EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)

Hizkuntzaren eta Literaturaren Didaktika Saila

Doktorego-tesia

Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa

Jon Alkorta Agirrezabala

Donostia, 2019

EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)

Hizkuntzaren eta Literaturaren Didaktika

Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa

Jon Alkorta Agirrezabalak Koldo Gojenola Galletebeitia eta Mikel Iruskieta Quintianen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Aitari, Amari eta Josuri

Eskerrak

Lehenik eta behin, nire zuzendariak eskertu nahiko nituzke: Koldo eta Mikel, eman didaten laguntzagatik. Eskerrik asko beti nire atzean egoteagatik eta baita tesian zehar sortu diren zalantzetan laguntza eskaintzeagatik ere. Azkenik, eskerrik asko egin dizkidazuen iruzkinengatik (asko lagundu baitidate) eta ikerkuntzaren mundua nolakoa den erakusteagatik ere. Halaber, eskerrak Maxuxi eta Rodriri tesi-txostena hobetzen laguntzeagatik.

Eskerrik asko tesian lagundu didaten beste pertsonei ere. Eskerrik asko Arantxari *Eustagger*rekin izan ditudan une zailetan laguntzeagatik eta Intxari *include*k Latexen duen garrantzia erakusteagatik. Eskerrak Estherri Kanadatik Ixarako konexioa ahalbidetzeagatik, Kikeri emandako laguntzagatik eta Amaiari administrazioaren munduan laguntzeagatik. Eskerrik asko, halaber, Itziar Gonzalez-Diosi *Murriztapen Gramatika*rekin eta tesi-txostena vs. Latex borrokan laguntzeagatik. Eskerrik asko zerrenda honetan aipatu gabe gelditu diren baina lagundu didaten beste guztiei ere.

Eskertzekoa da tesia hasi eta amaitu arte inguruan mugitu denari ere: 318 bulegoari giro ona sortzeagatik, tupper txokoari askotariko gaiak (batzuetan, frikiak) aipatzeagatik, pintxo-poteetan eta egindako beste planetan atera zareten guztioi eta, azkenik, estatistika klaseetan, Oierri izandako pazientziagatik eta *nukleo duro*ri estatistika ikasteko laguntza morala emateagatik.

También tengo que agradecer a Maite Taboada por su ayuda en la estancia de Vancouver y en los trabajos relacionados con el discurso y análisis de sentimientos.

Familia ere ezin da aipatu gabe utzi. Eskerrik asko aitari, amari eta Josuri

beti hor egoteagatik, eta laguntza behar izan dudanean laguntzeagatik.

Eskerrik asko kuadrillari ere, asteburuetan, musa eta *Comunio*rekin une biziki ziragarriak emateagatik. Eskerrak baita Urkizu eta Gazteluko bazkariengatik eta Kanpezura eta Jakatik mendira egindako planengatik ere.

Bukatzeko eskerrik asko Amets eta Lierni pisukideei eta aipatzea ahaztu zaizkidan beste guztiei.

Mila-mila esker denoi!

Laburpena

Tesi-lan honetan, hizkuntzalaritza aplikatuaren ikuspegitik, euskarazko sentimenduen analisian lehen urratsak egin dira. Bi helburu nagusi egon dira tesi-lanean. Alde batetik, sentimenduen analisia egiteko oinarrizko baliabideak sortu ditugu euskararentzat. Zehatz esanda, Euskarazko Iritzi Corpusa, *Sentitegi* izeneko euskarazko sentimenduen lexikoia eta dokumentu mailako sentimenduen sailkatzailea garatu ditugu. Corpusak sei domeinutako 240 iritzi-testu biltzen ditu. Egitura Erretorikoaren Teoriaz (RST) baliatuta, corpuseko diskurtso-informazioa etiketatuta dago. Gainera, iritzi-testuen orientazio semantikoa ere etiketatuta dago. Sentimenduen lexikoiari dagokionez, 1.237 hitzez osatuta dago eta bertako sarrerek -5 eta +5 arteko sentimendu-balentzia dute. Sentimenduen lexikoia sortzeko itzulpen-metodologia zehatz bat jarraitu dugu. Azkenik, dokumentu mailako sentimenduen sailkatzailea ere garatu dugu. Tresnaren oinarrian aurretik aipatu dugun sentimenduen lexikoia dago.

Beste aldetik, sentimenduen analisiaren lanketa teoriko bat ere egin dugu. Sentimenduen sailkapena lexikoian oinarrituz egin nahi bada, hitzen sentimenduen balentzia jakitearekin ez da nahikoa, izan ere, testuetan badaude zenbait fenomeno hitz horien sentimenduen balentzia eragiten dutenak. Horiei testuinguruko balentzia-aldatzaileak deritze eta horiek euskaran nola agertzen diren landu dugu. Hizkuntza maila bakoitzeko balentzia-aldatzaile mota bat landu dugu: fonologian, bustidura adierazkorra; morfologian, hizkiak; sintaxian, ezeztapen-markak eta, azkenik, diskurtsoan, erlaziozko diskurtso-egiturak eta unitate zentrala. Emaitzek erakusten dutenez, balentzia-aldatzaileek hitzen edo sintagmen sentimenduen balentzia indartu

edo ahuldu egiten dute. Ahultze horren intentsitatearen arabera, sentimendu balentziaren zeinuan aldaketa gerta liteke, positiboa dena negatibo bilakatuz edo alderantziz. Azkenik, kasu batzuetan, balentzia-aldatzaileak ez du eraginik sortzen.

Aurkibidea

METODOLOGIA

ONDORIOAK

Eranskina

Irudien zerrenda

Taulen zerrenda

SARRERA

1. KAPITULUA

Proiektuaren nondik norakoak

1.1 Motibazioa

Iritziak giza jardueraren oinarrian daude eta gure portaeran eragin handia dute (Liu, 2012). Errealitatearen gure uste eta pertzepzioak nahiz egiten ditugun aukeraketak hein handi batean munduak ikusi eta ebaluatzen duenaz baldintzatuta daude. Egoera horren ondorioz, erabaki bat hartu behar dugunean, besteen iritziak bilatu ohi ditugu. Norbanakoetan eta erakundeetan gertatzen den egoera da.

Iritziaren inguruan ere badira zenbait kontzeptu subjektibitatearekin lotura dutenak. Besteen artean, sentimendua, ebaluazioa, jarrerak eta emozioak aipa daitezke eta horiek guztiak, iritziaz gain, sentimenduen analisia edo iritzi-meatzaritza deritzon arloaren ikergaiak dira.

Liuk (2012) gogorarazten duenez, arlo horren hasiera eta hazkundea erabat loturik daude webetako sare sozialen jaiotzarekin. Webetako sare sozialak era askotakoak dira: iruzkinak, foroetako eztabaidak, blogak, mikroblogak eta Twitter. Aplikazio horiek guztiek gizakiaren historian lehen aldiz digitalki jasotako iritzi bolumen handi bat edukitzea ahalbidetu dute. Hori horrela izanik, 2000tik aurrera hizkuntza naturalaren prozesamenduan hazkunderik handiena eta aktiboena bilakatu den arloa da. Gainera, datu-meatzaritzan, web-meatzaritzan eta testu-meatzaritzan ere erabiltzen da.

Oro har, enpresan eta gizartean garrantzizkoa bilakatu da sentimendu-analisia, eta horrek arloa konputazio-zientzietatik kudeaketa-zientzietara eta gizarte-

zientzietara zabaldu du. Bere inguruan, industria garatu da eta enpresek zerbitzuak ere sortu dituzte arloa lantzeko. Beraz, sentimendu-analisiak enpresa edo gizarte eremu askotan aplikazioa aurkitu duela esan daiteke.

Ikusi dugun moduan, nazioartean sentimenduen analisiak garapen handia izan du. Egun, munduko *lingua franca* ingelesa da eta sentimendu-analisiko lan gehienak hizkuntza horretan egin dira. Gurera etorrita, ordea, gutxi dira euskara hartuta egin diren lanak.

Sentimendu-analisia oso arlo zabala da. Ataza batzuek lotura handiagoa dute informatikarekin eta beste batzuek, berriz, hizkuntzalaritzarekin. Gainera, ataza batzuetan bi hurbilpen egoten dira: hizkuntzaren ezagutzan oinarritutakoa eta metodo estatistikoetan oinarritutakoa.

Hizkuntzalaritzaren ikuspegitik, lehen urratsa da zein hitzek duen iritzia edo, modu zabalago batean, informazio subjektiboa den jakitea. Hurrengo urratsean, ordea, hitz horien informazio subjektiboan eragiten duten hizkuntzari lotutako fenomenoak bilatu nahi dira. Azter ditzagun hurrengo adibideak.

(1) Pelikula hori [gustatu$_+$]$_+$ zait.

(2) Pelikula hori [asko gustatu$_+$]$_{++}$ zait.

(3) Pelikula hori ez zait [gustatu$_+$]$_-$.

(4) Pelikula hori [gustatuko$_+$ litzaidake]$_0$ kontakizuna biziagoa izango balitz.

Goian, pelikula bati buruzko zenbait iritzi agertzen dira. Denak antzekoak dira, baina ez berdinak. (1) adibidean, esaterako, pelikula gustukoa izan duela esaten da; beraz, esaldiaren balorazioa positiboa da. (2) adibidean ere esaldiaren balorazioa positiboa da, baina *asko* intentsifikatzailea[1] agertzen da eta, horregatik, esaldi honek aurreko esaldiak baino iritzi positiboagoa du. (3) adibidean, ordea, esaldian iritzia adierazten duen hitz bat, positiboa gainera, egon arren, esaldiaren balorazioa negatiboa da. Azkenik, (4) adibidean ere, iritzia duen hitz bakarrak balorazio positiboa egiten du, baina esaldiak

[1]Intentsifikatzaileak adberbioak edo adjektiboak izan ohi dira. Berezko eduki semantiko gutxi dute, baina aldatzen duen hitzaren edo esaldiaren esanahia intentsifikatzen du.

ez du iritzirik adierazten, erreala ez den egoera batean dagoelako iritzia[2].

Adibide horietan ikusten denez, esaldi guztietan iritzia adierazten duen hitz bat (*gustatu*) egon arren, esaldi osoek adierazten duten iritzia ezberdina da. Tartean, hitzen arteko interakzioak egon dira eta fenomeno linguistiko batzuek iritzia adierazten duen hitzaren balorazioa aldatu egin dute.

Hain zuzen ere, aipatu ditugun bi alderdi horiek dira gure lanaren motibazioak. Alde batetik, euskararentzat sentimendu-analisiaren tresnak eta baliabideak sortu nahi ditugu, gaur egun, euskarazko testuen eduki subjektiboa erauzi ahal izateko. Beste aldetik, berriz, iritzia edo informazio subjektiboa duten hitzetan eragiten duten fenomeno linguistikoak bilatu nahi ditugu, fonologian, sintaxian eta diskurtsoan.

Gure lehen helburua sentimendu-analisiaren arloan ikerketa egiteko oinarrizko tresna eta baliabideak sortzea da. Zehazki, euskarazko iritzi-testuen corpusa eta euskarazko sentimendu lexikoia sortu nahi ditugu.

Izan ere, gutxi dira euskaraz dauden iritzi-testuen corpusak eta sentimendu lexikoiak, eta daudenak, berriz, ez dira baliagarriak gure helburuak betetzeko; izan ere, ez dituzte hizkuntzaren ezaugarriak aintzat hartzen eta hori ez dator bat gure ikusmoldearekin. Iritzi-testuen kasuan, txioak eta zerbitzuak (produktuen salerosketa webguneek eta hotelen web konparatzailea, adibidez) eskaintzen dituzten webguneetako iritziak biltzen dituzten corpusak edota datu-baseak badaude baina guri ez zaizkigu baliagarri testuak motzak direlako eta ondorioz, diskurtsorako zenbait elementu aztertzeko aukera zailduko ligukeelako. Lexikoiei dagokienez, berriz, hitzen orientazio semantikoa (positiboa edo negatiboa) adierazi arren, ez dute intentsitatea adierazten (hau da, *ondo* eta *bikain* positiboak dira, baina intentsitatean duten ezberdintasuna ez da adierazten) eta horregatik, ez zaizkigu baliagarriak. (Stone eta Hunt, 1963) sentimendu lexikoia, esaterako, tankera horretakoa da.

Gure bigarren helburua aurreko erronkan sortutako sentimendu lexikoiko hitzen balentzian eragiten duten hizkuntza-fenomenoak bilatzea eta beren eragina neurtzea da.

[2]Adibidean ematen den iritzia egoera errealean ez dagoela diogu, iritzia emateko erabiltzen den aditza ez dagoelako ez iraganean, ezta orainaldian ere.

(1), (2), (3) eta (4) adibideetan ikusi dugun moduan, lexikoian oinarritzen den sentimenduen sailkatzaile batean, lexikoia bere baitan bakarrik ez da baliagarria. Informazio linguistiko gehiago behar da sentimenduen sailkapen on bat egiteko. Horregatik, Polanyi eta Zaenenen (2006) lanean oinarrituz, euskaran dauden hizkuntza maila ezberdinetako testuinguruko balentzia-aldatzaileak[3] identifikatu nahi ditugu.

Gure hirugarren eta azken helburua dokumentu mailako euskarazko sentimenduen sailkatzaile bat sortzea da. Euskaran lan gutxi daude sentimenduen sailkapena egiten dutenak. Horietako bat EliXa (San Vicente *et al.*, 2015) da eta aspektu mailako sentimenduen sailkapena egiten du[4]. Gure erronka euskarari hizkuntza-prozesamendurako beste baliabide bat ematea da eta, horretarako, lexikoian oinarritutako dokumentu mailako sentimenduen sailkatzailea sortu nahi dugu. Esan behar da erronka hau betetzeko beharrezkoa dela aurretik aipatutako erronkak betetzea: sentimenduen lexikoi bat sortu behar da sentimenduen sailkatzailea inplementatzeko eta balentzia-aldatzaileak landu behar dira lexikoiko hitzetan eragiten duten aldaketa neurtzeko.

