko litzateke 'umorezko' gisa etiketatutakoekin eta ondoren batez bestekoa kalkulatu.

3.1.3 Ikasi Beharrekoa

Lan honetan ikasi nahi izan dena kontzeptu konplexua dugu, umorea kontu oso pertsonala baita, eta nahiz eta gai askotan adostasuna nahiko begi-bistakoa izan, badira gaiak zeinetan bat etortzea ez den horren erraza. Horregatik, bigarren pertsona batek ere etiketatu du test corpusa. Aurrerago emango da bi etiketatzaileen arteko adostasunaren inguruko informazioa.

3.1.4 Baseline

Emaitzak aztertzerako orduan, garrantzitsua izan ohi da oinarrizko neurria edo baseline-a ezagutzea. Gure kasuan 2. taulako informazioa begiratuz eta guztiak *dramatikoa*k direla esanda (gehienak halaxe baitira) *F-Measure* neurria kalkulatuko dugu. 3.1.2 atalean ikusita bezala kalkuluak eginda, horrela leudeke *Develop* eta *Test* corpusetako baseline-ak:

Develop corpuseko gai guztiak *dramatikoa*k direla esan, eta horri dagozkion kalkuluak egiteko kontingentzia taula bete, eta bertako informazioarekin kalkuluak egin ditugu.

	Zuzen=Drama	**Zuzen=Umore**
Eslei=Drama	59	46
Eslei=Umore	0	0

4 taula: Estatistikak kalkulatzeko *Develop* corpuseko kontingentzia-taula

$$Doitasuna(Drama) = 59/(59 + 46) = 0,5619$$

$$Estaldura(Drama) = 59/(59 + 0) = 1$$

$$F - Measure(Drama) = 2 * 0,5619 * 1/0,5619 + 1 = 0,7195$$

$$F - Measure(Develop) = 0,7195/2 = 0,3597$$

Kalkulu hauek dramatikoei dagozkienak bakarrik dira, eta umorezkoei dagozkienak 0-koak direnez (doitasuna eta estaldura 0 dira, zatikizuna bietan 0 baita) bien arteko batez bestekoa kalkulatuko dugu, hau da, *F-Measure* (Drama) / 2.

Test corpuserako eragiketa baliokideak gauzatuz, honakoa da *F-Measure* neurria:

$$F - Measure(Test) = 0,3823$$

3.1.5 .arff fitxategia

XML fitxategian gaiak formatu egokian eta etiketatuta izanik, Weka-rekin erabiltzeko prestatu behar izan dira datuak. Horretarako .arff fitxategia sortu dugu, lorturiko gaiaren lema aukeratuak (zatiak erabiliz) atributu gisa emanda, eta ikasi beharreko kategoria gisa DRAMATIKOA eta UMOREZKOA. Ondoren, @data gisa, komatxo bakun artean aurrez aipaturiko gaiari dagozkion lema aukeratuak eta bukaeran dagokion kategoria izango genituzke. Modu honetan, 5. taulan ikus daitekeen bezala geratuko litzateke .arff fitxategia.

```
@relation gaia
@attribute gaiaZatiak string
@attribute kat {DRAMATIKOA,UMOREZKOA}
@data
'OTAN izan boto eman aurre ukan jakin behar galdera'TRAGIKOA
'kartzela barru egon lagun entzun kanpo askatasun eskatu'TRAGIKOA
'Peagarikano futbol izan entrenatzaile Imanol joan aspaldiko igande ukan jokatu aukera eman'UMOREZKOA
'Euzkitze nerbio porru egin laskitu ibili joan psikiatra Telleriangan'UMOREZKOA
...
```

5 taula: XMLtik erauzi ondoren osaturiko .arff fitxategia

Gai guztiak batera harturiko .arff fitxategi hau Weka-rekin ireki, eta StringToWordVector filtroa pasa diogu. Filtro honekin emandako gaiaren hitz aukeratuen zerrenda bere horretan atributu gisa hartu ordez, hitz bakoitza hartzen da atributu gisa. Honen ondoren,

```
@relation 'gaia-weka.filters.unsupervised.attribute.StringToWordVector-...
@attribute 80 numeric
@attribute Aburuza numeric
@attribute Afganistan numeric
@attribute Ainhoa numeric
...
@attribute kat {DRAMATIKOA,UMOREZKOA}

@data
{134 1,330 1,404 1,484 1,585 1,746 1,993 1,1008 1,1536 1}
{296 1,383 1,549 1,594 1,659 1,1049 1,1064 1,1118 1}
{82 1,145 1,301 1,323 1,585 1,591 1,735 1,915 1,993 1,1029 1,1536 1,2112 1,2544 UMOREZKOA}
{60 1,545 1,909 1,1029 1,1385 1,1407 1,1700 1,2183 1,2263 1,2544 UMOREZKOA}
...
```

6 taula: StringToWordVector aplikatu ondorengo .arff fitxategia

kategoria bera azken atributu gisa jartzeaz oroitu behar gara, bestela ez baitigu ezertarako balioko. Aldaketa hauen ondorengo fitxategia 6. taulan ikus daiteke.

6. taulan ikus dezakeguna, *sparse* errepresentazio bat da. Hasierako zatian atributu zerrenda osoa dago, azken atributu gisa kategoria daukalarik. Ondoren, @data azpian gai bakoitzaren zenbakizko kodeketa bat ageri da. Kodeketa horretan, koma eta koma artean dagoen bikote bakoitza hitz bati dagokio. Adibide batekin argiago ikusiko dugu. Demagun ondorengo gaia eta gai honi dagokion *sparse* errepresentazioko lerroa ditugula:

```
'OTAN izan boto eman aurre ukan jakin behar galdera'DRAMATIKOA
{134 1,330 1,404 1,484 1,585 1,746 1,993 1,1008 1,1536 1}
```

Kasu honetan, 134 'OTAN' hitzaren indizea da, 330 'izan' hitzarena den bezala. Indizearen ondoren ageri den 1 zenbakiak, hitz hori gaian agertzen dela adierazten du. Gaia *umorezkoa* den kasuan, *sparse* errepresentazioan azken hitz gisa (beti 2544 indizearekin) UMOREZKOA balioa agertzen zaigu. Indize zenbaki horrek gaia bera adierazten du, hau da, gaiaren etiketa. Esan beharrekoa da gure kasuan 2544 atributu ditugula (0-2543 indexatuak) eta hortaz zenbaki hori indize gisa duen hitzik ez daukagula atributu gisa. Horregatik erabiltzen da 2544 zenbakia gaiarentzat. *dramatikoa* den kasuetan ez da ageri, era bitarrean erabiltzen delako, hau da, ez badago, *dramatikoa* da, eta agertzen bada, umorezko gisa agertzen da.

3.1.6 Sailkatzaileak

Atal honetan erabili ditugun sailkatzaileen azalpen labur bat emango dugu.

RandomForest: "*divide et impera*" algoritmoan oinarritutako sailkatzailea da. Erabaki zuhaitz honekin atributu bat aztertu eta balioen arabera adarrak definitzen dira. *RandomTree*-z osatutako "basoa" da eta *baggin* metodoa erabiltzen du (sailkatzaileak elkarlanean harturik).

NaiveBayesMultiNominal: Aldagai asko daudenean eta aldagai guztien arteko konbinaketen probabilitateak kalkulatzeko nahikoa adibide ez dagoenean, kalkuluak sinplifikatzen dira. Eredu multinominala erabilita, (p1, ... pn) multinominal batek sortutako gertaeren maiztasunak irudikatzen dituzte laginek, non pi, i gertaera gertatzeko probabilitatea den. Aldagai asko kontuan hartu behar direnean egokia izan daiteke.

SMO (*Sequential Minimal Optimization*): Bere forma sinpleenean bi klaseen arteko muga lineala ezartzen du, hala ere, linealak ez diren datu multzoentzat soluzio bat ematen du. Ezartzen zaizkion parametro desberdinen arabera, muga lineala alda daiteke. Atributuak normalizatzen ditu, eta izenezko atributuak bitar bihurtzen ditu. 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 (*maximum margin hyperplane*) deitzen zaion eredu lineal berezi bat baliatzen baitu. Hainbat atazatan erabiltzen da eredu lineal hau.

