Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena

Jon Alkorta[1], Koldo Gojenola[1], Mikel Iruskieta[2]

[1] Lengoaia eta Sistema informatikoak Saila, [2] Hizkuntza eta Literaturaren Didaktika Saila
IXA Taldea (UPV/EHU)

(Sentiment Analysis in Basque: a proposal from lexical level to relational discourse structure)

DOI: 10.1387/gogoa.15634

Abstract

Nowadays, opinion texts play an important role, in fact, people read opinions before they do an activity, buy a product or take a decision. However, the amount of opinion text is increasing rapidly and reading all opinions about a subject is unfeasible. 'Sentiment analysis' is a part of Natural Language Processing whose aim is to process opinion texts. This work is a part of 'sentiment analysis' and presents a first approximation to the assignment of polarity to texts written in Basque. Rhetorical Structure Theory has been used to assign different weights to text spans and an analysis has been performed at the lexical level. This method has been compared with other approaches and the results are promising.

Keywords: *Basque, sentiment analysis, Rhetorical Structure Theory, RST, lexicon.*

1. Sarrera eta ikerketaren helburuak

Hizkuntzaren Prozesamenduan (HP) pertsonen arteko edota pertsona eta makinen arteko interakzioak errazteko tresna konputazionalak sortzeko ikertzen da. Azken urteetan, 'sentimenduen analisia' izeneko ikerlerroa zabaldu da bertan eta testu baten barruan inplizitu dagoen informazio subjektiboa modu (erdi)automatikoan identifikatzea da asmo nagusia. Jendearen iritziak,

sentimenduak eta subjektibotasuna antzemateko ataza da sentimenduen analisia (Pang eta Lee, 2008).

Testu bateko informazio subjektiboa identifikatzea jendeak zerbaiti buruz zer pentsatzen duen jakiteko erabil daiteke. Esaterako, pelikulei buruzko iritziak (Pang eta Lee, 2004) edo politikari baten arrakasta neurtzeko erabil daiteke (Tumasjan et al., 2010), baita merkatal-marka batek bere produktuei buruz erabiltzaileek duten iritziak zein diren jakiteko ere (Xu et al., 2011). Beraz, sentimenduen analisirako tresnak onurak eta abantailak ekar ditzake albiste, dokumentu eta informazio andana denbora errealean aztertu behar denean. Gainera, HPko beste lan askotarako baliagarri izan daiteke, hala nola: iritzian oinarritutako laburpen automatikoan (Liu, 2012), galdera-erantzun sistemetan, edo sistema aholku-emaileetan (*recommendation systems*).

Gaur egun, teknologia berrietako eta Interneteko sare sozialak direla eta, gero eta ohikoagoa da iritzi-testuak aurkitzea eta irakurtzea. Iritzi horiek izan daitezke pelikula baten ingurukoak, produktu baten ingurukoak edo pertsona batek egindako ekintza jakin baten ingurukoak, besteak beste. Enpresak eta politikariak esaterako oso interesatuta daude jendeak (oro har, jendarteak) eurengan duten iritzian. Telebistako eztabaida baten ondoren edo kanpainaren fase ezberdinetan jendearen iritzia zein den oso baliagarria da etorkizunean erabaki hobeak hartzeko. Baina iritziak milaka dira eta ezin dira politikari baten edota politikari-talde baten irudiari buruzko iritzi guztiak irakurri. Horiek horrela, iritzi horiek automatikoki prozesatzeko beharra sortu da. Sentimenduen analisian, lan gehiena ingelesez egiten da, baina atazaren konplexutasuna dela eta, oraindik ikerlerro irekia da hainbat fenomenotan, batez ere diskurtsoan.

Euskaraz sentimenduen analisian egindako lanen artean, QWN-PPV izeneko metodoa (San Vicente et al., 2014) da erreferentziazkoa. Metodo honek hitzei lema eta adiera (*synset*) mailan polaritatea jartzen die eta gero testu osoko polaritatea kalkulatzen du.

Horiek horrela, gure helburua da automatikoki testuen sentimenduen analisia egingo duen tresnak behar duen informazio linguistikoa zein den aztertzea. Horretarako, sentimenduen analisian bereizi ohi diren hiru maila hartu nahi ditugu kontuan: a) maila lexikoa (QWN-PPV metodoak bezala), b) maila sintaktikoa eta c) diskurtso-maila.

Lan honetan maila lexikoan eta diskurtso-egituran oinarritutako metodo bat garatzeko azterketa linguistikoa egin dugu. Maila lexikoan, arloko zenbait lan erabiliko ditugu eta erlaziozko diskurtso-egituran Egitura Erretorikoen Teoria (*Rhetorical Structure Theory*, RST aurrerantzean).

Arloko egoera 2. atalean azalduko dugu. Lehenik eta behin, kategoria gramatikal bakoitzean (adjektiboak, izenak eta aditzondoak), intentsifikatzaile eta ahultzaileetan eta ezezkoetan egin den lana aztertuko dugu eta,

ondoren, erlaziozko diskurtso-egituran egin dena. Lan honen metodologia 3. atalean azalduko da; bai oinarri teorikoa eta baita lana egiteko burutu diren urratsak. Gure metodoak lorturiko emaitzak 4. atalean aztertuko dira eta beste metodoekin alderatu. Azkenik, ondorioak eta etorkizuneko lanak zeintzuk diren aipatuko dira 5. atalean.

2. Arloko egoera

Maila lexikoan, hainbat kategoria gramatikal aztertu izan dira eta kategoria horiei polaritatea ezartzeko irizpide ezberdinak proposatu izan dira.

Adjektiboa izan da gehien landu den kategoria gramatikala. Chesley et al.ek (2006) adjektiboak eta aditzak erabiltzen dituzte blogen polaritatea automatikoki sailkatzeko. Hizkuntza-ezaugarriak, aditz-motari buruzko informazioa, eta sareko tresna bat, Wictionary hiztegia, erabili dute adjektiboei polaritatea jartzeko. Adjektiboen polaritatea jartzerakoan adjektiboa definitzen duten hitzen polaritatea hartzen dute kontuan. Aditzen kasuan, aditz-motari jartzen diote polaritatea. Esaterako, adostasuna adierazten duten aditz guztiei polaritate positiboa ematen diete eta eztabaida adierazten dutenei polaritate negatiboa.

Halaber, aipatzekoa da Taboada et al.en (2011) lana. Eurek 'hazi-hitzak' (*seed words*) proposatzen dituzte eta hitzok eskala batean sartzen dituzte (–5 sentimendu negatibotik +5 sentimendu positiboraino). Testuingurua kontuan hartuz, adjektiboei polaritatea eskuz jartzen diete. Lan horretan, eskala bera darabilte izen, adberbio eta aditzekin[1].

Aurreko lanak ez bezala, Baccianella et al.ek (2010) garatutako *SentiWordNet* 3.0 tresnak automatikoki jartzen die polaritatea adjektiboei eta beste kategoria gramatikalei. *SentiWordNet* 3.0 *SentiWordNet* 1.0 tresnaren hobekuntza bat da eta automatikoki hitzen adierei (*synset*) polaritate positiboa, negatiboa edo neutrala ezartzen diete erdigainbegiratua den algoritmo bat erabiliz. Polaritatea anotatzeko modua eskalarra da, Taboada et al.ek (2011) egiten duten moduan.

Maila sintaktikoan, intentsifikatzaile edo ahultzaileei dagokienez, bi irizpide erabili izan dira sentimenduen analisian: *i*) Zenbakizko eskala. Polanyi eta Zaenenek (2006) hitz bati indarra handitzen diote intentsifikatzailea dagoenean, +1 balioa gehituz, eta indarra ahultzen duen ahultzailea dagoenean, –1 balioa kenduz. Baina metodo honek ez du intentsifikatzaile eta ahultzaileen indarra behar bezala islatzen fenomeno ezberdinekin. Irizpide honi jarraiki 'oso polita' (1 + 3 = 4); 'ikaragarri polita' (1 + 3 = 4) eta 'nahiko polita' (–1 + 3 = 2) kasuak edukiko genituzke; intentsifikatzaile edo ahul-

[1] Aditzek testuinguruaren arabera polaritate ezberdina dutela ondorioztatu dute.

tzaile guztiek indar bera ematen diote lotuta duten hitzari, nahiz eta berez horrela ez izan. *ii*) Ehuneko eskala. Intentsifikatzaile eta ahultzaileen indarra hobeto zehazteko, Taboada et al.ek (2011) ehuneko eskala erabiltzen dute. Lan horretan –%50 eta +%100 arteko eskala erabiltzen dute. Intentsifikatzaileek lotuta dauden adjektiboaren edo aditzondoaren esanahia indartzen dutenez balio positiboa dute eta ahultzaileek esanahia ahultzeagatik balio negatiboa. Adibidez, *pretty* 'nahiko' adberbioari –%10 balioa ezartzen diote eta *very* 'oso' adberbioari +%25 balioa. Hartara, intentsifikatzaile bakoitzak indar ezberdina du. Modu honetan, *zital xamarra* adjektiboa bagenu eta *zital* adjektiboak –3 balioa balu; ahultzailearen eraginez (–%30)[2], bere balio negatiboa –2, 1era murriztuko litzateke.