Laburbilduz, tesi honek euskara eta sentimenduen analisia uztartzen ditu. Euskara hizkuntzatzat hartzen duten lan gutxi egin dira sentimenduen analisian eta gehienak sentimendu lexikoia sortzera mugatzen dira. Guk euskarari baliabide eta tresnak eman nahi dizkiogu, baita euskarak arlo honetan dituen berezitasunak ikertu ere.

1.2 Hipotesi orokorrak

Aurreko atalean, sentimenduen analisia arloaren motibazioa aipatu dugu, baita berak gizartean duen erabilgarritasuna ere. Atal honetan, tesiaren irismena zehaztuko dugu. Lau hipotesi orokor ditu lan honek.

[3]Testuinguruko balentzia-aldatzaileak (*contextual valence shifters*, ingelesez) hitzen edota esaldien sentimendu balentzian aldaketak eragiten dituzten fenomenoak dira. Aldaketa hori balentziaren indartze bat edo ahultze bat izan daiteke.

[4]Aspektu mailako sentimenduen analisiak aztergai den entitateari loturiko aspektuak identifikatu (entitatea Londres bada, aspektuak ekonomia, turismoa etab. dira) eta horiei buruzko iritzia positiboa edo negatiboa den zehaztu nahi du.

1. ikerketa-hipotesia: *itzulpen bidez lortutako euskarazko sentimenduen lexikoiaren kalitatea corpusetik edo datu-base lexikaletatik sor daitezkeenak bezain baliagarria da.*
Sentimenduen lexikoiek testuetako hitzei beren orientazio semantikoa[5] eta sentimenduen balentzia esleitzen diete eta hori testu baten orientazio semantikoa kalkulatzeko lehen urratsa da. Lexikoiak sortzeko, ohiko bi hurbilpen daude: i) corpus bidez eta ii) datu-base lexikalen bidez.

Guk hautatutako bidea aipaturiko hurbilpenen nahasketa da. Oinarri bezala bi datu-base lexikal, hots, gaztelaniazko eta ingelesezko bi sentimenduen lexikoi erabiliko ditugu eta horiek itzuli egingo ditugu Euskarazko Iritzi Corpusak ematen duen informazioa aberastuz.

Gure ustez, jada sorturik dagoen sentimenduen lexikoia euskaratzea aurreko bi hurbilpenak (corpus bidezkoa eta datu-base lexikal bidezkoa) bezain baliagarria da eta kalitatean ez legoke ezberdintasun handirik. Izan ere, itzulpenaren metodologia aproposa bada, bidean ez da informaziorik galduko edo eraldatuko eta hasierako kalitatea mantenduko luke. Gainera, iritzi-testuen corpusa lagun izanda, sentimendua duten hitzen orientazio semantikoan nahiz sentimenduen balentzian okerreko esleipenak egitea eragotziko litzateke.

2. ikerketa-hipotesia: *hizkuntza maila guztietan (fonologia, sintaxia eta diskurtsoa) zenbait fenomeno linguistiko daude orientazio semantikoa duten hitzetan (eta esaldietan) eragiten dutenak.*
Eragin hori beren sentimenduen balentzia indartzea edo ahultzea izan daiteke. Batzuetan fenomeno linguistikoak ez du balentzian aldaketarik eragiten.

Gure ustez, fenomeno linguistiko horiek gramatika maila ezberdinetan ager daitezke: esateko moduan (hau da, fonologian), hizkien bidez, sintaxiko fenomeno ezberdinen bidez edota baita diskurtsoaren bidez ere. Lexikoian oinarritzen den dokumentu mailako sentimenduen sailkatzaile bati mota honetako informazioa gehitzea beharrezkoa dela uste dugu, bestela sailkatzailearen emaitzak ez baitira nahi bezain zehatzak izango.

[5]Orientazio semantikoa hitz batek duen informazio subjektiboa da eta positiboa edo negatiboa izan daiteke.

3. ikerketa-hipotesia: *diskurtso-egituran badaude osagaiak EDUen*[6]
Gure ustez, EDU edo erlaziozko diskurtso-egiturak orientazio semantiko positiboa edo negatiboa edukitzea ez da ausazko gertaera bat. Hau da, elementu horiek RST-zuhaitzean[7] dituzten guneak horretan eragina dutela uste dugu. Zehazki esanda, diskurtso-egiturako zenbait faktorek edo osagaik EDU eta erlaziozko diskurtso-egiturako orientazio semantikoa baldintzatzen dutela da gure hipotesia.

4. ikerketa-hipotesia: *testu baten domeinuak unitate zentralak ezaugarri jakin batzuk izatea eragiten du, bai bertan agertzen diren hitzen gramatika-kategoriari dagokionez, bai hitz horiek orientazio semantikoa edukitzea edo ez edukitzearekin ere.*
Nahiz eta unitate zentrala testu guztietan dagoen, bere ezaugarriak domeinuaren arabera ezberdinak direla uste dugu. Eta horrek unitate zentraletan orientazio semantikodun hitzak agertzeko moduan eragiten duela ere pentsatzen dugu. Testuen sentimenduen analisia egiterakoan, beharrezkoa domeinua zein den jakitea, horrek arrasto batzuk ematen baititu sentimendu balentziadun hitzak non agertzen diren jakiteko.

1.3 Helburuak

Aurreko ataletan ikerketa-lan honen testuingurua eta gure ikerketa gidatzeko hipotesi orokorrak eztabaidatu ditugu. Atal honetan, berriz, gure helburu zehatzak adieraziko ditugu gure hipotesiak sostengatzen dituzten egiaztatzeko.

- **1. helburua:** Sentimenduen analisia lantzeko oinarrizko tresnak eta baliabideak sortzea.

[6]Oinarrizko Diskurtso Unitatea (*Elementary Discourse Unit*, EDU) *diskurtso-egiturako unitaterik txikiena da.* eta erlaziozko diskurtso-egituren orientazio semantikoa duten bezalakoa izatea eragiten dutenak.

[7]Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory*, RST) testu-sorkuntza automatikorako sortutako teoria da. RSTn testuaren antolakuntza testu-zatien artean daude erlaziozko diskurtso-egitura ezberdinen bidez garatzen da eta koherentzia testuaren egiturak duen hierarkiaren bidez gauzatzen da.

  - ***1.1 helburua:*** euskarazko iritzi-testuak biltzen dituen corpusa sortzea, testuen orientazio semantikoari dagokionez orekatua eta sintaxiari dagokionez aberatsa.
  - ***1.2 helburua:*** euskarazko sentimenduen lexikoia sortzea, hitzen orientazio semantikoa adierazten duena eta horretarako hitzei zenbakizko balio bat esleitzen diena.
  - ***1.3 helburua:*** dokumentu mailako euskarazko sentimenduen sailkatzaile bat sortzea, oinarritzat sentimenduen lexikoia izango duena.

- **2. helburua:** hitzen sentimenduen balentzian eragiten duten euskarazko testuinguruko balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.1 helburua:*** fonologiako balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.2 helburua:*** morfologiako balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.3 helburua:*** sintaxian, ezeztapen-markek hitzaren edo esaldiaren sentimendu balentzian nola eragiten duten neurtzea.
  - ***2.4 helburua:*** diskurtsoan, erlaziozko diskurtso-egituretan nukleartasunaren eta iritzi-testuetan unitate zentralaren eragina neurtzea orientazio semantikoari dagokionez.
  - ***2.5 helburua:*** iritzi-testuetan, erlaziozko diskurtso-egiturek testuaren posizio jakin batean agertzeko joera duten aztertzea.
  - ***2.6 helburua:*** iritzi-testuetako unitate zentralen azterketa: hitzen gramatika-kategoriaren azterketa eta orientazio semantikoa duten hitzen banaketaren lanketa.

1.4 Argitalpenak

Tesi-lan honetan zehar garatutako zenbait lan aldizkari eta kongresu ezberdinetan aurkeztu dira. Atal honetan, argitalpen horiek zerrendatu ditugu tesi-laneko gaiaren arabera sailkatuta.

- **Corpusa**
  - Alkorta J., Gojenola K. eta Iruskieta M. Creating and evaluating a polarity-balanced corpus for basque sentiment analysis. In *Fourth International Workshop on Discourse Analysis*, 54-58, Santiago de Compostela, Spain, 2016a. ISBN 978-84-608-9305-9.
- **Sentimendu-hitzak eta sentimendu-lexikoia**
  - Alkorta J., Gojenola K. eta Iruskieta M. SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas* 22(4): 1295-1306, 2018b, ISSN 2007-9737, DOI 10.13053/CyS-22-4-3075
  - Alkorta J., Gojenola K. eta Iruskieta M. Sentimenduak deskribatzeko hurbilpen teoriko konputazionala euskaraz. *UPV/EHUko I. Doktorego Jardunaldiak. Elkarrekin ikertuz*: 245-246, 2017a, Euskal Herriko Unibertsitateko Argitalpen Zerbitzua, ISBN 978-84-9082-619-5.
  - Alkorta J., Gojenola K., Iruskieta M. eta Taboada M. Using lexical level information in discourse structures for Basque sentiment analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, 39-47, Santiago de Compostela, Spain, 2017b. ISBN 978-1-945626-78-4.
- **Ezeztapena**
  - Alkorta J., Gojenola K. eta Iruskieta M. Saying no but meaning yes: negation and sentiment analysis in Basque. In Balahur A., Mohammad S. M., Hoste V., eta Klinger, R., editoreak, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 85-90, Brussels, Belgium, 2018a. The Association for Computational Linguistic. ISBN 978-1-948087-80-3.
- **Diskurtsoa**

  - Alkorta J., Gojenola K., Iruskieta M. eta Perez A. El uso de la información de la estructura retórica en el análisis de sentimiento. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., eta Romá Ferri M.T., editoreak, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015b. Sociedad Española para el Procesamiento del Lenguaje Natural. ISBN 978-84-608-1989-9.
  - Alkorta J., Gojenola K. eta Iruskieta M. Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena. *Gogoa: Euskal Herriko Unibersitateko hizkuntza, ezagutza, komunikazio eta ekintzari buruzko aldizkaria (Xabier Arrazola Gogoan (1962-2015))*:14, 131-152, 2016b, ISSN 1577-9424
  - Alkorta J. El uso de información del discurso en el análisis de sentimientos en euskera. In Almela A., Alcaraz-Mármol G., Valencia R., Fernández G. T. *Proceedings of Doctoral Symposium of the 33rd Conference of the Spanish Society for Natural Language Processing (SEPLN 2017)*, Murcia, Spain, 2017c. ISSN 1613-0073
  - Alkorta J., Gojenola K. eta Iruskieta M. 2019b. Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, 144-152, Association for Computational Linguistics. ISBN 978-1-948087-98-8

- **Testuinguruko balentzia-aldatzaileak oro har**

  - Alkorta J., Gojenola K. eta Iruskieta M. 2019a. Sentimenduen tratamendu konputazionalerantz: gramatika maila ezberdineta-ko sentimendu balentzia-aldatzaileen bila. Olatz Arbelaitz, Urtzi Etxeberria, Ainhoa Latatu, Miren Josu Ormaetxebarria (arg.), *III. Ikergazte. Nazioarteko Ikerketa Euskaraz, Giza Zientziak eta Artea* (1. liburukia), 39-46. Udako Euskal Unibertsitatea (UEU). Bilbo. ISBN: 978-84-8438-682-7

- **Bestelakoak**

  - Alkorta J. Euskararako HPSG gramatikaren lehen proposamena. Alegria I., Latatu A., eta Omaetxebarria M.J., editoreak, *I. Ikergazte, Nazioarteko ikerketa euskaraz*, 24-31, Bilbao, Spain, 2015a. Udako Euskal Unibertsitatea. ISBN 978-84-8438-540-0.

- **Bidalitakoak**

  - Alkorta J., Taboada M., Gojenola K. eta Iruskieta M. Bidalita. The role of hierarchical structure and coherence relations in sentiment analysis: a study in Basque. *Journal of Corpora and Discourse Studies*

1.5 Tesi-lanaren antolakuntza

Tesi hau zazpi kapituluz osatuta dago. Hurrengo lerroetan, kapitulu bakoitzaren edukia laburbilduta azalduko dugu.

1. kapitulua: proiektuaren nondik norakoak.
Lehen kapituluan, lanaren motibazioa, hipotesi orokorrak, helburuak eta tesiarekin lotutako argitalpenak azaldu ditugu.

2. kapitulua: sentimenduen analisiko baliabideak eta testuinguruko balentzia-aldatzaileak.
Bigarren kapitulua bost atalez osatuta dago. Lehenik eta behin, sentimenduen analisiaren arloan dauden atazak aipatu ditugu. Gero, ataza horiek gauzatzeko behar diren baliabideetako bi (iritzi-testuen corpusak eta sentimenduen lexikoiak) azaldu ditugu. Ondoren, testuinguruko balentzia-aldatzaileak aztergai dituzten lanen inguruan jardungo dugu. Laugarren atalean, sentimenduen analisia eta euskara uztartzen dituzten lanak aipatu ditugu. Azkenik, kapitulu guztia laburbildu dugu.