Praktikan, goi-muga antzeko bat markatzen duen parametro bat kalkulatzea da gakoa, eta horretarako, esperimentuak egitea beste aukerarik ez dago. Esan beharra dago ikasketa automatikoko eskema hau ez dela batere azkarra, batez ere adibide asko dituen ikasketa corpusekin lan egiterakoan. 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.

Logistic: Emaitza bitarrak aurresateko erabili ohi da, hainbat kategoriako baina bi klaseko predikzioetan. Erantzun kualitatiboko ereduen parametroen estimazioetan erabiltzen da. Proba bakarrak deskribatzen dituen emaitza posibleen probabilitateak, funtzio logistiko bat erabiliz aldagai explikatiboen funtzio gisa modelatzen dira.

Logistic sailkatzaileak aldagai kategoriko dependiente baten eta aldagai independiente bat edo gehiagoren arteko erlazioa neurtzen du. Neurketa hau aldagai dependientearen aurresandako balioen gisako probabilitate puntuazioak erabiliz egiten da.

3.1.7 Atributuen aukeraketa

Ikasketa automatikoan oinarrizkoa den ekintza dugu atributuen aukeraketa. Eskura ditugun atributu guztiak ez dira maila berean esanguratsu izaten. Horregatik batzuk baztertzea komeni izaten da, ondoren ikasketa-algoritmoen emaitzak hobeak izan daitezen. Erabilitako atributuak atazarekiko esanguratsuak izateak emaitza hobeak lortzeko bidea ematen du, hortaz, hori da aukeraketaren karia, aztertu beharreko atazarekiko esanguratsuenak diren atributuak aukeratzea. Atributu-aukeraketa bi pausotan gauzatzen da:

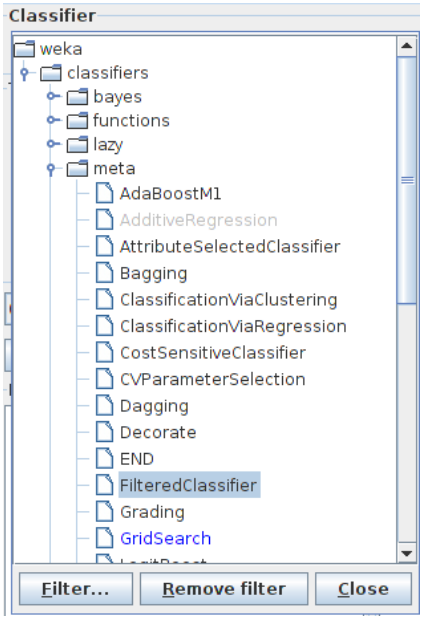
Bilaketa: Atributuen espazioan, ezaugarri multzo egokiena bilatu behar da. Hainbat aukera dauden arren, lan honetan atributuak nola ordenatuko diren eta horietatik zenbat aztertuko ditugun hartuko dugu kontuan.

Ebaluazioa: Atributu azpimultzo horren kalitate edo doitasuna ebaluatzeko metodo ezberdinak erabiltzen dira: informazio-irabazia, elkar-informazioa, *chi-square*...

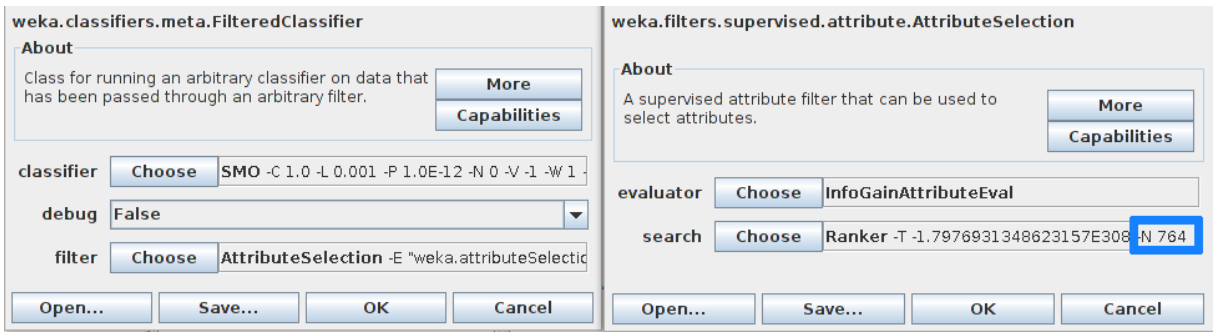
3.2 Egindako saioak

3.2.1 Hasierako probak

Train fitxategiarekin ikasi eta *Develop*-en proba egiteko meta ataleko *FilteredClassifier* sailkatzailea erabili dugu, honek, barnean *SMO, Logistic, NaiveBayesMultiNominal* eta *RandomForest* sailkatzaileak erabiliz, eta *AttributeSelection* filtroarekin atributuak aukeratuz. Filtroa *InfoGainAttributeEval* (atributuak banaka hartuta) ebaluatzailea eta *Ranker* "bilatzailea" (ez du bilatzen, atributuak ordenatu egiten ditu) aplikatuta baliatu dugu, aldiko atributu kopuru desberdina erabiliz. Hau aplikatuta, atributu guztiak erabiltzea ez da egokia; beraz, zenbat atributu aukeratu pentsatu behar dugu. Atributu kopuru totalaren (2544) portzentajeak hartzeari ekin diogu. Lehenengo probak honakoak izan dira: *Ranker*-ak ematen dituen balioetan zeroko guztiak kendu, eta balio positiboa duten atributuak soilik hartuta (91 atributu), eta atributu kopuru totalaren ehuneko hauek: % 5 (128 atributu),% 20 (509 atributu), % 30 (764 atributu), % 40 (1018 atributu) eta % 50 (1273 atributu)



4 irudia: Sailkatzaileak



5 irudia: *FilteredClassifier* sailkatzailearen ezarpenak

7. taulan ikus daitezke hasieran erabilitako sailkatzaile eta atributu kopuru desberdinekin lorturiko emaitzak.

Gauzak honela, atributu kopuru desberdin gehiago hartzea erabaki da. 8. taulan ikus daiteke informazio osotuagoa.

8. taulan ikusten den bezala, atributu kopuru berdinarekin nahiko antzerako emaitzak ematen dituzte sailkatzaileek (ikus 509 edo 1018 atributu hartuta). Horregatik sailkatzaile

F-Measure	SMO	LOGISTIC	NBMN	RF
91 att (0 kenduta)	0,659	**0,712**	0,68	0,65
128 att (%5)	0,663	**0,704**	0,68	0,663
509 att (%20)	0,714	**0,715**	0,705	0,687
764 att (%30)	**0,762**	0,602	0,676	0,685
1018 att (%40)	**0,741**	0,63	0,722	0,704
1272 att (%50)	0,69	0,611	**0,732**	0,696

7 taula: Hainbat atributu kopururekin lorturiko emaitzak

F-Measure	SMO	LOGISTIC	NBMN	RF	**BB4**
91 att (0 kenduta)	0,659	**0,712**	0,68	0,65	0,67525
128 att (%5)	0,663	**0,704**	0,68	0,663	0,6775
400 att	0,704	0,687	**0,724**	0,696	**0,70275**
509 att (%20)	0,714	**0,715**	0,705	0,687	**0,70525**
550 att	**0,753**	0,677	0,705	0,648	0,69575
600 att	**0,704**	0,687	0,696	0,62	0,67675
650 att	**0,733**	0,686	0,705	0,677	**0,70025**
700 att	**0,751**	0,658	0,715	0,578	0,6755
764 att (%30)	**0,762**	0,602	0,676	0,685	0,68125
800 att	**0,704**	0,563	0,714	0,667	0,662
850 att	**0,733**	0,647	0,695	0,677	0,688
900 att	**0,733**	0,62	0,703	0,63	0,6715
950 att	**0,723**	0,63	0,703	0,668	0,681
1018 att (%40)	**0,741**	0,63	0,722	0,704	0,69925
1100 att	**0,72**	0,6	0,712	0,658	0,6725
1272 att (%50)	0,69	0,611	**0,732**	0,696	0,68225

8 taula: Atributu kopuru desberdinekin egindako probak. Azken zutabean, lau sailkatzaileen emaitzen batez bestekoa.

desberdinen konbinazioa egitea erabaki da. Horretarako, atributu kopuru desberdin guztiekin proba egin ordez, batez besteko onena duten hirurak hartzea erabaki da, hau da, 400, 509 eta 650 att.