Ezezkoa bi modutan landu izan da. Batzuek alderantzizkatze irizpidea (*switch negation*) erabili dute. Sauríren (2008) lanean erabiltzen da eta, hor, ezezkoak berez hitzak duen polaritatea alderanzten du. Esaterako, *bikaina da* sintagmari +5 emango bagenio; *ez da bikaina* sintagmari –5 eman beharko genioke. Irizpide berari jarraituz, *ez da ona*ri –3 emango bagenio, orduan *ez da bikaina* egitura positiboari *ez da ona* egitura negatiboari baino balio negatibo handiagoa eman beharko genioke; nahiz eta *ez da bikaina* egitura *ez da ona* egitura baino positiboagoa izan. Kontraesan hori konpontzeko, beste irizpide edo ikuspegi bat dago: aldatze irizpidea (*shift negation*). Irizpide hau Taboada et al.ek (2011) proposatu zuten polaritate proportzioa mantentzeko eta ez alderantzizkatzeko. Horrela, negatiboa duen sintagmari –4 balioa ezartzen diete. Gainera, ezezkoa eta negatibodun hitza sintagma berean daudenean, ezezkoari +4 balioa jartzen diote, ezezko bateko hitz negatibo bati bere zentzu negatiboa kentzen diolako. Irizpide hau *bera ez da aditua baina ez-jakina ere ez da* perpausean aplikatzen badugu; lehenengo zatian, hitz positibo bat (*aditua*) ezezkoa duen sintagmarekin dago eta ondorioz, bere polaritateak indar positiboa galtzen du (5 – 4 = 1). Bigarren zatian, hitz negatibo bat (*ez-jakina*) ezezkoarekin agertzen da eta bere indar negatiboa galtzen du (–5 + 4 = –1). Modu honetan, alderantzizkatze irizpideko kontraesanak konpontzen dira, kasu honetan, *ez-jakina* ez izatea *aditua* ez izatea baino okerragoa (polaritatez negatiboagoa) delako eta ez alderantziz, aldatze irizpidearekin gertatuko litzatekeen bezala (hau da, *ez-jakina* ez izatea [+5], *aditua* ez izatea [-5] baino hobea izatea litzateke eta ez da horrela).

Bestalde, sentimenduen analisia egiteko, Heerschop et al.en (2011) eta Taboada et al.en (2008) lanak dira aipagarrienak RST erabiltzen dutenen artean. Bi lanotan diskurtso-egituran oinarritutako sentimenduen analisia egiten da; lehenik eta behin, testuak oinarrizko diskurtso-unitateetan (EDU)

---

[2] *Xamar* adjektiboak *zital* modifikatzen du. Beraz, $zital_{-3}$ $xamarra_{-\%30}$ = –3X (+%100 – %30/100) = –3X (70/100) = –2,1. Eragiketa honetan, +%100 hitzak berez duen esanahiaren indarra da eta ahultzaileak bere indarra %30ean jaisten du. Horregatik, kasu honetan, *zital* hitzak bere indarraren %70a du, hau da, –2,1.

segmentatuz eta bakoitzari pisu ezberdina emanez. Taboada et al.ek (2008) nukleoa eta satelitea bereizten dituzte eta lehena hartzen dute testuko zati garrantzitsutzat. Testuko zati garrantzitsuenei 1,5*X* eko pisua jartzen diete eta garrantzitsuak ez diren zatiei 0,5*X*eko pisua. Hitzetan, diskurtso-unitateek testuan duten garrantziaren arabera, horien polaritatearen garrantzia handiagoa edo txikiagoa da. Bi lanotan, lexiko-mailan oinarritutako sentimenduen analisian lortu diren emaitzak gainditu dituzte, erlaziozko diskurtso-egiturako informazioa baliagarria dela erakutsiz. Asher et al.ek (2009) ere erlazio erretorikoek sentimenduan nola eragiten duten aztertzen dute *Segmentaturiko Diskurtso-Errepresentazioaren Teoria*n (*Segmented Discourse Representation Theory* ingelesez, SDRT) oinarrituta.

Euskal hizkuntzalaritzan diskurtso-erlazioak Gómezek (1997) aipatu zituen lehendabizi SDRTn oinarrituz. Gómezen (2002) lanean, berriz, fokua eta gaia kontzeptuak ikertu zituen eta SDRT teorian oinarrituz bi kontzeptu hauen informazio banaketa egin zuen. Teoria honetako erlazio guztietatik bi erlazio (NARRAZIOA eta ZUZENKETA) hartu eta erlazio horiekin egin zuen fokua eta gaiaren arteko bereizketa. Ondoren, RST baliatuz Iruskietak (2014) laburpen zientifikoek (medikuntzan, terminologian eta zientzian) duten koherentzia-egitura zuhaitz hierarkikoekin deskribatu du; horretarako, erabili du RSTko sailkapen hedatua, 30 koherentzia erlaziok osatzen dutena.

Euskaran, San Vicente et al.ek (2014) QWN-PPV metodoa garatu dute. Metodo horrek testuetan polaritatea ezartzen die lemei eta adierei (*synset)*, Baccianella et al.ek (2010) proposaturiko metodoa erabiliz. Hauxe da metodoaren formula eta azalpena:

$$\frac{\textit{Testuko hitz positiboak + Testuko hitz negatiboak}}{\textit{Testuko hitz guztiak}} = \textit{Testuaren polaritatea}$$

Lehendabizi, lexikoaren polaritea automatikoki sortzen da Q-WordNet erabiliz eta ondoren, lexikoi horretako hitzak testuan bilatzen dira eta testuko hitzei definizioz ematen zaien polaritatea ezartzen zaie Q-WordNeti esker. Ondoren, polaritate positibodunen eta negatibodunen batuketa egin eta batuketaren emaitza testuan dagoen hitz kopuruarekin zatitzen da bataz bestekoa lortuz.

Alkorta et al.ek (2015) QWN-PPV metodoa eta erlaziozko dirkurtso-egitura erabiliz, sentimenduen analisiko emaitzak hobetu dituzte. Lan horretan 28 testuko corpusa bildu eta QWN-PPV metodoa aplikatu dute bi modutara. Batean, tresna bere horretan aplikatu da. Bestean, diskurtso-egiturako informazioa aplikatuz erabili da tresna hori. Diskurtso-egitura finkatzeko RST teoria oinarritzat hartuta bere unitate zentrala (testuko gai nagusia) eta EBALUAZIOA erlazio erretorikoa erabili dira eta 3 kategoriatako polaritatea (negatiboa, neutrala eta positiboa) aplikatu da. Diskurtso-egitura erabilita

lortu den emaitza 0,84ko F-neurria da eta lexiko mailako informazioa soilik erabilita 0,59koa; beraz, hobekuntza nabarmena egon da.

3. Metodologia

Atal honetan, lehenik eta behin Egitura Erretorikoen Teoria azalduko dugu eta ondoren, metodoa garatzeko burututako urratsak.

3.1. Oinarri teorikoa: Egitura Erretorikoen Teoria (RST)

Lan honetako oinarri teorikoa da Egitura Erretorikoen Teoria (RST) (Mann eta Thompson, 1988, 1987). Teoria horrek testuen egitura koherentziaren bidez deskribatzen du. Hasiera batean testu-sorkuntzarako *(text-generation)* sortu zen, nahiz eta gerora beste atazetarako erabili den, tartean sentimenduen analisia. Teoria honen bitartez hainbat hizkuntzatako eta hainbat domeinu edo eremutako testuak landu izan dira (ingelesezko kazetaritza testuak Carlson et al.ek (2002); alemanez kazetaritza testuak Stedeek (2004), portugesez informatikari buruzko testu zientifikoak Pardo eta Senok (2005), gaztelaniazko gai ezberdinetako testu zientifikoak da Cunha et al.ek (2011)) eta medikuntza, terminologia eta zientzia domeinuetako 60 testu-laburpen euskaraz Iruskieta et al.ek (2013).