3. kapitulua: metodologiaren diseinua.
Kapitulu honetan, lehenik eta behin, euskarazko iritzi-testuen corpusa eta euskarazko sentimenduen lexikoia sortzeko urratsak azaldu ditugu. Ondoren, berriz, euskarazko balentzia-aldatzaileak identifikatzeko metodologia zehaztu dugu. Fonologian, morfologian, sintaxian eta diskurtsoan egin dena bananbanan aipatu dugu. Gero, dokumentu mailako sentimenduen sailkatzailea

den SO-CAL tresnaren ingelesezko bertsioa euskaratzeko prozesua nolakoa izan den azaldu dugu. Azkenik, kapituluaren laburpena egin dugu.

4. kapitulua: sentimenduen analisirako baliabideak.
Laugarren kapituluan, tesi-lan honetan zehaztutako metodologia jarraitzearen ondorioz garatutako baliabideen eta beren ebaluazioaren berri eman dugu. Aipatu ditugun baliabideen artean, Euskarazko Iritzi Corpusa, *Sentitegi* izeneko sentimenduen lexikoia eta dokumentu mailako euskarazko sentimenduen sailkatzailea daude. Bukatzeko, kapituluaren laburpena egin dugu.

5. kapitulua: balentzia-aldatzaileak.
Bosgarren kapitulu honetan, lorturiko emaitzen berri eman dugu. Baina aurreko kapitulukoan ez bezala, hemengo emaitzak teorikoak dira. Zehazki, maila fonologiko eta morfologikoan, sintaktikoan eta diskurtsokoan aurkitu ditugun testuinguruko balentzia-aldatzaileak azaldu ditugu eta hitzen edota sintagmen orientazio semantikoan edo sentimenduen balentzian zer-nolako eragina duten ere aipatu dugu. Kapituluarekin amaitzeko, laburpena egin dugu.

6. kapitulua: ekarpenak, mugak eta etorkizuneko lanak.
Kapitulu honek tesi-lanaren ondorioak aipatuko ditu lehenik. Hasierako helburuak bete diren aipatu dugu eta ikerketa-galderei erantzuna emango diegu. Lanaren ekarpenak ere aipatu ditugu. Azkenik, etorkizuneko lanen inguruan aritu gara.

Terminologia eta laburdurak.
Eranskin honetan, tesi-txosten honetan aipatutako terminologia eta laburdurak zerrendatu ditugu.

AURREKARIAK

2. KAPITULUA

Sentimenduen analisiko baliabideak eta testuinguruko balentzia-aldatzaileak

Aurrekarien kapitulu hau bost zatitan banatuta dago. Lehenik eta behin, 2.1 atalean, sentimenduen analisiaren ataza nagusiak zeintzuk diren esplikatzen da. Ondoren, 2.2 atalean, iritzi-testuen corpusak eta sentimenduen lexikoiak nola sortu diren azaltzen da. 2.3 atalean, berriz, testuinguruko balentzia-aldatzaileekin egin diren lanak aipatzen dira. Euskara eta sentimenduen analisia uztartzen dituzten lanak 2.4 atalean zehazten dira. Azkenik, 2.5 atalean, kapitulu osoa laburtuta dago.

Kapitulu honetan, 2.1 eta 2.2 atalak Liuren (2012) lanean eta bertako sailkapenean oinarrituz egin dira. Sentimenduen analisia eta iritzien erauzketa arloen barnean, hainbat ataza eta hurbilpen daude eta egile bakoitzak sailkapen ezberdin bat egiten du. Guk Liuren (2012) sailkapenaren alde egin dugu; izan ere, arloaren ikuspegi orokorrago bat emateaz gain, garrantzi berezia ematen dio lexikoian oinarritutako edo hizkuntzalaritzaren ikuspegitik gertuago dauden hurbilpenei. Sentimenduen analisiaren inguruan egin diren beste sailkapen batzuk honako lanetan aurki daitezke: Pang eta Lee (2008), Vinodhini eta Chandrasekaran (2012), Cambria *et al.* (2017), Mohammad (2016) edo Westerski (2007), besteak beste.

2.1 Sentimenduen analisiaren atazak

Sentimenduen analisian, jarraian azaltzen diren hiru ataza nagusi bereizi ohi dira eta ataza horietako bakoitza hizkuntza maila batekin (hitz, esaldi eta diskurtso mailekin) lotuta dago (Liu, 2012).

- Entitate eta aspektu mailako ataza: ezaugarri maila (*feature level*, ingelesez) lantzen da. Hau da, iritzia positiboa edo negatiboa den jakiteaz gain, iritziaren jomuga (entitatea eta bere aspektuak) identifikatu nahi da. Beraz, ataza hau hitz mailakoa da. Lan aipagarrien artean Wiebe *et al.*ena (1999) dago.

- Esaldi mailako ataza: subjektibitatearen sailkapena (*subjectivity classification*, ingelesez). Ataza honetan, esaldiak informazio subjektiboa edo objektiboa (hots, informazio faktuala) duen bereiztea da helburua. Ataza hau ere lantzen duen lanetako bat Wiebe *et al.*ena (1999) da.

- Dokumentu mailako ataza: dokumentu mailako sentimenduen sailkapena (*document-level sentiment classification*, ingelesez). Ataza honen helburua dokumentu baten subjektibitateak balorazio positiboa edo negatiboa duen identifikatzea da. Ataza honetan aipagarriak dira Pang *et al.*ren (2002), Turneyren (2002) eta Pang eta Leeren (2008) lanak. Sentimenduen analisiko hiru ataza nagusi hauen artean, gure lanak dokumentu mailakoarekin du lotura; izan ere, dokumentu mailako sentimenduen sailkatzailea garatzea dugu helburu.

- Sentimenduen analisiko beste atazak. Ataza hauek sentimenduen analisiaren eta HPko beste atazen arteko nahasketa dira. Atazak mota askotarikoak dira: iritzi-testuen laburpena, konparaziozko iritzien analisia, iritzi-bilaketa eta -berreskuraketa, iritzi *spam*en detekzioa eta iritzien kalitatearen neurketa.

Jarraian, lau atazak banan-banan esplikatuko ditugu.

2.1.1. Entitate eta aspektu mailako sentimenduen analisia

Sentimenduen analisiko atazetako bat entitate eta aspektu mailako sentimenduen analisia da. Ataza honetan, iritzia positiboa edo negatiboa den jakiteaz gain, iritziaren hartzailea zein den ere jakin nahi da.

Entitate bati buruzko iritzi bat ematean, entitate horren eta aspektu guztien iritziak beti ez datoz bat. Hori dela eta, beharrezkoa da entitatearen aspektu bakoitzak nolako iritzia duen jakitea. Adibidez, film bat entitate bat da eta bere aspektuak izango lirateke pertsonaiak, atrezzoa, soinu-banda eta argumentua, besteak beste. Aspektu horietako batzuk positiboak izan daitezke eta beste batzuk, berriz, negatiboak. Entitate horren eta aspektu horien orientazio semantikoa identifikatzea da ataza honen motibazioa.

Aspektuak hitz edo sintagma mailan agertzen dira. Horregatik, aspektu bakoitzaren orientazio semantikoa lortzeko, hitz edo sintagmaren orientazioa kalkulatzen da. Horretarako, beharrezkoa da iritziak erreferentzia egiten dion entitatea bere aspektuetan deskonposatzea.

Aspektuan oinarritutako sentimenduen analisian sei urrats daude. Urratsak esplikatzeko Liuren (2012) (5) adibidea erabiliko dugu.

(5) Posted by: bigJohn Date: Sept. 15, 2011. [1] I bought a Samsung camera and my friends brought a Canon camera yesterday. [2] In the past week, we both used the cameras a lot. [3] The photos from my Samy are not that great, and the battery life is short too. [4] My friend was very happy with his camera and loves its picture quality. [5] I want a camera that can take good photos. [6] I am going to return it tomorrow.

1- Entitateen erauzketa eta kategorizazioa. Entitate guztiak erauzi eta ezaugarrietan oinarrituta multzokatu egiten dira. (5) adibidean, *Samsung*, *Samy* eta *Canon* entitateak erauzten dira eta *Samsung* eta *Samy* entitateak multzokatzen dira entitate bera errepresentatzen dutelako.

2- Aspektuen erauzketa eta kategorizazioa. Aspektu guztiak erauzi eta

multzokatu egiten dira. (5) adibidean, *picture*, *photo* eta *battery life* aspektuak erauzten dira eta *picture* eta *photo* aspektuak multzokatzen dira automatikoki, kamerari dagokionez sinonimoak direlako.

3- Iritziaren jabearen erauzketa eta kategorizazioa. Iritzi-hartzaile guztiak erauzi eta beren ezaugarrietan oinarrituta, kategorizatu egiten dira. (5) adibidean, iritziaren jabea [3] esaldian *bigJohn* (blogaren egilea) da eta [4] esaldian, aldiz, iritziaren jabea *bigJohn*en laguna da.

4- Denboraren erauzketa eta kategorizazioa. Denbora-adierazpenak erauzi, formatu bakar batean jarri eta sailkatu egin behar dira. Aztertzen ari garen adibidean, erauzitako denbora-adierazpena Sept-15-2011 da.

5- Aspektuan oinarritutako sentimenduen analisia. Urrats honetan egiten da sentimenduen analisia, hots, iritziaren aspektuak orientazio positiboa edo negatiboa duen zehazten da eta iritziaren hartzaileari kalifikazio bat esleitzen zaio.

   [3] esaldian, Samsung kameraren irudi-kalitateari eta bateria-bizitzari iritzi negatiboa ematen zaio. [4] esaldian, iritzi positiboa ematen zaio bai Canon kamerari, bai haren irudi-kalitateari. Gainera, esaldi honetan, *its* eta *this camera* hitzek nori egiten dioten erreferentzia ere jakin behar da. Azkenik, [5] esaldian, iritziak positiboa dela ematen du, baina ez da horrela. Esaldi horretan, hiztunak bere desio edo nahi bat adierazten baitu eta ez objektu baten (kasu honetan, kamera baten) iritzia edo ebaluazioa.

6- Aspektu bakoitzari lotutako iritziaren sorkuntza. Azken urratsean, aspektuak, aspektuari loturiko entitatea, aspektuari buruzko iritzia, iritziaren jabea eta iritzia eman den denbora zehazten dira.

   (Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
   (Samsung, battery_life, negative, bigJohn, Sept-15-2011)
   (Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
   (Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

Hu eta Liuren (2004) lana erreferentziazkoa ataza honetan. Izan ere, bezero askoren iritzi-testuak laburtzeaz gain, laburpen horietatik balorazioren

batzuk dauden zatiak bakarrik hartzen dituzte. Horrela, iritzi datu-base bat hartu eta, lehenik eta behin, POS etiketatzea, ezaugarrien maiztasunaren identifikazioa eta ezaugarrien kimatzea burutzen dute. Ondoren, ohiko ezaugarrietatik iritzidun hitzak lortzen dituzte eta azken honetatik, ezohiko ezaugarriak ere lortzen dituzte. Gero, iritzidun hitzak eta ezohiko ezaugarriak baliatuz, iritzidun esaldiaren orientazio identifikazioa egiten dute eta azkenik, berriz, iritzi-testuaren laburpen bat lortzen dute.

2.1.2. Esaldi mailako sentimenduen analisia

Sentimenduen analisiko ataza honek esaldi mailako sentimenduen sailkapena egiten du. Ataza honek bi urrats ditu: i) subjektibitatearen sailkapena eta ii) sentimenduen sailkapena.

Ataza lantzeko azter dezagun Liuren (2012) adibidea hau:

(6) "[1] I bought a Motorola phone two weeks ago. [2] Everything was good initially. [3] The voice was clear and the battery life was long, although it is a bit bulky. [4] Then, it stopped working yesterday."

Subjektibitatearen sailkapena izeneko atazean, esaldiak iritzia adierazten duen edo ez zehazten da. Esaldiak iritzia ez du adieraziko baldin eta bertan azaltzen den informazioa faktuala edo objektiboa baldin bada (hau da, informazioa egiazkoa bada). Esaldi batean iritzia, ebaluazioa, emozioa, usteak, espekulazioa, epaia, salaketa edota jarrera, besteak beste, agertzen bada, aldiz, esaldia subjektiboa da. Hala ere, batzuetan posible da esaldi objektibo batek ere iritzia adieraztea[1].

(6) adibidea aztertzen badugu, [1] esaldiak informazio faktuala du; beraz, esaldia objektiboa da. [2] eta [3] esaldiek, berriz, iritziak adierazten dituzte *good*, *clear*, *long* eta *bulky* adjektiboak agertzen baitira. Azkenik, [4] esaldia iritzia adierazten duen esaldi objektiboa da; izan ere, bateriak funtzionatzeari uztea informazio faktuala da baina zerbait negatiboa da.

[1]Esaterako, *mugikor honek bateria azkar xahutzen du* esaldia objektiboa da baina iritzi negatiboa ere badu.

Lehen urratsean, subjektibitatearen sailkapena egiteko hainbat teknika aipatu ditugu. Teknika horiek 2.1 taulan laburbilduta eta, ondoren, azalduta daude.

Lana	**Teknika**
(Wiebe *et al.*, 1999)	Ikasketa gainbegiratua
(Wiebe, 2000)	Ikasketa gainbegiratu gabea
(Yu eta Hatzivassiloglou, 2003)	Esaldien arteko antzekotasuna eta Naïve Bayes
(Riloff *et al.*, 2006)	Unigrama, n-grama eta patroi lexiko-sintaktikoen arteko harremanak

2.1 taula: Esaldien subjektibitatearen sailkapena egiteko teknika ezberdinak.