Konbinatzeko aukeraturiko modua *Vote* izan da. Sailkatzaile desberdinak konbinatzeko balio duen klasea da hau, probabilitateen konbinaketa desberdinak daude sailkapena ahalik eta egokiena izan dadin. Hauek dira hautatu ditugun hiruak: *Average of Probabilities* (probabilitateen batez bestekoa erabiliz kalkulatzen da zein aukeratu), *Majority Voting* (sailkatzaile gehienek erabakitzen dutena aukeratzen da) eta *Maximum Probability* (probabilitate altuena daukana aukeratzen da).

9. taulari begiratuz, emaitzak hobetzen direla ikusita, lau sailkatzaileak hartu ordez, kasu bakoitzean hiru emaitza onenak ematen dituzten sailkatzaileak harturik, batez bestekoa kalkulatu, eta onenekin *Vote* erabiltzea erabaki da, aurrez bezala *Average of Probabilities,*

					VOTE		
F-Measure	SMO	LOGISTIC	NBMN	RF	**AVG**	**MAJOR**	**MAX**
400 att	0,704	0,687	**0,724**	0,696	**0,743**	0,696	0,704
509 att (%20)	0,714	**0,715**	0,705	0,687	0,725	**0,753**	0,723
650 att	**0,733**	0,686	0,705	0,677	0,733	0,725	**0,761**

9 taula: *Vote* eta lau sailkatzaileak erabiliz lorturiko emaitzak

Majority Voting eta *Maximum Probability* konbinaketa arauak erabiliz.

						VOTE		
F-Measure	SMO	LOGISTIC	NBMN	RF	**BB3**	**AVG**	**MAJOR**	**MAX**
509 att (%20)	0,714	**0,715**	0,705	~~0,687~~	**0,711**	0,734	**0,763**	0,723
550 att	**0,753**	0,677	0,705	~~0,648~~	**0,712**	0,734	0,725	**0,752**
1018 att (%40)	**0,741**	~~0,63~~	0,722	0,704	**0,722**	**0,761**	0,753	0,741

10 taula: Vote eta unean uneko hiru sailkatzaile onenak erabiliz lorturiko emaitzak

3.2.2 Onenen aukeraketa

11. taulan ikus daitezke emaitza guztiak batera. *Test*-ean probak egiterakoan, ordea, onenak errepikatu ohi dira, beraz, atributu kopuruaren % 20 eta % 30 artean dauden atributuak garrantzitsuak direla dirudienez, tarte horrekin egingo da proba. Batetik, sailkatzaile soilekin, eta ondoren hiru onenak hartu eta *Vote* aplikatuz, *Average of Probabilities, Majority Voting* eta *Maximum Probability* konbinaketa arauak erabiliz.

F-Measure	SMO	LOGISTIC	NBMN	RF	BB4	VOTE (4) AVG	VOTE (4) MAJOR	VOTE (4) MAX	BB3	VOTE (3 onenak) AVG	VOTE (3 onenak) MAJOR	VOTE (3 onenak) MAX
91 att (0 kenduta)	0,659	**0,712**	0,68	0,65	0,67525	0	0	0	0,683666667	0	0	0
128 att (%5)	0,663	**0,704**	0,68	0,663	0,6775	0	0	0	0,682333333	0	0	0
400 att	0,704	0,687	**0,724**	0,696	**0,70275**	**0,743**	0,696	0,704	0,708	0	0	0
509 att (%20)	0,714	**0,715**	0,705	0,687	**0,70525**	0,725	**0,753**	0,723	**0,711333333**	0,734	**0,763**	0,723
550 att	**0,753**	0,677	0,705	0,648	0,69575	0	0	0	**0,711666667**	0,734	0,725	**0,752**
600 att	**0,704**	0,687	0,696	0,62	0,67675	0	0	0	0,695666667	0	0	0
650 att	**0,733**	0,686	0,705	0,677	**0,70025**	0,733	0,725	**0,761**	0,708	0	0	0
700 att	**0,751**	0,658	0,715	0,578	0,6755	0	0	0	0,708	0	0	0
764 att (%30)	**0,762**	0,602	0,676	0,685	0,68125	0	0	0	0,707666667	0	0	0
800 att	**0,704**	0,563	0,714	0,667	0,662	0	0	0	0,695	0	0	0
850 att	**0,733**	0,647	0,695	0,677	0,688	0	0	0	0,701666667	0	0	0
900 att	**0,733**	0,62	0,703	0,63	0,6715	0	0	0	0,688666667	0	0	0
950 att	**0,723**	0,63	0,703	0,668	0,681	0	0	0	0,698	0	0	0
1018 att (%40)	**0,741**	0,63	0,722	0,704	0,69925	0	0	0	**0,722333333**	**0,761**	0,753	0,741
1100 att	**0,72**	0,6	0,712	0,658	0,6725	0	0	0	0,696666667	0	0	0
1272 att (%50)	0,69	0,611	0,732	0,696	0,68225	0	0	0	0,706	0	0	0

11 taula: Emaitza guztien taula osotua (*Develop* corpusean)

3.2.3 *Test*-ean probatu

Horretarako, *Train* eta *Develop* corpusak bateratu egingo ditugu, gero *Test*-aren gainean probatzeko. 12. taulan ikus daitezke lortu ditugun emaitzak:

					VOTE		
F-Measure	SMO	LOGISTIC	NBMN	RF	AVG	MAJOR	MAX
509 att (%20)	0,622	**0,629**	~~0,568~~	0,622	**0,632**	0,631	0,622
550 att	0,604	**0,605**	~~0,568~~	0,568	0,615	**0,633**	0,604
600 att	0,604	**0,615**	~~0,596~~	**0,615**	**0,634**	**0,634**	0,613
650 att	**0,603**	0,587	0,587	~~0,572~~	0,606	0,606	**0,612**
700 att	**0,614**	~~0,549~~	0,568	0,549	**0,614**	0,578	**0,614**
764 att (%30)	**0,642**	~~0,521~~	0,568	0,586	**0,642**	0,596	**0,642**

12 taula: *Test*-ean probak egitean lorturiko emaitzak

3.2.4 Etiketatzaileen arteko adostasuna

Ikusiriko portzentajeak etiketatzaile bakarrarekin eginikoak dira, eta ondorioz, etiketatzailearen ziurtasuna jakiteko eta, batez ere, atazaren zailtasuna neurtzeko, *test*-eko atala bigarren etiketatzaile bati eman zaio, honek etiketatu dezan. Etiketatzea amaitzean, bi etiketatzaileen adostasuna kalkulatu dugu.

		2. Etik.	2. Etik.	
		UMORE	**DRAMA**	**GUZTIRA**
1. Etik.	**UMORE**	24	16	40
1. Etik.	**DRAMA**	4	61	65
	GUZTIRA	28	77	105

13 taula: *Test* corpusaren confusion matrix-a

Etiketatzaileen adostasuna kalkulatzeko koefiziente desberdinak existitzen dira, eta guk erabilienak diren 3 koefizienteren kalkulua egin dugu (Artstein eta Poesio, 2008). Hiru koefizienteetan aldatzen den parametroa *Expected agreement* (A_e) da, hortaz, aurrena horien kalkuluen formulak azalduko ditugu.