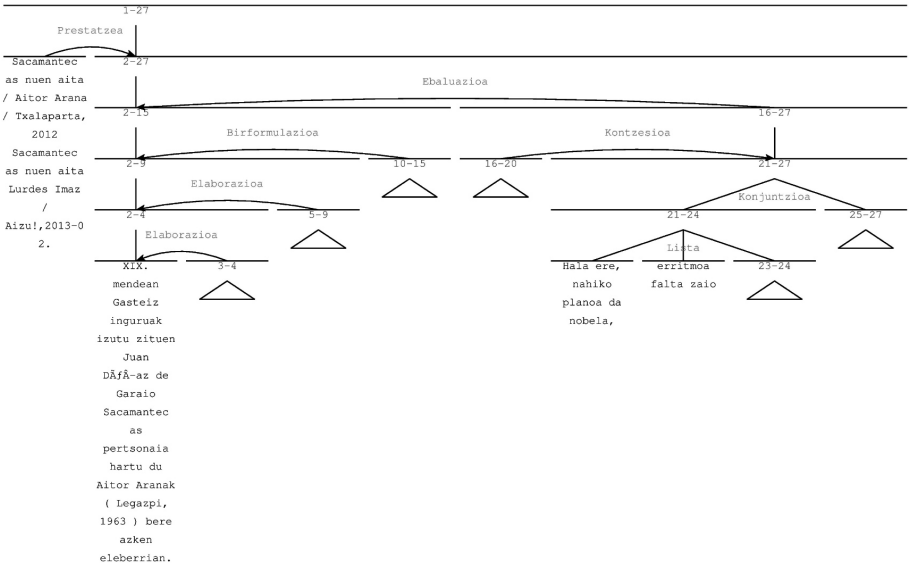
Iruskieta et al.ek (2013) aipatzen duen moduan, teoriaren abiapuntuan testuaren segmentazioa dago. Segmentatutako testu-zati bakoitza diskurtso-unitatetzat (*discourse unit)* hartzen da. Hurrengo urratsean, testuko unitate zentrala identifikatu behar da (testuko gai nagusia identifikatzen da) eta, azken urratsean, diskurtso-unitateek beste diskurtso-unitateekin duten harremana finkatzen da (erlazio-egitura deskribatzen da). Harreman horiei koherentziako erlazio (*coherence relation)* esaten zaie; RSTn, zehazkiago, erlazio erretoriko (*rhetorical relation)*.

Diskurtso-unitate bakoitzak testuan duen garrantzia nukleartasunaren (*nuclearity)* araberakoa da. Erlazio-egituran diskurtso-unitate guztiek ez dute garrantzi bera. Diskurtso-unitate bat bestea baino garrantzitsuagoa denean, garrantzitsuago denak nukleo-unitate (*nuclear unit*, N) izena hartzen du eta garrantzi gutxi duenak satelite-unitate (*satellite unit*, S).

Erlazio erretorikoak bi motatakoak izan daitezke. Nukleo bat eta satelite bat (N-S) erlazio erretoriko baten bidez lotuak daudenean, nukleo bakarreko (*nuclear unit)* erlazio bat dago eta honi *erlazio hipotaktikoa* deitzen zaio. Bestalde, bi nukleo loturik daudenean (N-N), erlazio nukleoaniztun bat (*multinuclear)* dago eta erlazio horrek *erlazio parataktiko* izena hartzen du. Erlazio nukleobakardunak (N-S) bi motatakoak izan daitezke: *a*) edukikozko erlazioak (*subject matter)*, diskurtso-unitateen artean erlazio bera dagoenean eta

b) aurkezpenekoak (*presentational)*, diskurtso-unitateen arteko erlazioak irakurlearengan eragin nahi duenean.

RSTko zuhaitz baten egitura partziala ageri da 1 Irudian.[3] Ezkerraldean, testuko diskurtso-unitaterik garrantzitsuena dago: unitate zentrala (*central unit*), ez baitago beste unitate batekin lotuta. Irudiaren eskuin aldean, LISTA eta KONJUNTZIOA erlazio erretorikoak daude eta hauek bi nukleo-unitate edo gehiago (21, 22 eta 23-24 LISTAk eta 21-24 eta 25-27 KONJUNZIOAk) lotzen dituzte, hau da, garrantzi bera duten bi diskurtso-unitate lotzen dituzte. KONJUNTZIOA erlazio erretoriko honen gainean, kontuan hartu behar da RST errekurtsiboa denez, beste erlazio erretoriko bat dago: KONTZETSIOA erlazioa. Erlazio hori nukleobakarra da, hau da, nukleo-unitate bat eta satelite-unitate bat daude. Nukleo-unitatea 21-27 diskurtso-unitatea da eta satelite-unitatea 14-20 diskurtso-unitatea.



1. Irudia

AIZ03 RST-zuhaitz partziala

1 irudiko testu-zatia literatura kritika baten zati bat da eta bere erlazio-egitura bat dator Horvath eta Egginsek (1995) deskribatu duten iritzi-testuen makroegiturarekin. Horien arabera elkarrizketetan iritzi-testuen egitura nahiko erregularra da eta iritzi nagusia beti hasieran agertzen da. Iritzi-

[3] Zuhaitz horren egitura osoa hemen ikus daiteke: http://ixa2.si.ehu.es/diskurtsoa/diskurtsoa_jpg/SENTAIZ03-A1. jpg.

testuetan honako egiturak bereizten ditu beti ere iritzia hasieran dela eta ondoren, zenbait ebaluazio.

i) IRITZIA-ERREAKZIOA-(EBIDENTZIA)-(ERRESOLUZIOA)
ii) IRITZIA-ERREAKZIOA
iii) IRITZIA-ERREAKZIOA: adostasuna-EBIDENTZIA
iv) IRITZIA-ERREAKZIOA-EBIDENTZIA-ERRESOLUZIOA

Beraz, RSTrekin irudikatutako adibidearen hasieran dago unitate zentrala (irudian ezkerrean) eta berari lotzen zaizkio EBALUAZIOzko erlazio erretorikoak. Beraz, sentimenduen analisia egiteko unitate zentrala diskurtso-unitateak eta EBALUAZIOAk lagunduko luketela uste dugu.

3.2. Ikerketako urratsak

Lan honetan, honako urrats hauek egin ditugu maila lexikoan oinarritutako metodoa hobetzeko.

i) Corpusa osatu. Alkorta et al. (2015) laneko corpus bera erabili dugu. Corpus hori Kritiken Hemeroteka[4] webguneko testuekin osatu da eta RSTko erlazio erretorikoekin etiketatu da. Guztira 28 testuko corpusa osatu dugu; 1038 oinarrizko diskurtso-unitate, 1008 erlazio erretoriko eta 8823 hitzekin. Testuak luzera ezberdinetakoak dira. Testurik laburrenak 108 hitz ditu eta luzeenak 485 hitz. Corpusa Euskal RST *Treebank*ean (Iruskieta et al., 2013) kontsulta daiteke.[5] Hasiera batean, 30 testuko corpusa osatzea pentsatu zen, polaritate kategoria bakoitzeko (oso negatiboa, negatiboa, neutrala, positiboa eta oso positiboa) sei testu baina polaritatea negatiboko testuak biltzea zaila izan zen eta azkenik, bost testu negatibo eta bi oso negatibo bildu genituen. Euskarazko iritzi-testu negatiboak bilatzeko zailtasuna Egañak (2013) emandako datuetan ere islatzen da, berak dioenez euskal literaturako iruzkinen %84k balorazio baikorra baitute. Alkorta et al. (2015) lanean bildutako 28 testuetatik 19 erabili ditugu lan honetan.

1. Taula

Corpusaren deskribapena

Testua	Dokumentuak	EDUak	Erlazio erretorikoak	Hitzak
Literatura kritikak	28	1.038	1.008	8.823

[4] http://kritikak.armiarma.eus/
[5] Euskal RST *Treebank*a hemen kontsulta daiteke: http://ixa2.si.ehu.es/diskurtsoa/.

ii) Testuaren batez besteko polaritatearen urre-patroia. Gure metodoaren emaitzak berez lortu beharko liratekeenarekin alderatzeko sortu dugu urre-patroia. Urre-patroia lortzeko, Google Forms formatuan oinarrituta 28 testuei dagozkien galdetegiak osatu ditugu, testuko galdetegi bana. Galdetegian, lehenengo zatian etiketatzaileari buruzko informazioa dago; ondoren, EDUetan segmentatutako testua dago, eta azkenik testuari buruzko zenbait galdera, unitate zentrala eta testuko polaritatea zein den erabakitzeko. Polaritateko galderetan, batetik bosteko eskala batean (bat oso negatiboa eta bost oso positiboa) testuaren polaritatea etiketatzaileen ustez zein den galdetzen zaie, hau da, polaritate kuantitatiboa. Galdetegi bakoitza 10 pertsonek bete ostean, testu bakoitzaren batez besteko polaritatea kalkulatu dugu; baita zein izan den eskalan zenbakirik hautatuena. Polaritate kualitatiboa ere kalkulatu dugu, hau da, ea testua positiboa, neutrala edo negatiboa den.

iii) Corpuseko elementu lexikoen polaritate automatikoa. QWN-PPV metodoak automatikoki hainbat hizkuntzatarako, tartean euskararako, polaritatedun lexikoiak sortzen ditu Q-WordNet erabiliz. Testuei metodo hau aplikatu aurretik beharrezkoa da testuok lematizatzea eta horretarako Eustagger (Ezeiza et al., 1998) tresna erabili dugu. Testua lematizatuta dagoenean, metodo horrek hitza Q-WordNetetik eratorritako lexiko-zerrendan dagoen ikusten du eta baldin badago zerrendan duen polaritatea jartzen dio (positiboa: +1 edo negatiboa: –1) 2 Taulako ezkerreko AIZ03 testuan ikusten den bezala. Taulan kolorez ikusten den bezala, metodo hau gauza da hitza bere ezaugarrien arabera sailkatzeko: berdez polaritate positiboko hitzak, gorriz polaritate negatibokoak, laranjaz neutralak, morez intentsifikatzaileak, lilaz ahultzaileak eta urdinez aldatzaileak.