Wiebe *et al.*ek (1999) subjektibitatearen sailkapena egiteko urre-patroia nola sortzen duten azaltzen dute. Etiketatzean desadostasunak agertu zaizkienez, aurreiritziak zuzentzeko etiketak sortzen dituzte bi helbururekin: i) eskuz etiketatzeko gidalerroen bertsio hobea sortzeko eta ii) sailkatzaile automatikoa sortzeko.

Urtebete geroago, Wiebek (2000) subjektibitatearen sailkapena egiteko ikasketa gainbegiratu gabea[2] egiten du. Kasu honetan, hitz-multzoen distribuzio semantikoa aintzat hartuz, subjektibitateari loturiko aztarnak lortzen ditu.

Yu eta Hatzivassiloglouk (2003), berriz, iritzidun esaldiak bilatzeko hiru teknika ezberdin erabiltzen dituzte: i) esaldien arteko antzekotasuna, ii) Naïve Bayes sailkatzailea[3] eta iii) Naïve Bayes anizkuna.

Bukatzeko, Riloff *et al.*ek (2006) ezaugarrien errepresentazioa egiteko sub-

[2]Ikasketa gainbegiratu gabea ikasketa automatikoko metodo bat da, behaketan oinarritzen den eredua da. Bertan, modelatze prozesu guztia sistemako sarreran dauden adibide multzo baten bidez gauzatzen da. Ez du adibide horien kategorien inguruko informaziorik. Hori horrela izanik, sistemak gauza izan behar du patroiak identifikatu eta, modu horretan, sarrera berriei etiketak esleitzeko.

[3]Naïve Bayes sailkatzaile probabilistiko bat da eta Bayes-en teoreman eta aldagaien arteko independentziaren hipotesian oinarritzen da. Sailkatzaile honen arabera, klase aldagai bat emanda, ezaugarriak elkarren artean independienteak dira eta, beraz, ez dago korrelaziorik.

suntzio hierarkia (*subsumption hierarchy*) erabiltzen dute, bertan, ezaugarri lexikal ezberdinak eta beren arteko harremanak finkatzeko. Modu horretan, esaldien subjektibitatearen sailkapena lortzen dute.

Bigarren urratsean, berriz, sentimenduen sailkapena egiten da eta esaldiak iritzia adierazten badu, iritzia positiboa edo negatiboa den zehatzen da.

(6) adibidearen kasuan, [2] esaldiak iritzi positiboa du, *good* ("ona") hitza dagoelako, baina [3] eta [4] esaldiek iritzi negatiboa dute: [3] esaldiaren kasuan, *bulky* ("tamaina handikoa") hitza ageri da eta [4] esaldian, azkenik, mugikorrak funtzionatzeari utzi diola aipatzen da.

Sentimenduen sailkapena egiteko ere hainbat teknika daude eta horiek 2.2 Taulan laburbilduta eta, ondoren, azalduta daude.

Lana	**Teknika**
(Yu eta Hatzivassiloglou, 2003)	Egiantz-arrazoiaren logaritmoa (*log-likelihood ratio*, LLR)
(Hu eta Liu, 2004)	Lexikoian oinarritutako algoritmoa
(Gamon, 2004)	Ikasketarako algoritmo erdi-gainbegiratua
(McDonald *et al.*, 2007)	Sekuentzia hierarkikoen ikasketa automatikoa

2.2 taula: Esaldi mailako sentimenduen sailkapena egiteko zenbait teknika.

Sentimenduen sailkapenaren urratsa gauzatzeko, Yu eta Hatzivassiloglouek (2003) hitzen orientazio semantikoa kalkulatzen dute eta horretarako Turneyk (2002) proposaturiko teknika erabiltzen dute. Hurrengo urratsean, esaldi osoaren orientazio semantikoa kalkulatzeko, berriz, egiantz-arrazoiaren logaritmoa[4] erabiltzen dute.

[4]Egiantz-arrazoiaren logaritmoak (*log-likelihood ratio*, ingelesez) bi eredu estatistikoen egokitzeko egokitasuna ebaluatzen du. Bi ereduok beren probabilitateen proportzioaren arabera lehiatzen dira; zehazki, bata parametroen espazio osoaren maximizazioarekin aurkitzen da eta bestea murriztapen batzuk ezarri ondoren aurkitzen da. Behatutako datuek murriztapena babesten badute (hau da, hipotesi nulua) bi probabilitateek ez dute laginketa-errore bat baino gehiagotan diferitu behar. Beraz, egiantz-arrazoiak erlazio hau bata bestearengandik oso ezberdina den edo bere logaritmo naturala zerotik guztiz ezberdina den ziurtatzen du.

Hu eta Liuek (2004), aldiz, hiru urrats egiten dituzte sentimenduen sailkapena gauzatzeko: i) kontsumitzaileen iruzkinetan agertu diren produktuaren ezaugarriak erauzi, ii) iruzkin bakoitzean iritzidun esaldia identifikatu eta positiboa edo negatiboa den adierazi eta iii) emaitzak laburbildu.

Bestalde, Gamonek (2004) ikasketa erdi-gainbegiratuen teknika erabiltzen du. SMV algoritmoa ezaugarri askoko bektoreekin erabiltzen du, zeintzuek bektore murrizketa jasan duten. Halaber, azterketa linguistiko sakonaren ezaugarriak hitzen n-grama funtzioei ere aplikatzen dizkie.

Azkenik, McDonald *et al.*ek (2007) testuaren granularitate-maila ezberdinetan testuen sailkapen semantikoa egiteko eredu egituratuak erabiltzen dituzte. Ereduaren inferentzia Viterbiren sekuentzia estandarren sailkapen tekniketan oinarritzen dira.

Ikusten den moduan, teknika ezberdinak daude esaldi mailako sentimenduen analisia egiteko, hots, esaldien iritzi positiboa edo negatiboa duten jakiteko.

2.1.3. Dokumentu mailako sentimenduen sailkapena

Aurretik esan bezala, ataza honetan dokumentu baten eduki subjektiboa aztertzen da, duen balorazioa positiboa edo negatiboa den identifikatuz. Ataza honetarako hainbat hurbilpen daude (Liu, 2012): i) sentimenduen sailkapena ikasketa gainbegiratua erabiliz (2.1.3.1 atala), ii) sentimenduen sailkapena ikasketa gainbegiratu gabea erabiliz (2.1.3.2 atala), iii) sentimenduen estimazio iragarpena (2.1.3.3 atala), iv) domeinu arteko sentimenduen sailkapena eta (2.1.3.4 atala) v) hizkuntza arteko sentimenduen sailkapena (2.1.3.5 atala).

2.1.3.1. Sentimenduen sailkapen gainbegiratua

Sentimenduen sailkapena ikasketa gainbegiratua egitean, dokumentu baten sentimendu sailkapen bitarra egiten da, hots, haren iritzia positiboa edo negatiboa den zehazten da. Zehaztapen hori kalifikazio bat emanez egiten da. Ikasketa gainbegiratuan, entrenamenduko eta testeko datuak iritzi-testuak dira eta bertan, iritzia 1 eta 5 bitarteko eskala erabiliz adierazten da non 1 eta 2 kalifikazioak negatiboak diren eta 4 eta 5 kalifikazioak, aldiz, positiboak.

Ikertzaile gehienek ez dute iritzi neutroa aintzat hartzen, modu horretan, sailkapena egitea erraza delako, baina beharko balitz, 3 kalifikazioa izango litzateke iritzi neutroa.

Sentimenduen sailkapena bereziki dokumentu mailako arazo bat da. Testu-sailkapen tradizionalean, hots, testu bat domeinu batean sailkatzeko arazoan, domeinu ezberdinetako dokumentuak sailkatu izan dira: politika, zientzia eta kirola, besteak beste, eta mota horretako sailkapenetan, domeinuari loturiko hitz gakoak garrantzitsuak dira, modu horretan, testua gai nagusiari edo bigarren mailako gai bati buruzko iritzia ematen ari den jakin daitekeelako. Baina sentimenduen sailkapenean, iritzi positiboa edo negatiboa adierazten duten sentimendu eta iritzidun hitzak garrantzitsuagoak dira (esaterako, *ederra*, *bikaina*, *txarra*), Liuren (2012) arabera.

Dokumentu mailako sentimenduen sailkapena testu-sailkapenaren ataza bat denez, ikasketa gainbegiratuan dauden teknika ezberdinak erabil daitezke. Dokumentu mailako sentimenduen sailkapena egiteko ikasketa automatiko gainbegiratuko aplikazioek zenbait ezaugarri erabili izan dituzte:

1- Hitzak eta horien maiztasunak. Banakako hitzek (unigramek) eta beren n-gramak[5] maiztasun-kontaketarekin osatzen dute ezaugarri hau. Batzuetan, posizioa ere kontutan hartzen da. Ezaugarri honek eraginkortasun handia erakusten du sentimenduen analisian.

2- *Part-Of-Speech* (POS). Hitz bakoitzaren gramatika-kategoria jakitea ere garrantzitsua da; izan ere, sentimenduen analisian gramatika-kategoria bakoitzak garrantzi ezberdina du. Hala, adjektiboak iritziaren adierazle garrantzitsuak dira. Adjektiboek halako garrantzia dutenez, batzuetan ezaugarri berezi moduan tratatzen dira. Beste batzuetan, aldiz, n-gramak eta POS batera erabiltzen dira ezaugarri berean.

3- Sentimendu-hitzak eta -sintagmak. Sentimendu hitzak iritzi positiboak eta negatiboak adierazteko erabiltzen dira. Adibidez, *on*, *zorioneko* eta

[5]N-grama sekuentzia baten parte den azpi-sekuentzia bat da. Azpi-sekuentzia hori hainbat n elementuz osatuta dago eta elementu horiek hainbat motatakoak izan daitezke; letrak edo hitzak, esaterako.

eder hitz positiboak dira. *txar*, *zoritxarreko* eta *itsusi*, aldiz, hitz negatiboak dira. Sentimendu-hitz gehienak adjektiboak eta adberbioak dira baina izenak (*maitasuna*) edota aditzak (*gorrotatu*) ere izan daitezke. Sentimendu-hitzez gain, sentimendu-sintagmak ere badaude; *leher eginda* lokuzioa oso nekatuta zaudela adierazteko bezalakoak.

4- Iritzi-erregelak. Sentimendu-hitz eta -sintagmez gain, badaude iritziak adierazteko zenbait adierazmolde iritzia edo sentimendua adieraz edo inplika dezaketenak. Esaterako, *bateria-bizitza luzea du* eta *bateria-bizitza laburra du* esaldietan, aurkako bi orientazio semantiko daude; baina halaber, *bateria-bizitza* hitzak berak orientazio semantiko positiboa inplikatzen du. Hau da, mugikorrak *bateria-bizitza* bera edukitzea positiboa da baina horren baitan, *bateria-bizitza luzea* positiboa den bezala *bateria-bizitza laburra* negatiboa da.

5- Sentimendu-aldatzaileak. Badaude zenbait adierazmolde orientazio semantikoa aldatzeko erabiltzen direnak. Orientazio semantiko positiboa negatibo bilaka dezakete edota alderantziz. Sentimenduen aldatzaileetako bat ezeztapena da. *Kamera hau ez zait gustatzen* esaldian, esaterako, *gustatu* hitzaren orientazio semantiko positiboa negatibo bihurtzen du.

6- Dependentzia sintaktikoak. Dependentzia-zuhaitzetatik sortutako hitzen dependentzia-ezaugarriak ere erabiltzen dira dokumentu mailako sentimendu-sailkapenean.

Dokumentu mailako sentimenduen sailkapena egiteko erabilitako teknikak 2.3 Taulan daude laburbilduta.

Aurretik aipatutako ezaugarriak erabiliz, zenbait teknika erabili izan dira. Pang *et al.*ek (2002) filmeen iruzkinak testu positibo eta negatibo bezala sailkatzen dituzte eta horretarako n-gramak (hitz-poltsak) erabiltzen dituzte sailkapeneko ezaugarri moduan Naïve Bayes nahiz SVM sailkatzaileak erabiliz.

Lana	Teknika
(Pang *et al.*, 2002)	1) Naïve Bayes sailkatzailea 2) Euskarri bektoredun makina (*Support Vector Machines*, SVM)
(Mullen eta Collier, 2004)	SVM + Erlazio sintaktikoak eta ohiko ezaugarriak
(Nakagawa *et al.*, 2010)	Dependentzia-zuhaitzean oinarritutako sailkatzailea + baldintzazko ausazko eremuak (*Conditional Random Fields*), CRF
(Kouloumpis *et al.*, 2011)	Ezaugarri linguistikoak + mikroblogetako hizkuntza kreatiboa eta ez-formala identifikatzeko ezaugarria

2.3 taula: Dokumentu mailako sentimenduen sailkapena egiteko erabilitako zenbait teknika.

Mullen eta Collierek (2004) SVM sailkatzailea[6] hainbat ezaugarrirekin konbinatzen dute, hala nola PMI[7] bidez lorturiko orientazio semantikoak, Wordneten oinarrituz egindako desberdintze semantikoa eta gaiarekiko hurbiltasuna eta erlazio sintaktikoen ezaugarriak.

Nakagawa *et al.*ek (2010), berriz, ingeleseko eta japonierako datu subjektiboak sailkatzeko dependentzia-zuhaitzak eta baldintzazko ausazko eremua[8]

[6]Euskarri bektoredun makina (ingelesez, *Support Vector Machine*, SVM) sailkapen nahiz erregresiorako erabiltzen den algoritmo multzoa da. Bere oinarrian bektore-espazioak daude eta sarreran, bere n dimentsio ereduan adieraz daitezkeen hainbat datu jasotzen ditu, bi kategorietatik batean sailkatuta daudenak. Ondoren, teknika honek bere ereduan dauden puntuak bereizteko hiperplanoa bilatzen du. Teknika honen kasuan gertueneko puntuetarako distantziarik handiena duena aukeratzen du eta hori da emaitza.