		2. Etik.	2. Etik.	
		UMORE	**DRAMA**	**GUZTIRA**
1. Etik.	**UMORE**	U_u	U_d	U_1
1. Etik.	**DRAMA**	D_u	D_d	D_1
	GUZTIRA	U_2	D_2	G

14 taula: Formulen azalpenerako confusion matrix-a

Etiketatzeko daukagun kategoria kopuruari k deituko diogu (gure kasuan 2 dira). 14. taulan ikus daitezke formuletarako erabiliko ditugun D eta U karaktereak. Letra larriak

lehenengo etiketatzaileak eman dion erregistroa adierazten du, eta azpi-indizeak bigarren etiketatzaileak esleitutako erregistroa. Hortaz, esaterako, U_u bi etiketatzaileek umorezko gisa etiketatutako gaien kopurua litzateke.

All Categories Are Equally Likely; S: Koefiziente hau bi etiketatzailek (bakoitza bere aldetik) lan bat egitean distribuzio uniforme bat lortuko dugulakoan oinarritzen da.

$$A_e = k * \left(\frac{1}{k}\right)^2 = 2 * \left(\frac{1}{k}\right)^2$$

A Single Distribution; π: Etiketatzaileak bakoitza bere aldetik lanean badabiltza, bakoitzarentzat distribuzio berdina lortuko dugu.

$$A_e = \left[\frac{(U_1 + U_2)}{2 * G}\right]^2 + \left[\frac{(D_1 + D_2)}{2 * G}\right]^2 = \left[\frac{(40 + 28)}{2 * 105}\right]^2 + \left[\frac{(65 + 77)}{2 * 105}\right]^2$$

Individual Coder Distributions (Kappa); k: Etiketatzaileak bakoitza bere aldetik lanean badabiltza, etiketatzaile bakoitzarentzat distribuzio banatu bat lortuko dugu.

$$A_e = \left[\left(\frac{U_2}{G}\right) * \left(\frac{U_1}{G}\right)\right] + \left[\left(\frac{D_2}{G}\right) * \left(\frac{D_1}{G}\right)\right] = \left[\left(\frac{28}{105}\right) * \left(\frac{40}{105}\right)\right] + \left[\left(\frac{77}{105}\right) * \left(\frac{65}{105}\right)\right]$$

Azken taula honetan (15. taulan) dauzkagu koefizienteon kalkuluak. Lehen zutabean ageri dira koefiziente bakoitza aipatzeko erabili ohi den ikurra. Bigarrenean bi etiketatzaileak bat datozen gaien portzentajea edo *Observed agreement* (A_o). Hirugarren zutabeak aurrez aipatu dugun *Expected agreement* (A_e) kalkuluaren emaitza erakusten digu. Azken zutabean dago koefizientearen emaitza edo *Chance-corrected agreement* (A_{cc}). Emaitzak ikusita, esan genezake ez dela erraza ataza honetako etiketatzea zuzen egitea. Kappa neurriaren balioak interpretatzeko irizpide anitz egonik ere, Carletta-k (Carletta, 1996) berak % 80tik gorako balioak jotzen ditu fidagarritzat; % 67 eta % 80 artekoak, berriz, zalantzagarriak direla dio. Hala ere, neurri hauek medikuntzan erabili ohi direnak dira, eta beraz, fidagarritasun oso altua eskatzen zaienak. Gure kasuan ordea, badirudi ataza nahiko zaila dela, izan ere umorea kontu subjektiboa dela baiezta dezakegu, eta ondorioz Kappa neurri altua lortzea lan zaila dela. Gure kasuan Kappa neurria % 60 azpitik dagoen arren *Observed agreement*-a % 81 izateak ados datozen gaien kopuru altua dela adierazten digu. Ondorioz, nahiz eta neurri txikiak lortu, esan genezake aztertzen ari garen atazarako ez direla batere zenbaki txarrak.

Coefficient	A_o **= 24+61/105**	A_e	A_{cc}=$(A_o - A_e)/(1 - A_e)$
S	**0,8095238095**	**0,5**	**0,619047619**
π	**0,8095238095**	**0,5620861678**	**0,5650372825**
k	**0,8095238095**	**0,5555555556**	**0,5714285714**

15 taula: Koefiziente desberdinen balioen kalkulua

4 Bertsoa eta bertsotan egiteko emandako ariketaren arteko antzekotasun semantikoa

Esperimentu hau garatzeko *Textual coherence in a verse-maker robot* (Astigarraga et al., 2014) lanean oinarritu gara, 2.7 atalean azaldu dugun moduan. Automatikoki bertsoak sortzeko helburua duen *Bertsobot* proiektuaren baitan, bertso barneko ahalik eta koherentzia handiena lortzean datza aipaturiko lana. Bertsoa sortzerako orduan, bertsolariak gai baten inguruan kantatzen du, horretarako ematen zaion ariketaren arabera. Aurrez aipaturiko lanaren kasuan makinari lau oin ematen zaizkio, eta hitz horietan oinarrituta beraien artean erlazio semantiko handiena duten lau esaldiekin bertsoa osatzea da helburua. Bertsoa osatzeko emandako oinek estrofa amaieran joan behar dute, eta estrofen neurriak *Zortziko Txikia*n (7, 6, 7, 6, 7, 6, 7, 6) bete behar dira. Gauzak honela, lortuko diren esaldi bakoitzak bi lerro bete beharko ditu, esaldiaren azken hitza emandako oinetako bat izanik eta 13 silabaz osaturik.

Bertsoa kantatzeko emandako hitzaren edo gaiaren eta bertsoa beraren arteko erlazioaren zenbatekotasuna lortzean datza hurrengo lerroetan azalduko dugun esperimentua. Bertsotarako gai bat edo hitz bakarra eman, desberdina izan ohi da bertsolariak egin behar izaten duen lana, hortaz, bakoitzari dagokion azterketa ere desberdina izango da. Erlazio semantikoaren neurketa beti estrofaka egingo da, izan gaia emanda ala hitza emanda, eta erabili dugun teknika *Latent Semantic Analysis* izan da.

4.1 Esperimentua

Esperimentu honetan bertsoen sorkuntzan baino, bertsoa eta bertsoa osatzeko emandako ariketaren analisian jarriko dugu arreta; zehazki erlazio semantikoaren analisian. Analisi horietarako, momentuko bi bertsolari onentsuenen bertsoak aukeratzea erabaki dugu, kasu, Amets Arzallus eta Maialen Lujanbio. Bi dira analizatzea erabaki dugun ariketa motak: batean bertsolariari hitz bakarra ematen zaio, eta hitz horrekin erlazionatuta dagoen bertso bat botatzea izango da haren lana. Bestean, gai bat emango zaio, eta gai horren inguruko hiru bertso garatu beharko ditu bertsolariak. Bi ariketak gehienetan kartzelako gai gisa ematen dira, eta modalitate honetan bi bertsolariek gai edo hitz berari erantzun behar izaten diote, bestearena entzun gabe. Bat kantatzen ari den bitartean bestea 'kartzelan' dago ezer entzun gabe, eta lehenak bertsoak kantatzen bukatzean bigarrenak egingo du saioa. Kartzelako gaiari erantzunez botatako bertsoak izateak, bi pertsona desberdinek gai edo hitz berari jarritako bertsoak aztertzeko aukera ematen digu.

Latent Semantic Analysis (LSA)

Latent Semantic Analysis teknika (Deerwester et al., 1990) corpus batetik abiatuz matrize bat sortzean oinarritzen da. Matrizea sortu ahal izateko, beharrezkoa da corpusa dokumentuetan banatzea, eta dokumentu hauek tamaina desberdinetakoak izan daitezke: esaldi batetik artikulu oso baterainokoak. Gure kasuan dokumentuarentzat aukeratu du-

gun tamaina paragrafoa da, ingurune ez oso murritz ezta oso zabala sortzeko. Aipaturiko matrizea sortzeko hitz bakoitza ageri den ingurunea (berau agertzen den dokumentuko beste hitzak) hartzen da kontutan. Adibidez, 'sua' hitza maiz agertzen bada 'txispa', beraien arteko erlazio semantikoa esanguratsuagoa izango da, bestalde, 'sua' eta 'teklatua' hitza oso gutxitan agertzen badira dokumentu berean, erlazioa ez da batere nabarmena izango. Metodo honek arazoak eman ditzake hitz bat oso sarritan agertzen bada testuaren gaia edozein delarik ere. Izan ere, ez da askorik nabarmenduko gaiaren arabera. Hortaz kendu beharreko hitzen zerrenda baten beharra dago. Horretarako, esperimentuetan erabiliko dugun *Euskaldunon Egunkaria*-ko corpus etiketatua analizatu dugu eta bertatik maizen agertzen diren 100 hitzak hautatu ditugu *zstopword* gisa.