Metodoak testuko hitzei polaritatea jarri ondoren, polaritatea kalkulatzen du modu honetara: hitz positiboen eta negatiboen batura testuko hitz kopuruarekin zatitzen da. 2 Taulako ezkerreko testuan polaritatea duten 46 hitz daude eta 204 hitz guztira testuan. Datu hauetatik abiatuz egindako zatiduraren emaitza +0,025 da, beraz, hori da testuaren polaritatea.

iv) Metodoaren ebaluazioa. Ebaluazioa egiteko eskala hamar mailatakoa izango da. Testu baten polaritatea bostetik beherakoa bada, testuak polaritate negatiboa izango du, bost eta sei bitartean neutrala eta seitik gorakoa bada positiboa.

Ebaluazioa egiteko; bi metodoek ematen dituzten emaitzak egokitu egin behar izan ditugu. Pang eta Leek (2005) dioen bezala, arazotsua da eskala ezberdinak konparagarri bihurtzea, faktore asko baitaude kontuan. Guk bi metodoek ematen dituzten emaitzak eskala batean kokatzeko honako hau egin dugu: corpuseko testuek dituzten emaitza maximoa eta minimoa hartu ditugu kontuan eta testuen emaitzak maximo eta minimo horretatik ateratako eskalan kokatu ditugu. On-

dorioz, emaitzetan beti egongo da testu bat 10eko polaritatea duena eta beste bat 0ko polaritatea duena.

Urre-patroiaren arabera, AIZ03 testuak polaritate negatiboa du, hamarreko eskala batean, 4,64ko polaritatea duelako eta QWN-PPV metodoak polaritate negatiboa jarri dio (0,71) testu berari. Kasu honetan, bi emaitzetan polaritate negatiboa dago baina emaitzen arteko aldea handia da (4,64 –0,71 = 3,93). Diskurtsoa aplikatzen den metodoak ere testu berari 5,20ko polaritatea ematen dio. Kasu honetan, berriz, aldea txikia da (0,56) baina testuari ez dio polaritatea ondo esleitzen.

v) Polaritatearen eskuzko etiketatzea. Maila lexiko-sintaktikoko eta diskurtsoko elementuen polaritatea eskuz etiketatzeko, honako hiru urrats hauek eman dira:

a) Polaritatedun hitzak identifikatu eta elementu bakoitzari balioa ezarri. Polaritatean eragiten duten hitzak azpIindize bidez adierazi dira: hitz positiboak$_{POS}$, hitz negatiboak$_{NEG}$, ezezkoak$_{EZ}$ eta intentsifikatzaileak$_{INT}$. Ez dira polaritatedun hitz guztiak identifikatu, kategoria gramatikal jakin batzuetakoak baizik: adjektiboak; izenak eta adberbioak; intentsifikatzaileak eta ezezkoak. Lan honetan ez ditugu aditzak kontuan hartu.[6]

Polaritatea jartzeko irizpideak Taboada et al. (2011) lanean azaltzen direnak izan dira. Polaritatea jartzerakoan, hitzak agertu diren testuingurua hartu dugu kontuan. Etiketatzeko irizpideak 3 Taulan aurkezten ditugu.

Adjektibo, izen eta adberbioei, –5 eta +5 arteko polaritatea jarriko zaie; intentsifikatzaileek eta ahultzaileek –%50 eta +%100 artean indartu edo ahuldu ahal izango dute lotuta duten hitzaren polaritatea eta ezezkoak –4 eta +4 edo –2 eta +2 arteko balioa izan dezake testuinguruaren arabera. Intentsifikatzailea/ahultzailea eta ezezkoaren kasuetan izan ezik, beste guztietan testuingurua hartu da kontuan hitzei polaritatea jartzerakoan. Adjektibo, izen eta adberbioei polaritatea jartzerakoan, lehenik eta behin testuinguruak adierazten duen sentimendua hartu da kontuan eta ondoren, hitz horren antzeko esanahia duten hitzak. Esaterako, *aspirina ona da sendatzeko* perpausan, testuinguruagatik 'ona' positiboa dela zehaztu dugu eta kontuan hartuta antzeko esanahia dutenen artean 'bikain' egon daitekeela eta honek intentsitate gehiago duela, 'ona' hitzari +3 polaritatea jarri diogu eta 'bikain' hitzari +5.

Intentsifikatzaile/ahultzaileen ehunekoa intentsifikatzaile/ahultzaileek beste intentsifikatzaile/ahultzaileekin duten posizioan oinarrituz finkatu da. Hau da, 'nahiko' ahultzaileak 'zerbait' ahultzaileak baino indar gehiago du, beraz, –%50 balioa jarri zaio ('zerbait' ahultzaileari

[6] Izan ere, Taboada et al.en (2011) arabera, aditzena fenomeno konplexua da.