[7]Elkarrekiko Informazio Puntuala (ingelesez, *Pointwise mutual information*) estatistikan eta informazioaren teorian erabiltzen den asoziazio neurria da. Hitz batek (x) beste hitz batekin (y) duen probabilitatea hitzaren beraren (x) probabilitatearekin zatituz lortzen da. Esaterako, *puerto* hitza 1.938 aldiz agertzen da eta *rico* 1.311 aldiz. Biak batera, berriz, 1.159 aldiz agertzen dira. Horrenbestez, PMI neurria 10,03 da.

[8]Baldintzazko ausazko eremuak (ingelesez, *Conditional Random Fields*, CRF) dokumentuetatik datuak erauzteko edo datu-sekuentziak segmentatu eta etiketatzeko erabiltzen den eredu estokastikoa da. Eredu honek datu-sekuentzia bat emanda elementuetako bakoitzari (*Oi*) etiketa bat (*Si*) ematen dio. Eredu sortzailea denez, behaketen eta eti-

erabiltzen dituzte. Lanean, esaldi bakoitzaren azpi-zuhaitz baten orientazio semantikoa aldagai bat da eta esaldi osoaren orientazio semantikoa aldagai horien arteko interakzioa kalkulatuz lortzen da.

Kouloumpis *et al.*ek (2011), berriz, txioen sentimenduen sailkapena egiteko, hainbat ezaugarri erabiltzen dituzte: n-gramen ezaugarriak, lexikoiko ezaugarriak (hitzak positiboak, negatiboak edo neutralak diren ere barne hartuz), gramatika-kategoria ezaugarriak eta, bukatzeko, mikroblogetako testuen ezaugarriak.

2.1.3.2. Sentimenduen sailkapen gainbegiratu gabea

Dokumentu mailako sentimenduen sailkapena ikasketa gainbegiratu gabea erabiliz egiten duten lanak ere badaude. Sentimendu-hitzak oinarrizko faktoreak dira sentimenduen sailkapenean eta horregatik, horiek modu gainbegiratu gabean erabil daitezke. Bi hurbilpen daude ikasketa gainbegiratu gabean: i) erregela sintaktikoetan oinarritutakoa eta ii) lexikoian oinarritutakoa.

Bi hurbilpenak argitzeko Liuren (2012) *this piano produces beautiful sounds* esaldia erabiliko dugu.

	Lehen hitza	**Bigarren hitza**	**Hirugarren hitza (ez erauzia)**
1	JJ	NN edo NNS	edozer
2	RB, RBR edo RBS	JJ	ez NN edo NNS
3	JJ	JJ	ez NN edo NNS
4	NN edo NNS	JJ	ez NN edo NNS
5	RB, RBR edo RBS	VB, VBD, VBN edo VBG	edozer

2.4 taula: Dokumentu mailako sentimenduen sailkapen gainbegiratu gaberako erregela sintaktikoak (Turney, 2002).

Turneyk (2002) erregela sintaktikoen hurbilpena jarraitzen du. Erregela sin-

keten probabilitatearen banaketa aldi berean modelatzen ditu eta baldintzazko ausazko eremuek behaketek baldintzatutako etiketen sekuentzia zuzenen probabilitatea modelatzen dute (*P(S—O)*). Beraz, eredu diskriminatzaile bat da.

taktikoak gramatika-kategorien POS etiketetan (Liu, 2012)[9] oinarritzen dira eta guztira hiru urrats daude hurbilpen honetan:

1. urratsa. Jarraian dauden bi hitz erauzi, baldin eta bi hitz horien POS etiketek 2.4 taulako patroi bat osatzen badute. Esaterako, bigarren patroiak adierazten du lehen hitzak adberbioa, bigarrenak adjektiboa eta hirugarrenak (erauzten ez denak) izena izan behar duela. *This piano produces beautiful sounds* esaldiaren kasuan, *beautiful sounds* dugu, lehen hitza adjektiboa da eta bigarrena, berriz, izena. Beraz, 2.4 taulako lehen patroiarekin bat egiten du eta bi hitzak erauzi egiten dira. JJ, RB, RBR eta RBS POS etiketak dituzten hitzek askotan iritzia adieraztea da patroiak erabiltzearen arrazoia eta berekin agertzen diren izen eta aditzek beren testuingurua dira. Izan ere, adjektiboek esaterako (*ikaragarri* kasu) testuinguruaren arabera sentimendu positiboa edo negatiboa izan dezakete.

2. urratsa. Aurretik erauzitako hitz-multzoaren orientazio semantikoa kalkulatzen da elkarrekiko informazio puntuala (PMI) neurria (2.1 irudia) erabiliz.

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right)$$

2.1 irudia: *Pointwise mutual information*, PMI.

2.1 irudiko PMI neurriak bi hitzen arteko dependentzia estatistikoa neurtzen du. Pr(term1∧term2) *term1* gertatzeko ko-okurrentzia probabilitatea eta Pr(*term1*)Pr(*term2*) bi hitzen ko-okurrentzia probabilitatea da, baldin eta bi hitzak estatistikoki independienteak badira.

(7) SO(phrase) = PMI(phrase, “beautiful”) - PMI(phrase, “ugly”).

[9]Hauexek dira erregeletan azaltzen diren laburpenen esanahiak: JJ: adjektiboa, RB: adberbioa, RBR: konparaziozko adberbioa, RBS: superlatibozko adberbioa, NN: izen singularra, NNS: izen plurala, VB: aditza erro moduan, VBD: aditza iraganaldian, VBN: aditza iragan partizipioan eta VBG: aditza orainaldian eta orainaldiko partizipioan.

Liuren (2012) *this piano produces beautiful sounds* sintagmaren orientazio semantikoa (SO) kalkulatzeko, (7) adibidean ikusten den moduan, erreferentzia positiboko hitza den *beautiful* eta negatibokoa den *ugly* hitzaren arteko asoziazioa neurtzen da.

Probabilitatea kalkulatzeko, bilaketa-motor[10] bati kontsulta eskatzen zaio eta bisita kopurua biltzen du hark. Bilaketa bakoitzeko kontsulta batean, bilaketa-motor batek kontsultari lotutako dokumentu kopuru aipagarriak eskaintzen ditu eta hori emaitza kopurua da. Beraz, bi hitz batera eta bereiz bilatuta 2.1 irudiko probabilitatea kalkula daiteke.

3. urratsa. Iritzi-testu bat emanda, algoritmoak iritzi-testuan dauden sintagma guztien orientazio semantikoa neurtzen du eta iritzi-testua positibotzat (orientazio semantiko guztien batezbestekoa positiboa bada) edo negatibotzat (orientazio semantiko guztien batezbestekoa negatiboa bada) sailkatzen du.

Taboada *et al.*ek (2011) ere ikasketa gainbegiratu gabea erabiltzen du. Bertan, dokumentu mailako sentimenduen sailkapena egiteko SO-CAL tresna aurkezten du[11]. Tresna honek hitz, esaldien eta testuen orientazio semantikoa kalkulatzen du.

Tresna honen abiapuntua −5 eta +5 arteko orientazio semantikoa duten hitzen lexikoia da. Guztira, lexikoiak lau gramatika-kategorietako hitzak biltzen ditu: izenak, adjektiboak, adberbioak eta aditzak. SO-CAL tresnak hitz horiek iritzi-testu batean bilatzean erauzi egiten ditu. Hurrengo urratsean, tresnak hizkuntzaren beste ezaugarri batzuk erauzten ditu, hala nola intentsifikatzaileak, ezeztapena edota puntuazio-zeinuak. Erauzketa egin ostean, lehen urratsean erauzitako hitzen sentimendu-balentziak aldaketa batzuk jasatzen ditu, aipatu berri ditugun hizkuntza fenomenoen ondorio. Aldaketa hori sentimendu-balentzia indartzea edo ahultzea izan daiteke.

[10]Bilaketa-motorra informazioa bilatzeko helburuarekin sortutako informazioa eskuratzeko sistema da. Baldintza batzuk zehazten dira motorrean eta horren araberako emaitzak lortzen dira zerrenda batean garrantzi handienetik txikienera.

[11]Sentimenduen sailkapena egiten duen SO-CAL tresnari buruzko informazio gehiago: atal hauetan dago: 4.3.1 atalean non tresnari buruzko informazio gehiago ematen den eta 4.3.2 atalean non funtzionamendu bera duen euskarazko bertsioa deskribatzen den.

Esaterako, intentsifikatzaileek hitzaren orientazio semantikoa indartu edo ahuldu dezake. Bestalde, sintagman edo esaldian ezeztapen-marka agertuz gero, horren orientazio semantikoa ahuldu egiten du tresnak, ezeztapen-markaren ±4 balentzia aplikatuz. Hau da, sintagma edo esaldiaren orientazio semantikoa negatiboa bada eta ezeztapena badago, ezeztapen-markaren +4 aplikatuz, haren orientazio semantikoa indartu egingo du. Sintagma edo esaldiaren orientazio semantikoa positiboa bada eta ezeztapena badago, aldiz, ezeztapen-markaren −4 balentziak, haren orientazio semantikoa ahulduko du.

Bukatzeko, SO-CAL tresnak puntuazio-zeinuak ere kontuan hartzen ditu eta, esaterako, komen artean dauden hitzei ez die orientazio semantikoa esleitzen, izenburuak edota aipuak bezalakoak direlako.

Liuren (2012) esaldia hurbilpen honetan aplikatzen badugu, lehenik eta behin, SO-CALen lexioiak esaldiko *beautiful* (+4) hitza erauzten du. Hurrengo urratsean, inguruan intentsifikatzailerik eta ezeztapenik ez dagoela antzematen du, eta horrenbestez, hitzaren orientazio semantikoari ez dizkio hizkuntza-fenomeno hauen eraginak aplikatzen. Bukatzeko, tresnak hitza komen artean ez dagoela egiaztatzen du eta horrenbestez, hitzaren orientazio semantikoa bere horretan mantentzen du. Esaldian, orientazio semantikodun beste hitzik ez dagoenez, *beautiful* hitzaren orientazio semantikoa da (+4) esaldiaren orientazio semantikoa.

Gure tesi-lanak sentimenduen analisi gainbegiratu gabearekin du zerikusia. Izan ere, garatu nahi dugun sentimendu-sailkatzailea lexikoian oinarritutakoa izango da.

2.1.3.3. Sentimenduen estimazio-iragarpena

Dokumentu mailako sentimenduen sailkapeneko beste arloetako bat sentimenduen estimazio-iragarpena (*Sentiment Rating Prediction*) da. Arlo honetan, dokumentu batek iritzi positiboa edo negatiboa duen sailkatzetik urrunago joan nahi da eta zenbakizko kalifikazio bat (1 eta 5 artean) esleitu nahi zaio. Kasu honetan, ataza erregresio[12] moduan trata daiteke; izan ere, kalifikazio-puntuazioak ordinalak dira.

[12]Erregresioa aldagai dependente baten eta independente zenbaiten arteko erlazioa zehaztu eta aztertzen duten teknika estatistikoa da.

Sentimenduen estimazio iragarpena-egiteko aipatuko ditugun teknika ezberdinak 2.5 taulan daude laburbilduta.

Teknika	**Lana**
SVM erregresioa	Pang eta Lee (2005)
Hitz-poltsen errepresentazioa	Qu *et al.* (2010)
Aspektuaren sentimenduen estimazio-iragarpena: Erregresio estandarra	Snyder eta Barzilay (2007)
Aspektuaren sentimenduen estimazio-iragarpena: Sare Bayesiarrean oinarritutako sailkatzailea	Long *et al.* (2010)

2.5 taula: Sentimenduen estimazio-iragarpena egiteko teknika ezberdinak.

Pang eta Leek (2005) SVM erregresioa erabiltzen du. Zehazki, klase anitzeko SVM sailkatzailea[13] erabiltzen du.

Bestalde, Qu *et al.*ek (2010) dokumentuko hitz-poltsen errepresentazioa darabilte iritzia duten n-gramen indarra neurtzeko. Berau ezberdina da ohiko hitz-poltsen errepresentaziotik. Teknika honetan iritzietako bakoitza hirukoitza da: sentimendu hitza, aldatzailea eta ezeztapena baititu. Esaterako, *ez oso ona* adibidean, *ez* ezeztapena da, *oso* aldatzailea eta *ona*, berriz, sentimendu-hitza. Modu honetan, sentimendu-hitzetik haratago dauden baino berengan eragiten duten fenomenoak identifika ditzakete.

Snyder eta Barzilayk (2007) beste ikuspegi batekin lantzen dute ataza. Iritzi bakoitzari zenbakizko kalifikazio bat esleitu beharrean, aspektuetako bakoitzari esleitzen diote kalifikazioa eta, horretarako, erregresio estandarra edo sailkatze-teknikak erabiltzen dituzte.

Azkenik, Long *et al.*ek (2010) iritzietako aspektu bakoitzari kalifikazioa esleitzeko sare Bayesiarra darabilen sailkatzailea erabiltzen dute.

[13]Klase anitzeko SVM teknikan, datuak bi kategoria baino gehiagotan sailkatzeko bi hurbilpen daude: i) kategoria bakoitza beste batzuetan zatitzen da eta denak konbinatzen dira eta ii) k(k-1)/2 ereduak sortzen dira non k kategoria kopurua den.

2.1.3.4. Domeinuen arteko sentimenduen analisia

Dokumentu mailako beste arlo bat domeinu arteko sentimenduen sailkapena da. Ikasketan erabilitako domeinuek eragin handia dute sentimenduen sailkatzaileetan. Horregatik, domeinu jakin bateko iritzi-testuekin ondo dabilen sentimenduen sailkatzaile bat, beste domeinu bateko iritzi-testuekin ez da ondo ibiltzen (Liu, 2012).