Corpusa

Aztertuko ditugun bertso motak aurkeztuta, ikasketarako erabiliko dugun corpusarekin jarraituko dugu. Hautatu dugun corpusa, *Euskaldunon Egunkaria*ren corpus lematizatua da. Corpus honetan egunkariko artikulu, albiste eta orokorrean hainbat testu dauzkagu, eta corpuseko hitz guztiak aurrez lematizatuta dauzkagu (*stemmer* bat erabilita). *Latent Semantic Analysis*-en dokumentu gisa erabiliko ditugun 95580 paragrafok osatzen dute corpusa. Paragrafo guztiak kontuan hartuta 4.192.616 hitz daude corpusean, eta horietatik 87.963 dira hitz-lema ezberdinak.

4.1.1 Hitza emanda

Ariketa mota honetan bertsolariari hitz bat ematen zaio, eta honek hitzarekin erlazionatutako bertsoa kantatu behar du. Gure kasuan emandako hitza eta bertsoko estrofa bakoitzaren arteko erlazio semantikoa kalkulatuko da (banan-banan), eta honela, 0 eta 7 bitarteko puntuazioak lortuko ditugu (azterturiko bertsoek 7 estrofa dituzte, eta konparaketa bakoitzak 0 eta 1 arteko zenbaki erreala ematen du).

4.1.2 Gaia emanda

Bertsolariak gai baten inguruko hiru bertso kantatu behar izaten ditu beste modalitate honetan, eta horregatik, neurketak beste era batera egingo dira. Hasiera batean gaiak bertso bakoitzarekin zeukan erlazioa kalkulatu genuen arren, neurketa modua aldatzea erabaki genuen. Izan ere, ariketa mota honetan bertsolariak hiru bertso osatzeko betebeharra duenez, ez du hasierako bertsotik gaiarekin zuzenean lotura egiten. Bertsoz bertsoko kontakizun bat osatzen du, hasierako bertsoan testuinguru bat sortuz askotan, eta ondorengoetan kontakizuna garatuz. Honekin, bertso bakoitza guztiz ulertzeko aurrekoaren beharra dagoela sumatu dugu, eta horrek neurketa modua moldatzeko beharra dakarkigu. Ondorioz, estrofa bakoitza emandako gaiarekin eta aurrez kantaturiko bertsoekin konparatuko da, dagoen erlazio semantikoa aztertze aldera. Lehen bertsoaren kasuan, gaiarekin bakarrik egingo da neurketa.

4.2 Emaitzak

4.2.1 Hitza emanda

Lau dira analizatu ditugun lanak, bi bertso 'sua' hitzari kantatuak eta beste bi 'altzoa'-ri. Biak ala biak Bertsolari Txapelketa Nagusiko lanak dira, 2009 eta 2013 urteetakoak hurrenez hurren. Bietan, bata Amets Arzallusena eta bestea Maialen Lujanbiorena dira. 16 eta 17. tauletan ikus daitezke 'sua' hitzari kantatutako bertsoak aztertzean lorturiko emaitzak.

Amets Arzallus	**'Sua'-rekiko**
Pospolu batek pizten badu lehen ditxa	0,1338325739
gero bota sastraka eta zumitza	0,0656301156
tximinitikan gora doa bere gisa	0,0350185186
baina neretzat sua dugu bertsogintza	0,9995695353
bertsoa balitza su baten baldintza	0,7073464394
pitz dezagun hitza alaituz bizitza	0,0683775246
hemen su horren bueltan dantzan gabiltza	0,3949621618
Batura	**2,404736869**
Batez bestekoa	**0,343533838**

16 taula: Amets Arzallusen bertsoak 'sua'-rekiko duen erlazio semantikoa.

Maialen Lujanbio	**'Sua'-rekiko**
Hura asmakizuna homo habilisena	0,1093494743
bazun intentzioa bazuen sena	0,0449577048
bi harrik elkar jota txispa bat aurrena	0,0074977744
asmakizun haundina mendeetan barrena	0,0689423233
sua da problema zenbait basorena	0,9966166019
edo jaki dena berotzen duena	0,0930994749
eta bi begiradek sortzen dutena	0,0208336413
Batura	**1,3412969951**
Batez bestekoa	**0,191613856**

17 taula: Maialen Lujanbioren bertsoak 'sua'-rekiko duen erlazio semantikoa.

Taulako estrofa bakoitzaren alboan zenbaki bat ageri da. Zenbaki honek estrofa bera eta emandako hitzaren arteko erlazio semantikoa adierazten du. Azken aurreko lerroan estrofa guztien puntuazioen batura ikus daiteke eta azkenengoan estrofa guztien puntuazioen batez bestekoa. Batez bestekoa ariketa mota honetan ez dugu erabiltzen, bertso guztiak neurri berekoak direlako, baina kalkuluak ateratzen ditugu 4.1.2 ataleko emaitzentzat erreferentzia bat izateko.

Begiratu orokor batean badirudi bertsoa eta hitzaren arteko erlazioa ez dela oso handia, izan ere 'sua'-ri kantatutako bi bertsoetan 7ko puntuazio maximo batetik (1 da lerro bakoitzean lor daitekeen puntuazio handiena) 2,40 eta 1,34 lortzen dugu ondoz-ondo (ikus bi tauletako 'Batura' lerroak). Estrofaz estrofa begiratuta, 'sua' hitza bera agertzen diren

lerroek puntuazio handienak lortzen dituztela berdez markatuta ikus dezakegu 16. zein 17. tauletan. Hala ere, berez erlazionatuta dauden hitz batzuek (esaterako 'pospolu', 'tximinitikan' edo 'txispa') ez dute horrenbesteko pisurik hartzen (ikus kasu hauek 16. eta 17. tauletako gorriz nabarmendutako neurrietan). Badirudi corpusean ez direla elkarrengandik gertu agertzen. Goazen orain 18 eta 19. tauletako 'altzoa' hitzari kantatutako bertsoen emaitzak ikustera.

Amets Arzallus	**'Altzoa'-rekiko**
Haur nintzela banuen aurpegi itzala	0,0204290301
kopetan ximurtua banun azala	0,0131416023
jakin nere sendagai beti ama zala	0,0659430772
haren usai goxoa ta haren kresala	0,0550328977
amaren ahala da unibertsala	0,0453446023
goxoa apala ta zabal zabala	0,055473201
ez dut inoiz ahaztuko zure magala	0,0100778025
Batura	**0,265442213**
Batez bestekoa	**0,037920316**

18 taula: Amets Arzallusen bertsoak 'altzoa'-rekiko duen erlazio semantikoa.

Maialen Lujanbio	**'Altzoa'-rekiko**
Helduok bihotzean zenbat harramazka	0,0683355406
bizitzaren bizitzez indarrak gasta	0,0168083534
nekeak eta minak askotan arrasta	0,0604400486
uzten gaituzte eta hori nola aska	0,0927355066
Bihotzeko zasta bizkarreko lazta	0,0187029503
banoa arnaska hartzen nau xarmaz ta	0,0230260249
altzoa da helduok dugun sehaska	0,0970831662
Batura	**0,3771315906**
Batez bestekoa	**0,053875942**

19 taula: Maialen Lujanbioren bertsoak 'altzoa'-rekiko duen erlazio semantikoa.