2. Taula

AIZ03 testuaren polaritatea automatikoki eta eskuz etiketatua

Sacamantecas nuen aita / Aitor Arana / Txalaparta, 2012. Sacamantecas nuen aita / Lurdes Imaz / Aizu!, 2013-02	
Automatikoa	Eskuzkoa
Sentimendua duen hitz kopurua: 46 **Polaritate kalifikazioa:** 0,025423728 **Polaritatea (atalasea – > 0,0):** positiboa	
XIX. mendean Gasteiz inguruak izutu zituen$_{POS}$ Juan Díaz de Garaio Sacamantecas pertsonaia hartu du Aitor Aranak (Legazpi, 1963) bere azken eleberrian. Sei emakume bortxatu$_{NEG}$ eta erail$_{NEG}$ zituen$_{POS}$ arabarrak 1870 eta 1879 bitartean, eta garrotez hil$_{NEG}$ zuten$_{POS}$ 1881ean. Munstroa$_{NEG}$ gizatiartzeko asmoz edo, Adela alabaren ikuspuntua aukeratu du idazleak Sacamantecasen istorioa berreraikitzeko. Santa Klara komentuan moja, aitari buruzko egunkari txatalak errekuperatuko$_{POS}$ ditu$_{POS}$, eta haietatik abiatuta Daz de Garaioren istorioa kontatu **nahiko**$_{INT}$ du$_{POS}$, bere galbahetik pasatuta$_{ADB}$.	XIX**. mendean Gasteiz inguruak izutu zituen Juan Díaz de Garaio Sacamantecas pertsonaia hartu du Aitor Aranak (Legazpi, 1963) bere azken eleberrian.** Sei emakume bortxatu eta erail zituen arabarrak 1870 eta 1879 bitartean, eta garrotez hil zuten 1881ean. Munstroa$_{NEG}$ gizatiartzeko asmoz edo, Adela alabaren ikuspuntua aukeratu du idazleak Sacamantecasen istorioa berreraikitzeko. Santa Klara komentuan moja, aitari buruzko egunkari txatalak errekuperatuko ditu, eta haietatik abiatuta Daz de Garaioren istorioa kontatu nahiko du, bere galbahetik pasatuta.
Lanetik etxerakoan, Gasteiz inguruko errepideetan harrapatzen$_{NEG}$ zituen emakumeak bortxatzen$_{NEG}$ eta erailtzen$_{NEG}$ saiatzen$_{POS}$ zen Díaz de Garaio, baita lortu$_{POS}$ ere sei aldiz. Zahartuz zihoala, gero eta zailagoa$_{NEG}$ egiten zitzaion basatikeriak egitea, eta horri esker$_{POS}$ harrapatu$_{NEG}$ zuten$_{POS}$. Pasarte horiek eta alabaren kontaketak biltzen ditu$_{POS}$ nobelak 16 kapitulutan zehar. Euskal Herriko hiltzailerik$_{NEG}$ famatuenaren$_{POS}$ ibilerak ezagutzeko aukera$_{POS}$ ematen digu$_{POS}$ Aranak, baita garai$_{POS}$ haietako ohiturak ere: debekatuta$_{NEG}$ zegoen makila-jokoa, «bideetako mari» izena zuten$_{POS}$ prostituten$_{NEG}$ bizimodua, eta abar. Lan historiko interesgarria$_{POS}$ egin du$_{POS}$ idazleak pertsonaia horri buruzko dokumentazio bildu eta irakurleen esku$_{POS}$ ipintzeko.	Lanetik etxerakoan, Gasteiz inguruko errepideetan harrapatzen zituen emakumeak bortxatzen eta erailtzen saiatzen zen Díaz de Garaio, baita lortu ere sei aldiz. Zahartuz zihoala, gero eta zailagoa$_{NEG}$ egiten zitzaion basatikeriak$_{NEG}$ egitea, eta horri esker$_{POS}$ harrapatu zuten. Pasarte horiek eta alabaren kontaketak biltzen ditu nobelak 16 kapitulutan zehar. [Euskal Herriko hiltzailerik famatuenaren$_{POS}$ ibilerak ezagutzeko aukera ematen digu Aranak, baita garai haietako ohiturak ere: debekatuta$_{NEG}$ zegoen makila-jokoa, «bideetako mari»izena zuten prostituten bizimodua, eta abar. Lan historiko interesgarria$_{POS}$ egin du idazleak pertsonaia horri buruzko dokumentazio bildu eta irakurleen esku ipintzeko.
Hala ere, **nahiko**$_{INT}$ planoa da nobela, erritmoa falta$_{NEG}$ zaio eta bortxaketen$_{NEG}$ kontaketak aspergarriak$_{NEG}$ ere bihurtzen dira, ankerkeria$_{NEG}$ horien narrazioak inoiz$_{POS}$ horrela deskribatu badaitezke, behintzat. Bestalde, alabaren ikuspuntua ez da batere$_{NEG}$ argi$_{POS}$ geratzen, eta ez dio askorik$_{INT}$ eransten kontaketari; sobera dagoela esan daiteke.	*Hala ere, nahiko*$_{INT}$ *planoa*$_{NEG}$ *da nobela, erritmoa falta zaio eta bortxaketen kontaketak aspergarriak*$_{NEG}$ *ere bihurtzen dira*, ankerkeria$_{NEG}$ horien narrazioak inoiz horrela deskribatu badaitezke, behintzat. *Bestalde, alabaren ikuspuntua ez da batere argi*$_{POS}$ *geratzen,* eta ez dio askorik$_{POS}$ eransten kontaketari; sobera$_{NEG}$ dagoela esan daiteke.]$_{EBALUAZIOA}$

–%25) Ezezkoak bi taldetan banatu dira; alde batetik, 'ez (ezik)' –4tik +4ra balioekin eta beste aldetik, 'barik' , 'gabe' , 'eza', 'ezinik' eta 'ezin' –2tik +2ra balioekin. Bereizketa egin dugu lehenengo taldeak polaritatea indartsuagoa duelako eta perpaus osoei eragiten dietelako eta besteek hitzari edo gehienez sintagmari eragiten dietelako.

Ezezkoak perpaus edo sintagma berean polaritatea duten hitzei eragiten die, duten balioa igoz edo jaitsiz. Adibidez, 'zoragarria' (+5) ezezkoarekin 'ez da zoragarria' (5 – 4 = +1) da; hau da,

3. Taula

Elementu lexiko-sintaktikoak etiketatzeko irizpide orokorrak

Elementu lexiko-sintaktikoa	Polaritatea
Adjektiboa	–5etik +5era
Izen eta adberbioak	–5etik +5era
Intentsifikatzaileak/ahultzaileak	–%50etik +%100era
Ezezkoak	–4tik +4ra edo –2tik +2ra

zoragarria izatea hobea (polaritatea altuagoa) da ez izatea baino; eta 'inuzentea' (–3) ezezkoarekin 'ez da inuzentea' (–3 + 4 = +1); beraz, hobea inuzentea ez izatea izatea baino. 'Ez' partikula indartsuena da perpaus osoari eragiten diolako baina badaude beste batzuk eragin txikiagoa dutenak 'barik', 'gabe' eta 'eza' besteak beste. Kasu hauetan eragina sintagmara mugatzen da: 'laguntza' (+3) eta 'laguntza gabe' (3 – 2 = +1). Hau da, 'ez' partikularekin bezala, hobea da (polaritatea altuagoa) laguntza edukitzea ez edukitzea baino.

AIZ03 testuan, irizpide horiek aplikatuta 4 Taulako hitz-zerrendak eta balioak lortu ditugu. Lehenengo zutabeetan erlazio-egituran goren dagoen EBALUAZIOko elementuak (EBAL.) jarri ditugu eta bigarrenean elementu horien polaritate-balioa (Pol.). Hirugarren eta laugarren zutabeetan, bestalde, erlazio-egiturako beste EBALUAZIOetan daudenak eta horien balioak. Bukatzeko, bosgarren zutabean erlazio-egiturako beste diskurtso-unitateetan dauden elementuak eta balioak.[7] 4 Taulan ageri den moduan, polaritatedun hitz batzuk EBALUAZIO gorena eta beste EBALUAZIOa erlazio erretorikoetan agertzen dira aldi berean; EBALUAZIO gorena erlazio erretorikoaren barruan beste EBALUAZIO erlazio erretoriko bat dagoelako.

[7] AIZ03 testuko unitate zentralean aztertzen ari garen polaritatedun elementurik ez dagoenez, ez dugu zerrenda honetan jarri. Aipatu dugunez, polaritate negatiboa lukeen 'izutu' aditza ez da lan honetako aztergaia.

4. Taula

AIZ03ko elementu lexikoen eta intensifikatzaileen polaritatea

EBAL.	Pol.	Beste EBAL.	Pol.	Beste EDUak	Pol.
interesgarri	+5	interesgarri	+5	munstro	–5
asko	+5			zail	–5
famatu	+4			basatikeria	–5
debekatuta	–3			esker	+3
nahiko	–%50	nahiko	–%50		
plano	–3	plano	–3		
aspergarri	–5	aspergarri	–5		
ankerkeri	–5				
ez	–4	ez	–4		
argi	+5	argi	+5		
ez	–4				
sobera	–3	sobera	–3		

Urrats honetan maila lexiko-sintaktikoko polaritate-balioen eragiketak egin ditugu. Esaterako, ezezkoa eta horrek modifikatzen duen hitzak duen balioa ebatzi dugu (1 eta 2), baita intentsifikatzailearen eragina ere (3).

(1) Alabaren ikuspuntua ez_{-4} da $argi_{+5}$ geratzen. –4 + 5 = +1
(2) Ez_{-4} dio $askorik_{+5}$ erakusten kontaketari. –4 + 5 = +1
(3) $Nahiko_{-\%50}$ $planoa_{-3}$ da nobela. (+%100 – %50/100)X – 3 = (50 /100)X – 3 = –1.50

Batetik, 'argi' eta 'asko' hitzen kasuan ezezkoarengatik +5eko balioa izatetik +1 balioa izatera igaro dira (1) eta (2) adibideetan. Bestetik, intentsifikatzailearen eraginez, 'planoa' hitzak indarra galdu du eta –3 balioa izatetik –1,50 balioa izatera igaro da (3).

b) Erlaziozko diskurtso-egitura etiketatu. Corpusa eskuz etiketatu da RSTrekin. Polaritatea erauzteko, ordea, ez dira corpuseko erlazio erretoriko eta diskurtso-unitate guztiak hartu kontuan. Erlazioei dagokienez, EBALUAZIOa[8] erlazio erretorikoan ageri baita testuko polaritaterik esanguratsuena (Alkorta et al., 2015). Diskurtso-unitateei dagokienez, unitate zentralarekin adierazten da testuko gai nagusia non dagoen eta berau ere kontuan hartu behar da, gai nagusiari buruz ematen diren beste iritziak baino garrantzitsuagoak direlako. Horregatik, testuko unitate zentrala belztu, EBALUAZIO

[8] EBALUAZIOaren definizioa hau da (Mann eta Taboada, 2005): sateliteko eta nukleoko arauetan «idazleak nukleoarekiko duen aldeko **iritzia** aurkezten du sateliteak».

gorenaren satelitea kako artean sartu eta beste EBALUAZIOen sateliteak azpimarratu ditugu 2 Taulan.