Hitz batek aurkako bi orientazio semantiko eduki ditzake bi domeinu ezberdinetan eta hori da arazo honen jatorria. Adibidez, *luzea* adjektiboaren kasuan, telefono mugikorren domeinuan bateria-bizitzari zuzenduta dagoenean orientazio positiboa du. Adjektibo berak, aldiz, zinemaren domeinuan orientazio negatiboa du: *filma luzea egin zitzaidan.*

Arlo honetan bi hurbilpen daude. Horietako bat iturri-domeinua (*source domain*) da. Gamon *et al.* (2005) da lan aipagarrietako bat hurbilpen honetan, etiketatutako datu ikasiak domeinu berri batean erabiltzen dira. Beste hurbilpena helburu-domeinua (*target domain*) da. Blitzer *et al.* (2007); Tan *et al.* (2005) dira lanik nabarmenenak eta kasu honetan, domeinu berri baterako ez dute etiketatutako daturik erabiltzen.

2.1.3.5. Hizkuntza arteko sentimenduen analisia

Orain arte deskribatu eta aipatu ditugun hurbilpenak eta lanak hizkuntza bati bideratuta egon dira. Hala ere, badira hizkuntza arteko sentimenduen analisia egiten duten hurbilpen eta teknikak ere, bai esaldi mailan, bai dokumentu mailan.

Esaldi mailako hizkuntza arteko sentimenduen analisian ere bi ataza daude: i) hizkuntza arteko subjektibitatea eta ii) sentimenduen sailkapena.

Arlo honetan, hizkuntza askotan baliabide falta dagoenez, ingelesetik beste hizkuntzetara sentimenduen analisirako baliabide bat automatikoki itzultzea da helburua. Liu (2012) lanaren arabera, hiru estrategia nagusi daude baliabideak itzultzerakoan:

1- Helburu-hizkuntzako test-multzoko esaldiak iturri-hizkuntzara itzuli eta bertan, iturri-hizkuntza iturriko hizkuntzan dagoen baliabidea edo sis-

tema erabiliz test-multzoko esaldiek informazio objektiboa edo subjektiboa duten sailkatu.

2- Iturri-hizkuntzako ikasketarako corpusa[14] helburu-hizkuntzara itzuli eta corpusean oinarritutako sailkatzailea sortu helburu-hizkuntzan.

3- Iturri-hizkuntzako sentimenduen lexikoia helburu-hizkuntzara itzuli eta lexikoian oinarritutako sailkatzailea sortu helburu-hizkuntzan.

Hurbilpen honetan aipaturiko lanak 2.6 taulan laburbilduta daude.

Teknika	**Lana**
Helburu-hizkuntzako datuak jatorri-hizkuntzara itzuli eta sentimenduen sailkapena egin	Kim eta Hovy (2006) Banea *et al.* (2008)
Sentimenduen sailkapena egiteko lexikoia helburu-hizkuntzara itzuli eta, gero, sentimenduen sailkapena egin	Kim eta Hovy (2006) Banea *et al.* (2008) Bautin *et al.* (2008)
Sentimenduen sailkapena jatorri-hizkuntzan eta, ondoren, itzulpena egin	Banea *et al.* (2008)
Eleaniztasunaren konparagarritasuna	Kim *et al.* (2010)

2.6 taula: Esaldi mailako hizkuntza arteko sentimenduen analisirako zenbait teknika.

Kim eta Hovyek (2006) bi esperimentu gauzatzen dituzte: i) alemanerazko emailak ingelesera itzuli eta itzulpenei ingelesezko sentimendu-hitzak aplikatu, horien orientazio semantikoa ikusteko eta ii) ingelesezko sentimendu-hitzak alemanerara itzuli eta itzulitakoa alemanerazko emailetan aplikatu.

Banea *et al.*ek (2008) hiru esperimentu egin dituzte: i) ingelesezko datu etiketatuak (iturri-hizkuntza) automatikoki errumanierara itzuli (helburu-hizkuntza), ii) jatorri-hizkuntza automatikoki subjektibitatez etiketatu eta ondoren helburu-hizkuntzara itzuli eta iii) helburu-hizkuntza jatorri-hizkuntzara itzuli eta bertan subjektibitatearen sailkapen tresna aplikatu.

[14]Ikasketa automatikoan, corpusa hiru zatitan banatu ohi da: garapena, ikasketa eta testa. Ikasketa zatia, hitzak berak adierazten duen moduan, ikasketa automatikoko teknikek esperientzia trebetasun edo jakintza bihurtzeko baliatzen dute.

Bautin *et al.*ek (2008), berriz, aspektuan oinarritutako hizkuntza arteko sentimenduen analisia egiten dute. Helburu-hizkuntza iturri-hizkuntzara itzultzen dute (ingelesera) eta ondoren, entitatea duen esaldi bakoitzari orientazio semantikoa esleitzen diote ingelesezko lexikoian oinarritzen den metodoa erabiliz.

Bukatzeko, Kim *et al.*ek (2010) eleaniztasunarem konpagarritasuna (ingelesez, *multilanguage-comparability*) metodoa aplikatzen dute. Hots, testu eleanitz bateko hitz-bikoteek subjektibitate esanahi bera dutela baliatuz, sailkapen emaitzen hitz-bikoteek duten adostasun-maila neurtzen dute.

Dokumentu mailako hizkuntza arteko sentimenduen analisia ere badago. Hizkuntza arteko sentimenduen analisia egiteko bi motibazio nagusi daude:

i) Ikertzaile batek sentimenduen analisirako sistema bere hizkuntzan sortu nahi du eta, bere hizkuntzan horrelakorik sortzeko baliabiderik ez dagoenez eta sistema gehienak ingelesez daudenez, ingeleseko sisteman oinarrituz bere hizkuntzan sistema sortzeko erabiltzen du.

ii) Aplikazio edo produktuetan, enpresek pertsonen iritziak jaso nahi dituzte eta, hori lortzeko, ingelesez duten sistema beste hizkuntzetara itzultzen dute.

Dokumentu mailan, aipatuko ditugun hizkuntza arteko sentimenduen analisirako teknika ezberdinak 2.7 taulan laburbilduta daude.

Teknika	**Lana**
Hainbat itzultzaile automatiko	Wan (2008)
Ko-ikasketa metodoa (SVM barne)	Wan (2009)
Transferentzia-ikaskuntza metodoa	Wei eta Pal (2010)
Gai-eredua	Guo *et al.* (2010)

2.7 taula: Dokumentu mailan, hizkuntza arteko sentimenduen analisia egiteko zenbait teknika.

Wanek (2008) ingeleseko baliabideak ustiatzen ditu txinerazko iritzi-testuen sentimenduen sailkapena egiteko. Bere algoritmoan, txinerazko iritzi-testuak

hainbat itzultzaile erabiliz ingelesera itzultzen ditu eta itzulitako testuei ingelesezko sentimenduen lexikoia aplikatzen die.

Wanek (2009), ordea, ko-ikasketa metodoa[15] erabiltzen du. Zehazki esanda, txinerazko iritzi-testuen sailkapena egiteko etiketatuta dagoen ingelesezko corpusa erabiltzen du modu gainbegiratu batean (SVM teknika erabiliz).

Wei eta Palek (2010), berriz, transferentzia-ikaskuntza metodoa[16] proposatzen dute hizkuntza arteko sentimenduen analisia egiteko. Itzulpen-prozesuan bertan, itzultzaile automatikoak sortzen duen zarata txikiagotzen du.

Azkenik, Guo *et al.*ek (2010) gai-ereduan oinarritzen den metodoa[17] proposatzen dute hizkuntza askotako iritzi-adierazpenak multzokatu eta, modu horretan, estatu askotako iritzien aspektuan oinarritutako sentimenduen sailkapena egiteko.

2.1.4. Sentimenduen analisiko beste atazak

Liuk (2012) adierazten duenez, dokumentu maila, esaldi maila eta entitate eta aspektu maila sentimenduen analisiko ataza nagusiak dira; izan ere, maila horietako bakoitza objektiboa edo subjektiboa den zehaztu, horien orientazio semantikoa zehaztu eta azkenik, iritziaren hartzailea eta entitate baten aspektu askoren orientazio semantikoa zehazten baita.

Hala ere, sentimenduen analisia zabalagoa da eta badaude bestelako atazak ere non sentimenduen analisia hizkuntza naturalaren prozesamenduko beste

[15]Ikasketa automatikoan, ko-ikasketa metodoa (ingelesez, *co-training*) etiketatutako datu-multzo txiki bat eta etiketatu gabeko datu-multzo handi bat erabiltzean datza.

[16]Transferentzia-ikaskuntza (ingelesez, *transfer learning*) ikasketa automatikoko ikerketa arazo bat da eta arazo bat ebatzi ostean lorturiko jakintza gorde egiten du; ondoren, arazo ezberdin baina antzeko batean aplikatzeko. Adibidez, autoak sailkatzeko erabilitako jakintza kamioiak sailkatzeko arazoan erabil daiteke. Psikologian dagoen transferentzia-ikaskuntzarekin nolabaiteko harremana du.

[17]Ikasketa automatikoan eta lengoaia naturalaren prozesamenduan gai-eredua (ingelesez, *topic model*) testu multzo batean ageri diren "gai" abstraktuak ezagutzeko eredu estatistiko mota bat da. Testuen baitan dauden ezkutuko egitura semantikoak ezagutzeko maiz erabiltzen den testu erauzketarako tresna da. Adibidez, eguraldia gaian den testuetan *eguzki* eta *tenperatura* hitzak maiztasun handiz agertzea espero da eta kirolean, aldiz, *irabazi* eta *talde*. Bestalde, *da* hitza bietan antzeko maiztasun batez agertzea espero da. Teknika honek sortzen dituen "gaiak" antzekotasuna duten hitz-multzoak dira.

arloekin uztartzen den. Ataza horiek gure tesi-lanaren helburuetatik urrutiago daudenez, azaletik esplikatuko ditugu.

2.1.4.1. Iritzi-testuen laburpena

Sentimenduen analisiko ataza ezberdinetan, pertsona askoren iritzia jakitea beharrezkoa da; bakoitzak bere ikuspegia duelako. Pertsona askoren iritzia bildu behar izateak iritzi-testuen laburpena egin beharra eragin du. Bestalde, aspektuan oinarritutako sentimenduen sailkapenak (2.1.1 atalean erakutsi bezala) iritzi-testuen laburpen bat egiteko adina informazio eskatzen du. Egoera honetan, iritzi-testuen laburpena eta aspektu mailako sentimenduen sailkapena batzearen ondorioz, aspektuan oinarritutako iritzi-testuen laburpena sortu da. Aspektuan oinarritutako iritzi-laburpenaren emaitza bi modutakoa izan daiteke: i) testu egituratua (2.2 irudian bezala) edota ii) testu desegituratua.

Digital Camera 1:
Aspect: **GENERAL**
Positive: 105 <individual review sentences>
Negative: 12 <individual review sentences>
Aspect: **Picture quality**
Positive: 95 <individual review sentences>
Negative: 10 <individual review sentences>
Aspect: **Battery life**
Positive: 50 <individual review sentences>
Negative: 9 <individual review sentences>
...

2.2 irudia: Aspektuan oinarritutako iritzi-testuen laburpenaren emaitza egituratua (Cambria *et al.*, 2017).

2.2 irudian, iritzi-testu baten laburpenaren emaitza egituratua ikus daiteke. Bertan argazki-kamera bati buruzko iritzia ematen da eta beraren zenbait aspektu azaltzen dira: orokorra, argazkien kalitatea eta bateriaren bizitza. Aspektu horietako bakoitzaren eskuinetara, zenbat iritzi-testuk positiboki edo negatiboki ebaluatu duten zehazten da. Adibidez, argazkiaren kalitatea 95 iritzi-testuk positiboki baloratu dute eta 10 iritzi-testuk negatiboki.

Aspektuan oinarritutako iritzi-laburpena dagoen moduan, badago dokumen-

tu askotako testu-laburpena ere. Carenini *et al.* (2006) eta Hu *et al.* (2017) eremu horretako lanak dira. Hala ere, iritzi-testuen laburpena ezberdina da dokumentu baten edo askoren laburpenetik. Izan ere, iritzi-testuen laburpena aspektu eta entitateetan eta aspektu eta entitate horietako bakoitzaren orientazio semantikoan oinarritzen da. Gainera, alderdi kuantitatiboa ere badago, hau da, horietako bakoitza positiboa edo negatiboa den zehaztu behar da. Testuen laburpen tradizionalean, aldiz, esaldirik garrantzitsuenak erauzten dira eta dokumentu askotako testu-laburpenean, berriz, errepikatzen ez den informazioa bilatu eta errepikatzen den informazioa ezabatzen da.

Aspektuan oinarritutako iritzi-laburpenaren lanik aipagarrienak Hu eta Liurenak (2004) eta Liu *et al.*renak (2005) dira. Dokumentu askotako iritzi-laburpenean, aldiz, oinarrizko lana Das eta Martinsena (2007) da.

2.1.4.2. Konparaziozko iritzien analisia

Entitate bati edo bere aspektuei buruzko iritzi positiboa edo negatiboa zuzenean adierazteko aukeraz gain, antzeko bi entitateen arteko konparazioa eginez ere iritzia adieraz daiteke. Horrelako iritziei konparaziozko iritziak deritze.