Batura bakarrik begiratzen badugu, badirudi oraingoan ez dagoela inolako erlaziorik hitza eta bertsoak kontatzen duenaren artean; izan ere, 18. eta 19. tauletako 'Batura' lerroetako balioei erreparatuz 0,26 eta 0,38 zenbakiak ikus ditzakegu hurrenez-hurren. Estrofaz estrofa begiratuta ere badirudi ez duela pisu gehiegirik hartzen emandako hitza estrofan bertan agertzeak. Hala gertatzen da 19. taulako gorriz markatutako puntuazioa duen lerroan; ez da 0,1 izatera ere iristen. 'Sua'-ri dagozkien bertsoetan, 'altzoa' hitza aipatzen den lerro honek bestelako puntuazioa lortzen du esaterako 16. taulako urdinez nabarmendurikoak. 'Altzoa' hitzaren agerpen honekin batera 'helduok' hitza ere ageri da, eta hori ez da emandako hitzarekiko batere adierazgarria gure indizearen arabera; honek estrofa horren puntuazioa txikiagoa izatea ekar dezake.

HAP masterra

Azterturiko lau bertsoekin egindako esperimentuetan lorturiko emaitzak txikiak izan dira. 'Sua'-ri dagokionez 'txispa' hitza agertzen den esaldiak lortzen duen puntuazio txikiaren arrazoiak edota 'altzoa'-ri dagokionez 'magala' hitzak pisu handiagoa ez hartzearen arrazoiak aztertzeko, gure indizearen portaera ikustea erabaki dugu. Horretarako corpus guztiko hitzen eta analizatu beharreko bi hitzen arteko konparaketa bat egin dugu, hitz hutsek beraien artean duten erlazioaren neurriak lorturik.

Sua	**Erlazioa**	**Altzoa**	**Erlazioa**
eten	0,9477	mielmari	0,6252
menia	0,7538	zaldibikoak	0,5447
paramilitarismoa	0,7380	sega	0,5326
baletor	0,7192	kamixeta	0,4566
tregua	0,6838	segalari	0,4028
wittgensteinek	0,6751	zituzketen	0,3801
ezterenzubiko	0,6650	fidek	0,3585
koalitatiborik	0,6605	txapelketa	0,3485
kontestuan	0,6552	ureta	0,3414
pirotecnia	0,6522	fideko	0,3257
nubako	0,6337	azkaraten	0,3179
etari	0,6309	busturiarraren	0,3173
angolar	0,6271	alper	0,3142
amata	0,6267	marlow	0,3133
montezin	0,6203	bereikua	0,3120
richek	0,6130	menda	0,3057
garenontzat	0,6001	twifi	0,3052
unitako	0,5999	zitzaizkion	0,3028
tenetek	0,5986	iiarekin	0,3011
tximinia	0,5912	duval	0,3006
lahar	0,5883	bista	0,2996
inausketa	0,5807	remontival	0,2994
gar	0,5802	jasoaldi	0,2986
kiskal	0,5782	valencianok	0,2982
aljeriatik	0,5673	aresok	0,2955
caballer	0,5613	arraiotz	0,2945
ke	0,5579	piraguismo	0,2944
tenet	0,5538	falguiere	0,2941
dostumen	0,5486	clesinger	0,2941

20 taula: 'Altzoa' eta 'Sua'-rekiko erlazio altueneko hitzak

20. taulan ikus daitezkeen emaitzak lortuta, azterturiko bertsoen emaitzak horren onak ez izatearen zantzuak hartzen hasten gara. 'Sua' hitzaren kasuan, corpuseko erreferentziazko hitz gehienak gerra edo borroka armaturen bati buruzkoak ditugu. Argiago ikusteko, zerrendako lehen hiru hitzak ikusi besterik ez dago: 'eten' (su-eten hitzaren parte), 'menia' (su-eten edo tregua) eta 'paramilitarismoa'. Normalean suarekin erlazionatuko genituzkeen hitzetan lehena 'pirotecnia' dugu, eta bera hamargarren lekuan dago. Atzerago agertzen dira 'tximinia', 'gar', 'kiskal' eta 'ke'. 'Altzoa' hitzaren zerrendari erreparatuz, espero

genituen 'haur', 'ama' edo 'magala' gisako hitzak ez zaizkigu ageri. Horrez gain, zerrendako lehen hitzaren puntuazioari erreparatzen badiogu, ohartzen gara antz handiena duena izateko oso puntuazio txikia duela (0,6252), eta honek emaitzak esanguratsuak ez izatea dakarrela uste dugu.

4.2.2 Gaia emanda

Oraingo honetan sei bertso dira aztertzeko dauzkagunak, bertsolari bakoitzaren hiru. Aurrez kantaturiko bertsoa (emandako gaiaz gain) ere kontuan hartuko da. Horixe izango da egiteko moduaren desberdintasuna. Alegia, lehenengo bertsoa gaiarekin bakarrik konparatuko da, baina bigarren bertsoa aztertzerako orduan gaiari aurrez kantaturiko bertsoa gehituko zaio, eta hirugarren bertsoa aztertzerakoan bigarren bertsoa ere gehituko zaie gaia eta lehen bertsoari. Ondorengoa da bertsolariei bertsoak kantatzeko eman zitzaien gaia: *'Hileko aurrez aurreko bisita daukazu. Heldu zara kartzelara eta funtzionarioak esan dizu sartu nahi baduzu, bere aurrean biluzi beharko duzula eta miatzen utzi.'*

Oraingo honetan batez bestekoa erabiliko dugu erreferente gisa, bi bertsoaldiak ez baitira neurri berdinean kantatuak (bigarren bertsoaldiko bertsoak lehenengokoak baino estrofa bat luzeagoak dira), eta ondorioz puntuazioen batura ez litzateke alderagarria izango. Ikus ditzagun aipaturiko gaiari erreferentzia eginez kantaturiko lehen bertsoaldiko hiru bertsoak (21, 22 eta 23. taulak).

Amets Arzallus: 1. Bertsoa	**Erlazioa**
A-8 pare horretan urte osoan obretan	0,0613
gu beraiekin zorretan bide zail ta lehorretan	0,0027
baina utzi dut Burgos atzean ta Madril ez dago bertan	0,0074
azkenerako Herrera ikusi pankarta handi batetan	0,0850
sartu naiz eta funtzionariak ez daude hitz samurretan	0,3895
biluztu behar dugula eta zer zabiltza gezurretan	0,0211
urteak dira Euskal Herria dagoela hezurretan	0,1164
Batura	**0,6834**
Batez bestekoa	**0,0976**

21 taula: Amets Arzallusen lehen bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Amets Arzallus: 2. Bertsoa	**Erlazioa**
Pentsa zuek ze komedi ze baldintza ze komeri	0,0320
ta funtzionari honeri nik esan behar bi egi	0,0612
oraintxe jarri garenez gero biak aurpegiz aurpegi	0,0565
guztia bistan dudala ez didak berriz egingo abegi	0,0210
barruko horrek ulertuko dik hala esan zidaan neri	0,0199
kanpora irten naiz ta oihu egin dut erdi gaixo erdi eri	0,1488
besarkada bat Eneko eta muxu handi bat deneri	0,1002
Batura	**0,4396**
Batez bestekoa	**0,0628**

22 taula: Amets Arzallusen bigarren bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Amets Arzallus: 3. Bertsoa	**Erlazioa**
Martxan jarri auto zaharra zaharra baina ez da txarra	0,0185
hemen goaz tirri-tarra aurrean dugu iparra	0,0149
iparralderantz Euskal Herrirantz ze pena ta ze negarra	0,0591
baina hala da Espainia aldeko justizien ezbeharra	0,0237
heldu etxera eta hartu dut paper baten zirri marra	0,1384
gutun txiki bat ta utzi ditut bi muxu ta irriparra	0,3122
hori delako nik biluzteko dakidan era bakarra	0,0717
Batura	**0,6386**
Batez bestekoa	**0,0912**

23 taula: Amets Arzallusen hirugarren bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Batez besteko neurriei erreparatzen badiegu, eta 16 (0,3434) edo 17 (0,1916) tauletako batez bestekoekin konparatzen badugu, ikus dezakegu bertsoetako bat ere ez dela emaitzotara gerturatzen. Badirudi agertzen diren hitzek ez dutela nahikoa indar emaitza onak emateko, nahiz eta hitz berak agertu gaian eta bertsoan. Baliteke hau bertsotarako ariketarekin ere lotuta egotea. Izan ere, honelako kasuetan bertsolariak bere kontakizuna asko xehatzen du, hiru bertsotan barrena garapen bat emanez. Hiru izatean kantatu beharreko bertsoak, harira zuzenean jo ordez, hasieran bestelako testuinguruen kontakizunekin hasi ohi du bertsolariak bertso-sorta. Ikus dezagun zein ondorio atera ditzakegun 24, 25 eta 26. tauletatik.