c) Elementu lexikosintaktikoei erlazio-egiturako polaritate balioak aplikatu. 5 Taulan, diskurtso-egituraren pisua aplikatu diegu polaritatedun hitzei. Unitate zentralari 1*X* eko pisua jarri zaio, EBALUAZIO gorenari 0,7*X* koa; beste EBALUAZIOei 0,5*X* koa eta beste erlazio erretoriko edo diskurtso-unitateei 0,2*X* ko pisua jarri zaie.[9]

5. Taula

Elementu lexikoak eta euren pisua
diskurtso-egituran duten posizioaren arabera

Lexikoa	Hiztegia	UZ	EBAL. gorena	beste EBAL.	beste EDUak
interesgarri	+5		+3,5	+2,5	
asko	+1		+0,7		
famatu	+4		+2,1		
debekatuta	−3		−2,1		
plano	−1,5		−1,05	−0,75	
aspergarri	−5		−3,5	−2,5	
ankerkeri	−5		−3,5		
argi	+1		+0,7	+0,5	
sobera	−3		−2,1	−1,5	
munstro	−5				−1
zail	−5				−1
basatikeria	−5				−1
esker	+3				+0,6

5 Taulan diskurtso-egiturak hitzen polaritatean duen eragina nabari daiteke, EBALUAZIO gorena, beste EBALUAZIOak eta beste EDUak zutabeetan. Esaterako, ‘interesgarri$_{+5}$ ’,‘aspergarri$_{-5}$’ eta ‘munstro$_{-5}$’ hitzei diskurtso-egiturako pisua aplikatuta, polaritate hau ezartzen diegu: ‘interesgarri$_{+3,5}$ ’ (hitzaren polaritatea +5etik +3, 5era igaro da EBALUAZIO gorena erlazioan dagoelako); ‘aspergarri$_{-2,5}$’ (hitzaren polaritatea −5etik −2, 5era bestelako EBALUAZIO erlazioan dagoelako) eta ‘munstro$_{-1}$’ ean bihurtu da, beste diskurtso-unitate batean dagoelako.

Testuaren polaritatea lortzeko, 6 Taulan ikusten den bezala, diskurtso-erlazio eta diskurtso-unitateen polaritatearen batura egin dugu.

[9] AIZ03 testuko unitate zentralean ez dago polaritatedun hitzik.

6. Taula

Diskurtso-erlazio edo EDU bakoitzaren polaritatea

Diskurtso-erlazioa edo unitatea	Polaritatea
Hiztegiko balioak	–19,5
Unitate Zentrala	0
EBALUAZIO gorena	–1,75
Beste EBALUAZIOak	–1,75
Beste EDUak	–2,40
Guztira	–5,9

Azkenik, testuaren polaritatea eskala batean jartzeko, honako urrats hauek egin ditugu (guk proposatutakoan, hau da, diskurtsoan oinarritutakoan nahiz QWN-PPV metodoan:

i) Testuetan lortu diren polaritate maximoa eta minimoa identifikatu ditugu. Gure kasuan, polaritate maximoa +42,2 izan da (EIB01 testuan) eta minimoa –54,2 (PUT01 testuan).
ii) Ondoren, polaritate maximoa eta minimoaren arteko aldea kalkulatu dugu. Aldea 96,4koa izan da (42, 2 + 54, 2 = 96, 4).
iii) Azkenik formula hau aplikatu dugu:

$$\frac{54{,}2 + (x)}{9{,}64} = \textit{Polaritatea eskalan kokatuta}$$

Formula hori AIZ03 testuan aplikatuta, $\frac{(+54{,}2)+(-5{,}9)}{9{,}64}$ lortu den emaitza +5,01 izan da. Beraz, AIZ03 testuaren polaritatea gutxigatik positiboa da guk proposatutako metodoaren arabera.

4 . Emaitzak

4.1. QWN-PPV metodoaren errore-analisia

QWN-PPV metodoa maila lexikoko eskuzko etiketatzearekin konparatu ostean, QWN-PPVk zenbait akats egiten dituela ohartuko gara. Hauek dira bertatik ateratako ondorio nagusiak:

a) Polaritate okerra. Aditz nagusi batzuek polaritate okerra dute. Taboada et al.ek (2011) egiten duten moduan metodo honek ere jartzen dio polaritatea aditz nagusiari. Urre-patroiarekin lortutako emaitzak ez datoz QWN-PPV metodoarekin bat; bada, balorazioan oinarrizkoa den banaketa ez du ondo egiten: balorazio positiboa duten testuei balorazio negatiboa jarri die, zehazki ARG02 (8,40), ARG04 (5,20) eta

HIR01 (8,20) fitxategiek urre-patroian balorazio positiboa dute, baina QWN-PPV metodoak balorazio negatiboa jarri die (3,57, 2,26 eta 4,28, hurrenez hurren).

b) Jarri gabekoak. Subjektibitatea duten hitz ia guztiei jartzen die polaritatea metodoak; hala ere, tarteka egiten du hutsen bat. PUT01 testuan 'xaxi-idaxle' ortografia normalizatu gabe duen hitz-elkartua polaritate gabe agertzen da eta negatiboa eduki beharko luke (beste testuetan baino arruntagoak dira helburu komunikatibodun hitz hauek). COR01 testuan, ordea, beste mota bateko arrazoi bategatik 'erasokorra' hitza polaritate gabe agertzen da, polaritate negatiboa dagokionean. AIZ03 testuan, berriz, 'nahiko' intentsifikatzailea antzeman du baina ez 'planoa' adjektiboa.

c) Gehiegi jarritakoak. Aditz laguntzaileei polaritatea jartzen die. Aditzek badute polaritatea, nahiz eta normalean ez den «hazi-hitzena» bezain indartsua; baina kasu honetan, aditz laguntzaileari jartzen zaio polaritatea eta hori ez da zuzena. AIZ03 testuan, esaterako, hainbat aldiz 'zituen' aditz laguntzaileari polaritate positiboa jarri zaio. Beste zenbait testutan ere polaritate positiboa jarri zaie aditz-laguntzaileei: 'du', 'digu', 'ditu'.

d) Entitateak. Metodoak ez ditu entitateak antzematen. Metodoak entitateak antzematen ez dituenez, entitate hitzei ere polaritatea jartzen die. BER06 testuan, «Hutsegiteen garaia» liburuaz hitz egiten da; baina 'hutsegiteen' hitzak polaritate negatiboa du, eta 'garaiak', berriz, polaritate positiboa. Beraz, beharrezkoa da metodoari entitateak antzemateko sistema bat gehitzea; entitate elementu bakoitzari polaritatea ez jartzeko, batez ere polaritate biak (informazio kontraesankorra) ez jartzeko.

e) Diskurtso-markatzaileak. Metodoak ez ditu diskurtso-markatzaileak kontuan hartzen, nahiz eta hauek testuaren polaritatean eragin dezaketen. BER05 testuan, adibidez, «bizitzan eroso sentitu nahi baina bere burua etengabe atsekabetzea baino lortzen ez duten horietakoa» testua dago. Lehen zatia («bizitzan eroso sentitu nahi») positiboa da baina bigarren zatiak («bere burua etengabe atsekabetzea»), negatiboa denak, 'baina' juntagailuak lehen zatiaren polaritatea 'indargabetzen' du eta beraz, perpausean nagusitu behar den balorea negatiboa da. Metodoak, ordea, perpausaren lehen eta bigarren zatiak modu berean tratatzen ditu. Hori horrela, beharrezkotzat jotzen dugu lokailuak kontuan hartzen dituen sistema bat garatzea, emaitzak ahalik eta zehatzenak izateko.

f) Elementu mailako polaritatea. Metodoak ez du testuaren egitura kontuan hartzen. Testuko hitz eta perpaus guztiek garrantzi bera balute bezala jokatzen du metodoak, diskurtsoa baztertuz. Esaterako, garrantzi bera dute adibideetako hitzek eta testua laburbiltzen duen perpauseko hitzek. FAR01 testuan, «gustura» hitza testuaren unitate zentralean dago eta +1 balioa du. Testu berean, «dibertigarriagoa» hitza

erlazio batean satelitean dago eta honek ere +1 balioa du. Kontraesan hori konpontzeko, RST darabilgu eta teoria horretan oinarrituta, testuko hitzei pisu ezberdina jarri diegu.

4.2. Gure metodoaren emaitzak eta errore-analisia

Guk proposatzen dugun metodoaren emaitzak ebaluatzeko, polaritatea neurtzeko QWN-PPV metodoarekin eta urre-patroiarekin konparatuko dugu:

i) Urre-patroia. 10 etiketatzaileren arteko batezbestekoa da urre-patroitzat hartu duguna.
ii) QWN-PPV metodo automatikoa.