Konparaziozko esaldiek beti ez dute zuzenean iritzia adierazten. Hau da, objektu bati buruzko iritzia zuzenean eman beharrean, beste objektu batetiko konparazioa da iritzia. Gainera, konparaziozko iritzien barruan, superlatibozko iritziak ere topa ditzakegu.

2.8 taulan, konparaziozko iritzietan aipatuko ditugun lanak laburbilduta daude.

Teknika	**Lana**
Gako-hitzak + SVM	Jindal eta Liu (2006)
Konparaziozko esaldiak identifikatzeko erregelak	Liu *et al.* (2010)
Lexikoian oinarritutako hurbilpena + aspektuan oinarritutako sentimenduen sailkapena	Ding *et al.* (2009)

2.8 taula: Konparaziozko iritziak lantzeko zenbait teknika.

Konparaziozko iritziak identifikatzeko hainbat teknika ezberdin erabili izan

dira. Jindal eta Liuk (2006) erakusten dutenez, konparaziozko iritziek gako-hitzak eduki ohi dituzte, mota horretako hitzak antzematen laguntzen dutenak. Adibidez, konparaziozko adjektiboak (*gehiago*, *gutxiago*) eta superlatibozko adjektiboak (*gehien*, *gutxien*) gako-hitzak dira.

Bestalde, Liu *et al.*ek (2010) konparaziozko galderak eta konparatzen diren entitateak ikertzen dituzte. Konparaziozko esaldiak identifikatzeko erregelak erabiltzen dituzte.

Azkenik, Ding *et al.*ek (2009) konparaziozko iritzi batean entitate gustukoena identifikatzea dute helburu eta bi urrats burutzen dituzte: i) entitatea ezagutzea eta ii) entitatea esleitzea.

2.1.4.3. Iritzi-bilaketa eta -berreskuraketa

Iritzi-bilaketa web-bilaketaren atazaz eta sentimenduen analisiko atazaz osatuta dago:

1- Web-bilaketako ataza: i) bilaketarekiko garrantzitsuak diren dokumentuak eta esaldiak bilatu behar dira eta, ondoren, ii) esaldi edo dokumentuak garrantziaren arabera sailkatu behar dira.

2- Sentimenduen analisiko ataza: i) bilatutako dokumentu edo esaldiak bilaketa-gaiari (entitatea edo aspektua) buruzko iritzia baduen zehaztu behar da eta, hala bada, ii) iritzia positiboa edo negatiboa den zehaztu behar da.

Iritzi-bilaketan aipatu ditugun teknika ezberdinak 2.9 taulan esplikatuta daude.

Iritzi-bilaketa egiteko zenbait teknika proposatu izan dira. Zhang *et al.*ek (2007) lehenik eta behin, bilaketarekiko garrantzitsuak diren dokumentuak bilatzen dituzte. Horretarako, kontzeptuen desanbiguazioa, kontzeptu horien sinonimoen bilaketa eta dokumentuen antzekotasuna neurtzen dituzte eta, ondoren, SVM erabiltzen dute sentimenduen sailkapena egiteko.

Eguchi eta Lavrenkok (2006), berriz, sorkuntzazko hizkuntza modelatzen[18]

[18]Modelatu esatean, hizkuntzaren eredua sortu nahi dela esan nahi da.

Teknika	Lana
Desanbiguazioa, sinonimoen bilaketa, dokumentuen antzekotasuna + SVM	Zhang *et al.* (2007)
Sorkuntzazko hizkuntza-eredua	Eguchi eta Lavrenko (2006)
Lexikoian oinarritutako iritzi-bilaketa	Na *et al.* (2009)
Lexikoen eta sentimenduen ezaugarriak + algoritmo ezberdinak	Liu *et al.* (2009)

2.9 taula: Iritzi-bilaketa egiteko aipatutako teknikak.

dute iritzi-bilaketa egiteko. Horretarako, bilaketako hitz multzo adierazgarri bat, intereseko sentimendu-balentziadun hitzak erabiltzen dituzte.

Na *et al.*ek (2009), ordea, ikuspegi ezberdina lantzen dute; izan ere, lexikoian oinarrituta egiten baitute iritzi-bilaketa. Domeinuari loturiko lexikoia sortzen dute horretarako, eta lexikoiaren oinarrian *feedback* bidezko ikasketa dago. Ikasketa egiteko, bilaketa bateko dokumentuak erabiltzen dituzte.

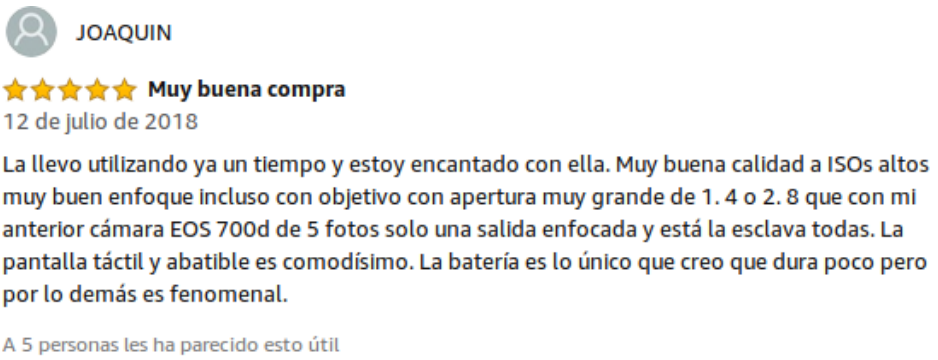
Azkenik, Liu *et al.*ek (2009) iritziak dauden blog batetik algoritmo ezberdinak eta lexikoen eta sentimenduen ezaugarriak ikasten dituzte. Modu horretan, sentimenduen analisia eta osagaien berreskuraketa uztartzen dituen estrategia garatzen dute.

2.1.4.4. *Spam*ak diren iritzien detekzioa

Norbanakoek eta erakundeek sare sozialetako iritziak gero eta gehiago hartzen dituzte kontuan. Iritzi horiek produktu baten erosketan, hauteskundeetan, marketinean edota produktu baten diseinuan eragin dezakete. Testuinguru honetan, iritzi positiboak arrakasta eta irabaziekin lotzen dira eta horregatik, norbanakoek gezurrezko iritziak eta iruzkinak idazten dituzte; produktu, zerbitzu, pertsona edo ideiei ospea emateko edo ospea kentzeko. Horrelako iritzi-motak antzemateko *spam*ak diren iritzien detekzioa garatu da. Li *et al.* (2011), Lim *et al.* (2010) eta Wang *et al.* (2012) dira lanik aipagarrienak iritzi *spam*ak diren iritzien detekzioan.

2.1.4.5. Iritzien kalitatearen neurketa

Sentimenduen analisiaren beste ataza bat iritzien kalitatearen neurketa da. Ataza honek iritzi-kritika baten kalitatea, lagungarritasuna, erabilgarritasuna edo baliagarritasuna neurtu nahi du. Lehenik eta behin, iritzi-testuak erauzi egiten dira eta, ondoren, iritzi-testuak sailkatu. Sailkapena egiteko irizpideak bere kalitatea eta bere erabilgarritasuna dira.


JOAQUIN

★★★★★ **Muy buena compra**

12 de julio de 2018

La llevo utilizando ya un tiempo y estoy encantado con ella. Muy buena calidad a ISOs altos muy buen enfoque incluso con objetivo con apertura muy grande de 1. 4 o 2. 8 que con mi anterior cámara EOS 700d de 5 fotos solo una salida enfocada y está la esclava todas. La pantalla táctil y abatible es comodísimo. La batería es lo único que creo que dura poco pero por lo demás es fenomenal.

A 5 personas les ha parecido esto útil

2.3 irudia: Iritziaren kalitatearen neurketa Amazon webgunean.

Gaur egun, helburu horrekin, zenbait webgunek (Amazon barne) kritikaren kalitatea edo erabilgarritasunari buruzko kalifikazioa lortzen dute irakurleei horien inguruan galdetuz. 2.3 irudian esaterako, kritika hori bost pertsonei baliagarria suertatu zaie. Hala ere, iritziaren kalitatea neurtzeak luze jo dezake eta tresna automatikoen beharra dago.

Iritzien kalitatea neurtzeko aipatuko ditugun teknikak 2.10 taulan laburbilduta daude.

Teknika	Lana
SVM erregresioa	Kim *et al.* (2006)
Ospe ezaugarriak + eduki ezaugarriak + ezaugarri sozialak + sentimendu ezaugarriak	Lu *et al.* (2010)
Hurbilpen gainbegiratu gabea	Tsur eta Rappoport (2009)

2.10 taula: Iritzien kalitatea neurtzeko zenbait teknika.

Kim *et al.*ek (2006) iritzien kalitatearen neurketa SVM erregresioa baliatuz egin dute. Horretarako, ezaugarri hauek erabili dituzte: egitura-ezaugarriak

(kritikaren luzera, esaldi-kopurua, galdera- eta harridura-ikurrak, etab.), lexiko ezaugarriak (unigramak eta bigramak), ezaugarri sintaktikoak (hitzen gramatika-kategoria eta horien agerpen-kopurua), ezaugarri semantikoak (produktuaren aspektuak eta sentimendudun hitzak) eta meta-datu ezaugarriak (kritikak dituen izar kopurua).

Lu *et al.*ek (2010), ordea, sailkatze hurbilpenean, kritika lagungarriak eta ez-lagungarriak sailkatzen dituzte. Erabiltzen dituzten ezaugarriak ospe-ezaugarriak, eduki-ezaugarriak, ezaugarri sozialak eta sentimendu-ezaugarriak dira.

Azkenik, Tsur eta Rappoportek (2009) liburu-kritiken erabilgarritasuna hurbilpen gainbegiratu gabea erabiliz neurtzen dute. Lehenik, kritiketako hitzik garrantzitsuenak identifikatzen dituzte eta horiekin bektore bat osatzen dute. Ondoren, kritika bakoitza bektore bilakatzen dute aurretik zehaztutako hitzik garrantzitsuen agerpenean oinarrituz. Azkenik, kritikak sailkatu egiten dituzte.

2.2 Sentimenduen analisirako baliabideen sorkuntza

Atal honetan, sentimenduen analisiko ataza ezberdinetan erabili ohi diren bi baliabideen inguruan jardungo dugu. Alde batetik, iritzi-testuen corpusen inguruan arituko gara (2.2.1 atala). Hauek aniztasun handia erakusten dute. Batzuk elebakarrak dira, besteak eleanitzak. Badaude anotatu gabekoak eta anotatutakoak ere. Eta batzuek testu-genero asko biltzen dituzten bitartean, besteek testu-genero bakarra lantzen dute. Bestetik, berriz, sentimenduen lexikoien inguruan arituko gara (2.2.2 atala). Baliabide hau sortzerakoan, bi hurbilpen nagusi daude: hiztegian oinarritutakoa eta corpusean oinarritutakoa.

2.2.1. Iritzi-testuen corpusak

Hizkuntza naturalaren prozesamenduko sentimenduen analisi-atazarako mota askotako eta helburu ezberdinetarako corpusak sortu izan dira. Leturia *et al.*ek (2014) corpus-mota hauek bereizten ditu.

- **Eremuaren arabera**
  - Orokorrak. Denetariko gaiez osatutakoak dira. Hizkuntzaren erabilera-eremu guztietarako baliagarriak edo "adierazgarriak" izatea dute helburu.
  - Espezializatuak. Hizkuntzaren erabilera-eremu espezializatu bateko testuak biltzen dituzte. Terminologiarako egokiak dira.
- **Erregistroaren arabera**
  - Denetarikoak. Genero edo erregistro guztiak biltzen dituzten corpusak dira.
  - Erregistro jakin batekoak. Testu guztiak genero edo erregistro jakin batekoak dira. Esaterako, kazetaritza eta administrazioko testuez osatutako corpusak badaude.
- **Adierazgarritasunaren arabera**

  - Orekatuak. Unibertsotik ausaz hartutako testuez osatutako corpusak. Unibertso horren ezaugarriak eta banaketa badituzte. Corpusik objektiboenak dira.
  - Ereduzkoak. Corpusaren egileak eredugarritzat jotzen dituen testuez osatuta dago. Helburu preskriptiboa dute horrelako corpusek.
  - Oportunistak. Irizpide bakartzat testuak erraz biltzea duen corpusmota da.

- **Hizkuntza-kopuruaren arabera**
  - Elebakarrak. Hizkuntza bakarrez osatutako corpusak dira. Lexikografian erabiltzen dira, batez ere.
  - Eleaniztunak. Hizkuntza bat baino gehiagoz osatutako corpusak dira. Itzulpen automatiko estatistikoko sistemak entrenatzeko egokiak dira.

- **Corpus eleaniztun motak lerrokatze mailaren arabera**
  - Paraleloak. Hizkuntza bakoitzeko azpicorpuseko testuak elkarren itzulpenak diren corpusei esaten zaie.
  - Konparagarriak. Azpicorpusek ezaugarri amankomunen bat dutenean esaten da. Testuak ez dira elkarren itzulpenak eta ez daude lerrokatuta. Amankomunean dutena generoa edo garaia izan daiteke.

Irizpide horietaz gain, badago beste bat corpusak sailkatzeko erabil daitekeena. Sailkapen hori aberastasunaren araberakoa da, hots, ea corpusa etiketatuta dagoen edo ez.

Jarraian, sentimenduen analisirako sortu diren zenbait corpus deskribatu ditugu. Horretarako, aurretik aipaturiko ezaugarrietan oinarritu gara.

Sentimenduen analisirako sortu diren eta aipatu ditugun corpusak 2.11 taulan laburbilduta daude.