Maialen Lujanbio: 1. Bertsoa	**Erlazioa**
Pasa ziren epaiketa prozesu luze ta auzi	0,0453
espetxetik espetxera orain egiten dut jauzi	0,1273
azken hau Okañakoa ez det gutxitan ikusi	0,0120
baina gaur arte ez nintzen sentitu horren biluzi	0,0078
maitea barruan daukat bakarrik ezin det utzi	0,0003
duintasun, askatasun hitzak dizkit erakutsi	0,0317
ta eutsi nahi diet baina ia ezin diet eutsi	0,0276
zenbat tasun ezpainetan eta zeinen muxu gutxi	0,0037
Batura	**0,2556**
Batez bestekoa	**0,0319**

24 taula: Maialen Lujanbioren lehen bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Maialen Lujanbio: 2. Bertsoa	**Erlazioa**
Etxean ni izan nintzen atxiloketan lekuko	0,0660
erauzi ta eraman zuten oihan nuena gertuko	0,1043
kondena hark zenbat gezur zenbat tranpa zenbat truko	0,1571
gure hitzek epaiari baina ezin egin uko	0,4492
badakit Espainia ez dela herri legedi justuko	0,0757
akaso haserrez nere lepazaina zait puztuko	0,0029
biluztuko nauzu eta nere zorroa hustuko	0,0591
baina bihotz barrukoa ez didazu biluztuko	0,1618
Batura	**1,0762**
Batez bestekoa	**0,1345**

25 taula: Maialen Lujanbioren bigarren bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Maialen Lujanbio: 3. Bertsoa	**Erlazioa**
Kartzelaren ataria umela eta ospela	0,1625
funtzionari poliziak doazela datozela	0,0075
ba ez dizut ba men egingo ta beteko erregela	0,0244
ez zuk nahi lez sinatuko konfesioan papela	0,0143
aitortzen det samin handiz uzten dudala kartzela	0,1461
atzera etxera noa etorri naizen bezela	0,1514
esperantza daukat bera indartsu egongo dela	0,0224
badakit badakiela berekin izan naizela	0,1048
Batura	**0,6333**
Batez bestekoa	**0,0792**

26 taula: Maialen Lujanbioren hirugarren bertsoak gaiarekiko duen antzekotasun semantikoaren emaitzak

Bigarren bertsoaldi honetako lehenean, hau da, 24. taulan, aurreko bertsoaldiko edozein bertsotan (21., 22. eta 23. taulak) baino puntuazio txikiagoak lortu ditugu. Hori, lehen bertsoak gaiari erreferentzia gutxi egiten diolako izan daiteke. Izan ere, hasiera batean bere momentu emozionala aurkezten hasi ohi da bertsolaria, eta horrek gure teknika

huts egitera darama. Bestalde, bigarren eta hirugarren bertsoetan jada ikus daiteke nola aurreko emaitza hobetzen den. Aurreko bertsoa gaiari gehituz aztertzen dugulako gertatzen da hau, izan ere, 25. eta 26. tauletako 'etxea' aipatzen diren estrofak aztertuta (ikus urdinez nabarmenduak), bietatik lehenengoan puntuazio txikiagoa lortzen dela ikus dezakegu. Gure susmoak egiaztatze bidean (hitz bat aurreko bertsoan agertzeak hurrengo bertsoaren puntuazioan gorakada bat ekar dezakeela), hirugarren bertsoa bigarrenaren baldintza berdinetan ebaluatu dugu (gaia eta lehen bertsoarekiko aztertuz), eta 'etxea' agertzen den estrofaren puntuazioa 0,1514-tik 0,0423-ra jaisten da. Ondorioz, ziurta dezakegu bigarren bertsoko 'etxea' hitzaren agerpenak eragina duela hirugarren bertsoko 'etxea' hitza agertzen den estrofan, eta lotura semantiko handiago bat lortzen dugula honela. Gauzak honela, bertsoaldia aztertzerako orduan aurrez kantaturiko bertsoak gaiarekin batera konparatzeko gehitzeak bere eragina duela egiaztatu dugu.

Horrez gain, beste emaitza batzuk ulertzeko bidean beste taula bat osatu dugu. Taula horretan (27. taula) gaiaren eta gaian agertzen diren hitz esanguratsuenen ('bisita', 'kartzelara', 'funtzionarioak', 'biluzi' eta 'miatzen' izan dira hitz aukeratuak) arteko erlazioak kalkulatu ditugu emaitza batzuk ulertzeko bidean. Emaitzak aztertzen hasiz, aukeratutako bost hitzetatik bik ('biluzi' eta 'miatzen' taulan gorriz) oso puntuazio txikia lortzen dute, eta beste hirurenak handiagoak diren arren ez dira oso esanguratsuak, izan ere, 'sua' hitzak corpusean gertuen duen 'eten' hitzak 0,95 lortzen duen bitartean 'kartzelara' hitzak gaiarekiko 0,22 lortzen du. Hitz hauen gaiarekiko pisuak bertsoen erlazio semantikoan duen garrantzia aztertzeko ere baliatu dugu taula. Esaterako, 24. taulako laugarren estrofan (gorriz nabarmenduta dagoena) 'biluzi' hitza agertzen den arren lortzen den emaitza oso txikia da, eta horren eragina 'biluzi' hitzak berak gaiarekiko duen pisu txikia (0,005) izan daiteke, edota corpusean oso gutxitan agertzean.

Hitza	**Erlazioa**
kartzelara	0,2229
bisita	0,1440
funtzionarioak	0,1203
miatzen	0,0125
biluzi	0,0052

27 taula: Bertsotarako gaiaren eta gaiko hitz esanguratsuenen arteko erlazioa

5 Ondorioak eta etorkizuneko lana

Master-tesi honetan bi lan egin ditugu, itxuraz independenteak, baina bertsolaritza eta informatika uztartzen dituztenak biak. Biak ala biak azterketa lanak izan dira, eta ez aurrez aipaturiko beste lan batzuk bezala, sorkuntzakoak. Batetik, bertsotan egiteko emandako gaia *umorezkoa* ala *dramatikoa* den automatikoki hautematean oinarrituriko lana gauzatu da. Ikasketa automatikoa baliatuz, corpus etiketatu batetik abiatu, eta bertatik ikasiz, bertsotarako ariketaren erregistro zuzena zein den asmatzen saiatu gara. Bestetik, bertsoa kantatzeko emandako bi ariketa mota aukeratu ditugu, horien gaiak eta bertsolariek bat-batean erantzundako bertsoak hartu eta ariketa eta bertsoaren arteko erlazio semantikoaren neurria lortzen saiatu gara. Master-tesi berean bi lan sartu ditugunez, eta lan bakoitzak bere gorabeherak izan dituenez, ondorengo bi ataletan xeheago azalduko ditugu zeintzuk izan diren bakoitzaren ondorioak:

Gaien erregistroaren identifikazioa

Ikasketa automatiko bidez % 76ko *F-Measure* lortzera iritsi gara *Develop* corpusean. Hasiera honetako emaitzak ikusita esan genezake lana oso ongi egiten dela, izan ere, *Develop* corpuseko *baseline*-a 40 puntuko aldearekin hobetzea lortu da. Bukaerako probetarako uneko hiru sailkatzaile onenak hartu eta bozketa bidez aukeratzea erabaki da. 28 taulan ikus daitezke *Develop* eta test corpusei dagozkien emaitza onenak, eta bakoitzeko *baseline*-a.