Gure metodoa 7 Taulan beste metodo batekin eta urre-patroiarekin konparatu dugu.[10] Urre-patroiko emaitzak metodoek eman beharko lituzketen emaitzak dira eta aldeak urre-patroiko emaitzen eta bi metodoek ematen dituzten emaitzen arteko aldea adierazten du.

7. Taula

Urre-patroia, QWN-PPV metodoa eta diskurtso-egituran oinarrituriko metodoarekin lorturiko balorazioak eta bi metodoen emaitzen eta urre-patroiaren arteko aldea

Testua	Urre-patroia	QWN-PPV	Aldea	Diskurtsoa	Aldea
AIZ01	9,80	10,00	0,20	7,87	1,93
AIZ02	9,02	6,43	2,59	6,88	2,14
AIZ03	4,64	0,71	3,93	5,01	0,37
ARG01	7,60	7,86	0,26	7,38	0,22
ARG02	8,40	3,57	4,83	6,38	2,02
ARG03	4,40	4,29	0,11	3,42	0,98
ARG04	5,20	2,26	2,94	5,90	0,70
ARG05	**8,80**	6,43	2,37	**8,80**	**0,00**
ARG06	5,30	3,57	1,73	5,80	0,50
BER05	3,20	4,28	1,08	1,50	1,70
COR01	5,80	5,00	0,80	4,99	0,81
COR02	8,20	7,86	0,34	6,17	2,03
EIB01	8,80	9,28	0,48	10,00	1,20
FAR01	9,80	5,71	4,09	7,68	2,12
HIR01	8,20	4,28	3,92	7,27	0,93
IRU01	8,60	5,71	2,89	3,95	4,64
PUT01	3,60	0,00	3,60	0,00	3,60
PUT02	4,00	2,26	1,74	3,01	0,99
QUE01	7,80	6,43	1,37	6,41	1,39
Aldea guztira		39,27		28,27	

[10] Hiru metodoak konparagarri egiteko eskala berean jarri ditugu: 1etik 10era. Beraz, 5etik beherako polaritatea duten testuak balorazio negatibotzat hartuko ditugu eta 5 eta 6 artekoak neutroak eta 6etik gorakoak balorazio positibotzat.

Urre-patroiko, QWN-PPV metodoaren eta diskurtsoan oinarrituriko balorazioak aztertzen baditugu, diskurtsoan oinarritutako metodoak QWN-PPV metodoak baino emaitza hobeak ematen dituela ohartuko gara. QWN-PPV metodoan urre-patroiko balorazioen eta metodoaren balorazien arteko aldeen batura 39,27koa da; diskurtsoan oinarritutakoan 28,27koa den bitartean. Guk proposatzen dugun metodoan egon da urre-patroiko balorazioarekiko alderik txikiena (0,00; ARG05), hau da, kasu batean emaitza ondo asmatu du. Alderik handiena, berriz, QWN-PPV metodoan egon da (4,83; ARG02).

Baina gure metodoak ere ez ditu ondo bereizten balorazio negatiboko eta positiboko testuak, QWN-PPV metodoak bezala. IRU01 testuak balorazio positiboa du (8,20) eta gure metodoak balorazio negatiboa (3,95) jarri dio. Gauza bera gertatzen da COR01 testuan. Testuak balorazio positiboa du (5,80) eta gure metodoak ere balorazio negatiboa (4,99) ematen dio. Alderantzizko kasua ere gertatzen da. AIZ03 testuak balorazio negatiboa du (4,64) eta diskurtsoan oinarritutako metodoak balorazio positiboa jartzen dio (5,01).

Beraz, emaitzak aztertuta diskurtso-egituran oinarrituriko metodoak analisi fidagarriagoa egiten du QWN-PPV metodoak baino. 19 testuetatik diskurtsoan oinarritutako metodoaren emaitzak 10 testutan egon dira gertuago urre-patroitik eta 8 testutan QWN-PPV metodoan. Testu batean (PUT01), berriz, bi metodoen eta urre-patroiaren arteko alde berdina da. Gainera, urre-patroiarekiko metodoen emaitzetan egon den aldearen batura txikiagoa da diskurtsoan oinarritutakoan (28,27) QWN-PPV metodoan baino (39,27).

Jarraian, zerrendatuta ageri dira hobetzeko dauden alderdiak.

a) Aditzei polaritatea jartzea, emaitzak hobetzeko. Esaterako, AIZ02 testuan, EBALUAZIO gorena erlazio erretorikoan, 'lagundu' aditzari polaritate positiborik ez zaio jarri. Beste zenbait testutan ere antzeko fenomenoa gertatu da: 'goxatu' (QUE01), 'erdietsi' (IRU01) eta 'gozatu' (FAR01).
b) Entitateak bereizteko teknika inplementatzea. Entitateek ez dute polaritaterik, baina metodoaren lehen proposamena maila lexikoko azaleko analisian oinarritu denez, token bakoitza hartu da kontuan eta horrek zenbait akats dakartza. Adibidez, QUE01 testuan, 'Gerra txikia' entitate bat da baina 'gerra' zein 'txikia' polaritate negatiboko hitz gisa etiketatu dira.
c) Testuinguruaren araberako polaritatea ezartzea elementuei. Hitz batek polaritate positiboa edo negatiboa izan dezake testuinguruaren arabera. Esaterako, AIZ02 testuan 'aberats' adjektiboak zentzu positiboa du; egindako literatura lana landua eta bertan baliabide asko erabili direla adierazteko, baina beste testuinguru batean; pertsona aberatsek pobreak baliatzen dituztenean eurak aberasteko eta 'aberats horiek diruagatik dena egiten dute' moduko adierazpenak egiten

'aberats' unitate lexikalak zentzu negatiboa hartzen du. EIB01 testuan ere, 'arin' hitza erabili da idazlearen gaitasuna deskribatzeko eta, beraz, zentzu positiboa du; baina 'buru arina' bezalako hitz-elkarketa edo antzeko sintagma batean 'arin' hitzak polaritate negatiboa leukake.

d) Diskurtso-mailan, berriz, interesgarria litzateke diskurtso-unitate eta erlazio erretorikoei pisu gehiago jartzea; darabilgun metodoak intentsifikatzaileek nahiz diskurtso-unitate edota erlazio erretorikoek antzeko pisua ematen diete hitzei (gehienez, hitzak duen pisua beste). Izan ere, diskurtso-egituraren eragina intentsifikatzaileen gainetik dago. Adibidez, Taboada et al.ek (2008) nukleoei 1,5ko pisua jartzen die eta sateliteei 0,5ekoa.

Bestalde, metodoa garatzerakoan zenbait hitz-zerrenda eratu ditugu: adjektiboena, izenena, adberbioena, intentsifikatzaileena eta ezezkoena. Hitz-zerrenda horiek unitate lexikalak eta bere polaritateak (+: positiboa, +−: neutroa edo bi polaritateduna eta −: negatiboa) osatzen dute. Honako ezaugarriak dituzte lortu ditugun hitz-zerrendek.

8. Taula

Hitz-zerrenden ezaugarriak

Zerrenda-mota	Unitate kopurua	Ehunekoa	+	+−	−
Adjektiboak	163	%42,01	72	8	83
Izenak	153	%39,43	64	6	83
Aditzondoak	49	%12,63	23	2	24
Intentsifikatzaileak	14	%3,61			
Ezezkoak	9	%2,32			
Guztira	388	%100	159 (%43,56)	16 (%4,39)	190 (%52,05)

5. Ondorioak eta etorkizunerako norabidea

Lan honetan, testuei polaritatea jartzeko maila morfosintaktikoa eta diskurtsoan oinarritzen den metodoa eskuz garatu dugu. Lexiko mailan adjektiboak, izenak eta adberbioak tratatu ditugu eta baita berauen polaritatea alda dezaketen intentsifikatzaileak eta ezezkoak ere. Diskurtsoa lantzeko eta testuko zatiei pisu ezberdina emateko Egitura Erretorikoaren Teoria (RST) baliatu dugu.

Lan honetako ondorio nagusiak honako hauek dira:

— Diskurtso-egitura erabiliz emaitza hobeak lortzen dira. Testu baten balorazioa asmatu egin da eta asmatzea lortu ez den kasuetan; 10 alditan hurbilago egon asmatzetik dikurtsoan oinarritutakoa QWN-PPV meto-

doa baino (8 aldiz egon da gertuago QWN-PPV eta kasu batean, bi metodoek hurbiltasun bera izan dute).

— Testuak balorazioa negatiboa duen kasuetan; bostetik behin egiten du huts erlaziozko diskurtso-egituraren informazioa darabilen metodoak eta inoiz ez QWN-PPV metodoak. Balorazio neutroa duten testuetan, hiru kasuetatik bitan asmatzen du diskurtsoan oinarritutakoan eta behin maila lexikoan oinarritutakoak. Balorazio positiboa duten hamaika testuetatik hamarretan asmatzen du guk proposaturiko metodoak eta zazpitan QWN-PPVk. Beraz, QWN-PPV metodoa polaritate negatiboko testuetan hobeto dabilen bitartean; diskurtsoan oinarritutakoa hobeto dabil polaritate positiboko testuetan.
— Izenak eta adjektiboak dira kopuruz polaritate gehiena duten kategoria lexikoak. Kopuruz gutxiago da aditzondoa eta gutxien intentsifikatzailea eta ezezkoa dira. Hau bat dator arloan egin diren ikerketekin; adjektiboak izan baitira gehien landu direnak.
— Sortutako lexikoian, polaritate negatiboa duten hitzak gehiago dira. Zerbait gutxiago dira polaritate positibodun hitzak eta askoz gutxiago neutroak edo bi polaritate dituztenak.

Bestalde, hauexek dira lehen proposamen honek izan dituen mugak.

— Entitateak antzematea. Beharrezkoa da entitateak antzemateko teknikak metodoan inplementatzea, emaitzak hobetu ahal izateko. Izan ere, entitateak ziren zenbait hitz, HAULak (Hitz Anitzeko Unitate Lexikalak) batez ere, polaritatea duten hitz moduan tratatu ditu metodoak eta horrek metodoaren emaitzak okertu egin ditu.
— Baliagarria den informazio gehiago gehitzea. Ikerlanen honen muina erlaziozko diskurtso-egituraren informazioa sentimenduen analisian erabiltzea izan da. Hala ere, emaitzak oraindik hobetu daitezke eta honek metodoan beharrezkoa den baina oraindik identifikatu ez dugun informazioa ez dugula aplikatu iradokitzen digu. Honen atzean arrazoi hauetako bat edo denen konbinaketa egon daitekeela uste dugu: diskurtso-egiturari pisu gutxi jarri izana, maila sintaktikoa landu ez izana, aditzak kontuan hartu ez izana, besteak beste.
— Lexikoari polaritatea jartzeko irizpidea gehiago finkatu beharra. Polaritatea jartzerakoan irizpide jakin bat jarraitu badugu ere, hau da, testuingurua eta lexikoen arteko antzekotasuna, berau ez da nahikoa. Azterketa linguistiko sakona beharrezkoa da polaritate-eskala kategoria lexiko hauei jartzerakoan.
— Metodoek ematen dituzten emaitzak eskala batean jartzerakoan ere mugatuta egon gara. Alde batetik, erabilitako corpusa oso txikia da sistema bat ondo ezagutzeko, hau da, bere maximo eta minimoak zein diren jakiteko. Bestalde, Pang eta Leek (2005) dioten moduan, eskala aldatzean beti egongo da konparaezintasuna; eta horregatik, erreferentzia gisa erabili ahal izango dira soilik.

Etorkizunean lan hauek egiteko asmoa dugu:

— Literatura-kritiketako testuetan, EDUen eta koherentzia erlazioen sailkapena egin, subjektibo-objektibo ardatzaren arabera.
— SentiWordNet euskarara itzuli eta eskuzko errebisioa egin. Aditzak SentiWordNet erabiliz inplementatuko ditugu metodoan.
— Lan honetan garatu den metodoa automatizatu.
— Taboada et al.ek (2008) egin duten moduan, diskurtso-egiturako beste faktore batzuei pisua jartzeko metodo ezberdinak aztertu, emaitzak hobetzen diren ikusteko.
— Metodoak ematen dituen emaitza eskalan kokatzeko modu ezberdinak aztertu eta fidagarriena aukeratu.

Erreferentzia bibliografikoak

ALKORTA, J., GOJENOLA, K., IRUSKIETA, M., eta PEREZ, A. (2015), «Using relational discourse structure information in Basque sentiment analysis». In *5th Workshop «RST and Discourse Studies», in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante, España.

ASHER, N., BENAMARA, F., eta MATHIEU, Y. Y. (2009), «Appraisal of opinion expressions in discourse». *Lingvisticæ Investigationes*, 32(2):279-292.

BACCIANELLA, S., ESULI, A., eta SEBASTIANI, F. (2010), «Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining». In *LREC*, volume 10, 2200-2204 orr.

CARLSON, L., OKUROWSKI, M. E., MARCU, D., CONSORTIUM, L. D., et al. (2002), *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

CHESLEY, P., VINCENT, B., XU, L., eta SRIHARI, R. K. (2006), «Using verbs and adjectives to automatically classify blog sentiment». *Training*, 580(263):233.

DA CUNHA, I., TORRES-MORENO, J.-M., eta SIERRA, G. (2011), «On the development of the rst spanish treebank». In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10. Association for Computational Linguistics.

EGAÑA, I. (2013), *Kritikarako hurbilketa literaturaren soziologiatik. Egunkari eta aldizkarietako euskal literatur kritiken analisia (1975-2005)*. Doktoretza tesia. Euskal Herriko Unibertsitatea, UPV/EHU, Vitoria-Gasteiz.

EZEIZA, N., ADURIZ, I., ALEGRIA, I., ARRIOLA, J. M., eta URIZAR, R. (1998), «Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *COLING-ACL'98*, volume 1, 380-384 orr., Canada.

GÓMEZ, I. (1997), «La partición informacional en el discurso». Doktoretza tesia. Euskal Herriko Unibertsitatea, UPV/EHU.

GÓMEZ, I. (2002), «Foco y tema. una aproximación discursiva». EHUko Argitalpen Zerbitzuak/Servicio Editorial de la UPV.

HEERSCHOP, B., GOOSSEN, F., HOGENBOOM, A., FRASINCAR, F., KAYMAK, U., eta de JONG, F. (2011), «Polarity analysis of texts using discourse structure». In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1061-1070 orr. ACM.

HORVATH, B. M. eta EGGINS, S. (1995), «Opinion texts in conversation». *Advances In Discourse Processes*, 50:29-46.

IRUSKIETA, M. (2014), *Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean*. Doktoretza tesia. Euskal Herriko Unibertsitatea, UPV/EHU, Donostia.

IRUSKIETA, M., ARANZABE, M. J., DE ILARRAZA, A. D., GONZALEZ, I., LERSUNDI, M., eta DE LA CALLE, O. L. (2013), «The RST basque treebank: an online search interface to check rhetorical relations». In *4th Workshop RST and Discourse Studies, Brasil, October*, 21-23 orr.

LIU, B. (2012), «Sentiment analysis and opinion mining». *Synthesis Lectures on Human Language Technologies*, 5(1):1-167.

MANN, W. C. eta TABOADA, M. (2005), RST web site. Available at http://www.sfu.ca/rst/.

MANN, W. C. eta THOMPSON, S. A. (1987), *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

MANN, W. C. eta THOMPSON, S. A. (1988), «Rhetorical structure theory: Toward a functional theory of text organization». *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243-281.

PANG, B. eta LEE, L. (2004), «A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts». In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271-278 orr. Association for Computational Linguistics.

PANG, B. eta LEE, L. (2005), «Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales». In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 115-124 orr. Association for Computational Linguistics.

PANG, B. eta LEE, L. (2008), «Opinion mining and sentiment analysis». *Foundations and trends in information retrieval*, 2(1-2):1-135.

PARDO, T. A. S. eta SENO, E. R. M. (2005), «Rhetalho: um corpus de referência anotado retoricamente». *Anais do V Encontro de Corpora*, 24-25 orr.

POLANYI, L. eta ZAENEN, A. (2006), «Contextual valence shifters». In *Computing attitude and affect in text: Theory and applications*, 1-10 orr. Springer.

SAN VICENTE, I., AGERRI, R., eta RIGAU, G. (2014), «Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages». In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden.

SAURÍ, R. (2008), *A factuality profiler for eventualities in text*. Doktoretza tesia. Brandeis University Waltham.Massachusetts.

STEDE, M. (2004), «The potsdam commentary corpus». In *Proceedings of the 2004 ACL Workshop onDiscourse Annotation*, 96-102 orr. Association for Computational Linguistics.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K., eta STEDE, M. (2011), «Lexicon-based methods for sentiment analysis». *Computational linguistics*, 37(2):267-307.

TABOADA, M., VOLL, K., eta BROOKE, J. (2008), «Extracting sentiment as a function of discourse structure and topicality». *Simon Fraser Univeristy School of Computing Science Technical Report*.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., eta WELPE, I. M. (2010), «Predicting elections with twitter: What 140 characters reveal about political sentiment». *ICWSM*, 10:178-185.

XU, K., LIAO, S. S., LI, J., eta SONG, Y. (2011), «Mining comparative opinions from customer reviews for competitive intelligence». *Decision support systems*, 50(4):743-754.