			VOTE			
	F-Measure	**SMO**	**AVG**	**MAJOR**	**MAX**	**Baseline**
Develop	**509 att**	0,714	0,734	**0,763**	0,723	**0,3597**
	764 att	**0,762**	-	-	-	
Test	**509 att**	0,622	0,632	0,631	0,622	**0,3823**
	764 att	**0,642**	**0,642**	0,596	**0,642**	

28 taula: Ondorioetarako *Develop* eta *Test* corpusetako emaitzak

Test-eko emaitzak begiratuz gero, *F-Measure* neurria ez da % 64tik pasatzen, eta *test*-eko *baseline*-a (guztiak *dramatikoa*k direla esanik) % 38 da. Hortaz, *Develop* corpusaren gaineko emaitzekiko 12 puntuko jaitsiera izan arren, test corpuseko baseline-a 26 puntutan hobetzen da. *Test* corpuseko emaitzaren beherakada, nahiz eta banaketa ausaz egin den, *Develop* eta test corpusen arteko portzentaje diferentziak (*Develop* corpusean % 56 dira *dramatikoa*k eta testean aldiz % 62) eragin dezake hein batean. Horrez gain, 3.2.4. atalean ikusi dugun bezala, aztertzen ari garen ataza erraza ez dela argudia dezakegu. Izan ere, lortu dugun Kappa-neurria, medikuntzaren alorreko literaturan irakurri dugunez erabili ohi denarekiko oso txikia da (fidagarritasun altu batek % 80ko Kappa-neurria eskatzen du, gurea % 60 azpitik dagoen bitartean), eta nahiz eta *Observed agreement*-a ia % 81 izan, umorea oso gauza subjektiboa dela medio, ataza zaila dela esan dezakegu. Ataza honetan emaitza hobeak lortzeko bidean, etiketatutako corpus adostuago bat izatea pauso

garrantzitsua litzateke.

Antzekotasun semantikoaren azterketa

Bigarren lan honen baitan ere bi azterketa egin ditugu, eta bakoitzeko emaitzekin gauza desberdinak ondorioztatu dira. Batetik bertsolariari hitza emanda kantatzeko ariketarekin eginiko azterketan 'sua' gai gisa emanda lortu ditugu emaitza onenak. 29 eta 30 taulei erreparatuz ikus dezakegu nola 'sua' hitza agertzen diren lerroek pisu handia hartzen dutela amaierako puntuazioari begira. Bestalde, gure ustez erlazio zuzena duten 'pospolu', 'pizten' edota 'berotzen' hitzek ez dute esperotako pisurik hartzen. Hau, 'sua' hitza eta espero genituen besteak, corpusean elkarrekin oso gutxitan agertzen direlako izan daitekeelakoan gaude.

Amets Arzallus	**'Sua'-rekiko**
Pospolu batek pizten badu lehen ditxa	0,1338325739
gero bota sastraka eta zumitza	0,0656301156
tximinitikan gora doa bere gisa	0,0350185186
baina neretzat **sua** dugu bertsogintza	**0,9995695353**
bertsoa balitza **su** baten baldintza	**0,7073464394**
pitz dezagun hitza alaituz bizitza	0,0683775246
hemen su horren bueltan dantzan gabiltza	0,3949621618
Batura	**2,404736869**
Batez bestekoa	**0,343533838**

29 taula: Amets Arzallusen bertsoak 'sua'-rekiko duen erlazio semantikoa.

Maialen Lujanbio	**'Sua'-rekiko**
Hura asmakizuna homo habilisena	0,1093494743
bazun intentzioa bazuen sena	0,0449577048
bi harrik elkar jota txispa bat aurrena	0,0074977744
asmakizun haundina mendeetan barrena	0,0689423233
sua da problema zenbait basorena	**0,9966166019**
edo jaki dena berotzen duena	0,0930994749
eta bi begiradek sortzen dutena	0,0208336413
Batura	**1,3412969951**
Batez bestekoa	**0,191613856**

30 taula: Maialen Lujanbioren bertsoak 'sua'-rekiko duen erlazio semantikoa.

Bi bertsoen arteko alderaketa eginda, hasiera batean espero genuen bezala Arzallus-en bertsoak makinarentzat semantikoki gertuagokoak dira 'sua'-rekiko Lujanbiorenak baino. Hala ere, lortzen diren emaitzak pisu gutxikoak direla iruditzen zaigu. Emaitzok hobetzeko erabili dugun Euskaldunon Egunkariaren corpusaren ordez zientzietako corpus bat edota *Wikipedia* bera erabiltzea egokiagoa izan daitekeela uste dugu. Bestalde, bertsolariari gai bat ematen zaionean, badirudi zailtasunak dituela bertsolariak kantatzen duena

eta gaia lotzeko. Gai-jartze mota honetan, bertsolariak sentimenduez kantatu ohi du sarri, baliteke emaitzak hobetzeko poesia edo sentimenduei buruzko testuez osatutako corpusek laguntzea. Bi kasuetarako interesgarria izango litzateke WordNet sare semantikoa erabiliz laguntzea, *POS-tag based poetry generation with WordNet* (Agirrezabal et al., 2013a) lanean egin den gisara. Sare semantikoaren laguntza interesgarria iruditzen zaigu, sinonimo diren hitzen arteko erlazioak hobeto hautemateko, izan ere, 'altzoa' hitzari kantatzeko ariketako emaitzetan 'magala' hitza ageri den lerroak 1etik 0,01eko puntuazioa lortzen du, eta bi hitzak sinonimo izanik, oso emaitzak txikia da hori. Emaitzak orokortuta, erabili dugun corpusa eta egin beharreko analisia ez datozela bat ondoriozta dezakegu. Corpus zabalago edota zehatzagoekin emaitzek hobera egin dezaketela sumatzen dugu, eta horrez gain, bertsotarako ariketa mota bakoitzak corpus mota desberdina beharko lukeela ere esan daiteke.

Eskerrak

Azkenik, eskerrik beroenak eman nahi nizkieke master-tesi honen baitan egin diren lanak garatzerakoan behar izan dudan laguntza emateko prest egon diren Manex Agirrezabal eta Aitzol Astigarragari.

Erreferentziak

Manex Agirrezabal. Bertsobot: lehen urratsak. 2012.

Manex Agirrezabal, Inaki Alegria, Bertol Arrieta, eta Mans Hulden. Bad: An assistant tool for making verses in basque. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 13–17. Association for Computational Linguistics, 2012.

Manex Agirrezabal, Bertol Arrieta, Aitzol Astigarraga, eta Mans Hulden. Pos-tag based poetry generation with wordnet. *ENLG 2013*, page 162, 2013a.

Manex Agirrezabal, Bertol Arrieta, Aitzol Astigarraga, eta Mans Hulden. Bota bertsoa, eta guk aztertuko dugu: azken urteetako bertsolari txapelketa nagusien analisia. *Elhuyar: zientzia eta teknika*, (300):46–49, 2013b.

Bertol Arrieta, Iñaki Alegria, eta Xabier Arregi. An assistant tool for verse-making in basque based on two-level morphology. *Literary and linguistic computing*, 16(1):29–43, 2001.

Ron Artstein eta Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008. ISSN 0891-2017. doi: 10.1162/coli.07-034-R2. URL `http://dx.doi.org/10.1162/coli.07-034-R2`.

A. Astigarraga, E. Jauregi, E. Lazkano, eta M. Agirrezabal. Textual coherence in a verse-maker robot. page 15, 2014.

J Carletta. Assessing agreement on classification task: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

Deerwester, Scott, Dumais, T. Susan, Furnas, W. George, Landauer, K. Thomas, Harshman, eta Richard. Indexing by latent semantic analysis. 1990.

Mikel Osinalde Agirre. Agur-bertsoetako egitura diskurtsiboaren xerka. 2013.