Corpusa	Tamaina	Eremua	Erregistroa	Adierazgarritasuna	Hizkuntzakopurua	Lerrokatzemaila	Aberastasuna
Taboada (2008)	300 testu	Orokorra (6 gai)	Bakarra (iritziak)	Orekatua	Ingelesa		Etiketatu gabea
Refaee eta Rieser (2014)	8.868 txio	-	Bakarra (txioak)	Oportunista	Arabiera		Hizkuntza maila ezberdinetan etiketatuta
Clematide *et al.* (2012)	270 esaldi	Orokorra	Denetarikoak	Orekatua	Alemana		Hizkuntza maila ezberdinetan etiketatuta + sentimendua
Shin *et al.* (2012)	8.050 esaldi		Bakarra (albisteak)		Koreera		MPQA hurbilpenaz etiketatuta
Li *et al.* (2012)	108 dokumentu 714 esaldi	Espezializatua (politika)	Bakarra (albisteak)		Alemana		Hizkuntza-fenomenoak + Sentimendua
Fernández *et al.* (2011)	300 testu 975 esaldi ~90.000 hitz	Orokorra	Bakarra (Web 2.0ko testuak)		Gaztelania Italiera Ingelesa	Konparagarria	Subjektibitate-adierazpenak (sentimendua)
Schulz *et al.* (2010)	750 iruzkin	Espezializatua	Bakarra (kritikak)	Orekatua	Ingelesa Gaztelania Alemana	Konparagarria	
Karoui *et al.* (2017)	38.262 txio		Bakarra (txioak)	Orekatua (ironiko eta ez-ironiko)	Frantsesa Ingelesa Italiera	Konparagarria	Sentimendua etiketatuta
Bond *et al.* (2016)	3.824 esaldi 74.732 hitz	Espezializatua	Bakarra (ipuinak)	Orekatua	Ingelesa Txinera Japoniera	Paraleloa	Kontzeptuak sentimenduarekin etiketatuta

2.11 taula: Sentimenduen analisirako sortu diren zenbait corpus.

Taboadaren (2008) *The SFU Review Corpus* izenekoa lanetako bat da. Corpus honek *Epinions* webguneko iritzi-testuak biltzen ditu, sei gai edo kategorien ingurukoak: liburuak, autoak, ordenagailuak, sukaldeko tresneria, hotelak, filmak, musika eta telefonoa. Kategoria bakoitzak 25 iritzi positibo eta 25 iritzi negatibo ditu. Beraz, guztira 300 iritzi-testuko corpusa da. Bere ezaugarriei dagokienez, corpus orokorra dela esan daiteke, baita erregistro bakarrekoa dela ere, eta adierazgarritasunari dagokionez, orekatua dela. Bestalde, corpusa ingelesez dago eta hainbat hizkuntza mailatan etiketatuta dago: diskurtsoa RST hurbilpena erabiliz, subjektibitatearen ebaluazioa *Appraisal Theory* erabiliz eta, azkenik, ezeztapena eta espekulazioa.

Bestalde, Refaee eta Rieserek (2014) Twitter erabiltzen dute eta arabierazko 8.868 txio biltzen dituzte. Arabieraren dialekto ezberdinak biltzen dituzte ausaz egindako bilaketa batean. Beraz, corpus oportunista da adierazgarritasunaren aldetik eta txioak bakarrik biltzen dituelako erregistro bakarrekoa da. Gainera, corpusa ezaugarri ezberdinez etiketatuta dago, hala nola, morfologia, sintaxia, semantika, estilistika[19] eta seinale sozialak[20].

Clematide *et al.*ek (2012), berriz, corpusa alemanerako sortzen du. Corpusa guztira 270 esaldiz osatuta dago eta *DeWaC Corpus* izeneko corpusean oinarritzen da. Corpus honek webguneetako hainbat generotako testuak biltzen ditu. Corpusa geruza ezberdinez osatuta dago eta horietako bakoitza etiketatuta. Lehena esaldi mailako geruza da eta esaldi bakoitzaren objektibitatea, subjektibitatea eta polaritatea zehazten da. Bigarren geruza hitz eta sintagma mailakoa da, eta informazio subjektiboa eta faktuala dago etiketatuta. Azken geruza adierazpen maila da. Hizketa-gertaerak objektiboak edo zuzenekoak diren zehazten du. Beste modu batera esanda, egoera pribatuko ingurunea zehazten da. Beraz, corpus hau eremuari dagokionez orokorra dela esan daiteke, eta erregistroari dagokionez orekatua eta etiketatua.

[19]Hizkuntzalaritzan, estilistika adierazkortasuna eta literaturan estiloa eragiten duten hizkuntza erabilerak aztertzen dituen arloa da. Erretorikazko figurak eta sintaxizko egiturak lantzen dituzte.

[20]Seinale sozialak gizarte gertakariei buruzko informazioa ematen duten seinale komunikatiboak edo informatiboak dira. Hainbat motariko informazioa eman dezakete: gizarte elkarrekintzak, gizarte-emozioak, gizarte-ebaluazioak, gizarte-jarrerak edota gizarte harremanak, besteak beste.

Li *et al.*ek (2012) albisteez osaturiko sentimenduen analisirako beste corpus bat aurkezten dute. Kasu honetan, alemanezko albiste politikoak bildu eta etiketatu dituzte. 108 dokumentuz eta 714 esaldiz osatuta dago corpusa. Guztira jarrerarekin lotura duten lau elementu anotatu dituzte. Horietako bat testu-aingura da eta jarrera adierazten duten testu-zatiak dira hitz, sintagma edota lokuzioak. Bigarren elementua helburua da, hots, jarreraren helburua edo jomuga. Anotatu beharreko hirugarren elementua jarreraren iturria da, hau da, iritzia adierazten duen pertsona. Azken elementua, berriz, laguntzailea deiturikoa da; sentimenduan eragin dezakeen hitza: ezeztapena, intentsifikatzailea edota ahultzailea.

Shin *et al.*ek (2012) koreerazko sentimenduen corpusa aurkezten dute. Corpusak albiste-artikuluak biltzen ditu eta guztira 8.050 esaldiek osatzen dute. Corpusa persperktiba anitzeko galde-erantzunen (*Multiperspective Question Answering*, MPQA) hurbilpenean oinarrituz etiketatuta dago. MPQAk esaldiaren orientazioa positiboa edo negatiboa den zehazteko aukera ematen du.

Orain arte, sentimenduen analisi-atazarako sortu diren corpus ezberdinak deskribatu ditugu. Ikusten denez, corpusek ezaugarri ezberdinak dituzte, bai bildutako testu edo esaldiei dagokienez, bai etiketatzeko moldeari dagokionez. Baina denek dute ezaugarri bat amankomunean eta hori corpusa hizkuntza batez osatuta dagoela da. Hala ere, badira zenbait corpus hizkuntza askoz osatuta daudenak, jarraian datozenak bezalakoak.

Fernández *et al.*ek (2011) aurrekoekiko zertxobait ezberdina den corpusa aurkezten dute. *EmotiBlog* izeneko corpusak Web 2.0 web-zerbitzuaren bidez sortutako testu-genero berriak biltzen ditu. Bertako adierazpen subjektiboak etiketatuta ditu corpusak. Corpusak 975 esaldi ditu. Horietatik 519 esaldi objektiboak dira eta gainerakoak, 456 esaldi subjektiboak. Hauen artean, 260 esaldi positiboak dira eta 188 esaldi, aldiz, negatiboak. Hiru gai lantzen dituzte testuek: Kioto protokoloa, Zimbabweko Mugaberen gobernua eta AEBetako 2008ko hauteskunde presidentzialak. Gai bakoitzeko 100 testu bildu dituzte. Bestalde, corpusa eleanitza da, gaztelaniazko, italierazko eta ingelesezko testuak baitaude. Hizkuntza bakoitzak 30.000 hitz inguru ditu.

Schulz *et al.*ek (2010) hizkuntza anitzeko corpus bat aurkezten dute. Corpus honek ingeleseko, gaztelaniako eta alemaneko testuak biltzen ditu eta

eskuz etiketatuta daude. Hu eta Liuen (2004) corpusean, ingelesez dagoena, oinarritzen dira corpusa sortzeko eta produktu berberen alemanezko eta gaztelaniazko beste 500 iruzkin gehitzen dituzte. Iritzien ezaugarri hauek etiketatu dira: i) produktuen ezaugarriak, ii) horien orientazio semantikoa eta iii) horien indarra (0-3 artekoa). Kasu honetan, corpus eleanitza lerrokatze mailari dagokionez konparagarria da; izan ere, azpicorpusek amankomunean dutena produktu berberen iritziak izatea da.

Karoui *et al.*ek (2017) ere hiru hizkuntzaz osaturiko corpusa sortu dute baina testu-mota ez da bera, txioz osaturiko corpus hirueleduna sortu baitute. Gainera, corpusa helburu batekin sortu dute eta helburua ironia detektatzeko anotazio-irizpide bat finkatzea da. Corpuseko hiru hizkuntzak frantsesa, ingelesa eta italiera dira. Frantseseko azpicorpusean, 2.073 txio ironiko eta 16.179 txio ez-ironiko daude. Ingeleseko azpicorpusean, 5.173 txio ironiko eta 6.116 txio ez-ironiko daude. Azkenik, italierakoan, 3.079 txio ironiko eta 5.642 txio ez-ironiko daude. Ikusten den moduan; beraz, corpusa ez da paraleloa, txioak ez direlako itzulpenak; baina konparagarria bada, ezaugarri amankomuna ironia edukitzea edo ez edukitzea delako. Geruza askotariko anotazio-irizpidea baliatuta etiketatu dute corpusa: ironia (esplizitua edo inplizitua), ironia-kategoriak eta ironia-markatzaileak.

Azkenik, Bond *et al.*ek (2016) ere beste corpus eleanitz bat osatu dute. Corpus hau ingelesezko bi istorio laburrez osatuta dago. Aurretik aipatutako beste bi corpusekin corpus honek duen ezberdintasuna horixe da; corpus eleanitza itzulpen bidez lortu denez, corpusa paraleloa dela. Anotazioa bi mailatan egin dute. Kontzeptu mailan, sentimendu positiboa edo negatiboa duten hitzak (kontzeptuak) etiketatu dituzte. *Chunk* mailan, berriz, sentimendu bera duten segidako hitz-multzoak etiketatu dituzte. Ingeleseko testuak originalak dira eta japonierara eta txinerara itzuli dituzte.

Gure iritzi-corpusak lotura gehien Shin *et al.*ekin (2012), Li *et al.*ekin (2012) eta Taboadarekin (2008) du. Taboadak (2008) bezala, gure corpusa orekatua da eta Shin *et al.* (2012) eta Li *et al.* (2012) lanek bezala, albisteak edo kazetaritzako testuak biltzen ditu gure corpusak.

2.2.2. Sentimenduen lexikoia

Sentimendua duten hitzei hainbat izen hartu izan dituzte: sentimendu-hitzak (*sentiment words*), iritzi-hitzak (*opinion words*) edota polaritate-hitzak (*polar words*), besteak beste. Liuk (2012) dioenez, sentimendu positiboko hitzek egoera edo kualitate desiratuak adierazten dituzte: *ederra*, *bikaina* edota *txundigarria*. Sentimendu negatiboko hitzek, aldiz, egoera edo kualitate ezdesiratuak adierazten dituzte: *txarra*, *beldurgarria* edota *pobrea*. Hitz horiek guztiek osatzen duten multzoari sentimenduen lexikoia esango diogu.

Sentimendu-hitzak bi motarikoak izan daitezke: oinarrizkoak (*base type*) eta konparaziozkoak (*comparative type*). Azken honetan, konparatiboak edo superlatiboak diren hitzak biltzen dira: *hobea*, *okerragoa*, *onena* eta *okerrena*, esaterako.

Liuren (2012) lanean sentimendu-hitzak biltzeko bi modu agertzen dira:

1- Hiztegian oinarritutako hurbilpena.

2- Corpusean oinarritutako hurbilpena.

Jarraian, bi hurbilpen horiek deskribatuko ditugu, tesi-lan honetan sortu dugun *Sentitegi*, euskarazko sentimenduen lexikoia, bi hurbilpen horietaz elikatu delako.

2.2.2.1. Hiztegian oinarritutako sentimenduen lexikoia

Sentimenduen lexikoia sortzeko hiztegia erabiltzea (Millerren (1995) *Wordnet* izeneko datu-base lexikala kasu) nahiko hurbilpen ohikoa izan da; izan ere, hiztegiek sinonimoak eta antonimoak biltzen dituzte eta horiek baliagarriak dira lexikoia sortzeko. Beraz, teknika sinpleena honako hau da: sentimendua duten hitz multzo txiki bat hartu eta, hiztegiaren sinonimo eta antonimo egituran oinarrituta hiztegiko egituran barrena mugitzea *boostrapping* teknika[21] erabiliz sentimendua duten hitz gehiago bilatzeko.

[21]Estatistikan, *boostrapping* berlanginketa egiteko metodo bat da eta laginen banaketa egiteko erabiltzen da. Laginen datu multzo bat baliatuz, populazio baten gainean inferentzia egitean; ondoren, inferentzia hori laginen datuen berlaginketa baten bidez modelatzean eta, azkenik, berlangiketako datuekin inferentzia berri bat egitean datza.

Wordnet erabilita lexikoi bat Hu eta Liuren (2004) arabera modu honetan sor daiteke:

1- Lehenik eta behin, eskuz, hazi-hitz positibo edota negatiboak bildu behar dira. Hazi-hitzak datu-base lexikaletik hartutako hasierako zenbait hitz dira. Ondoren, hitz horietan oinarrituta bilaketa gehiago egiten dira datu-basean hitz gehiago aurkitzeko. Beigman Klebanov *et al.*ek (2012), esaterako, *anxiety*, *conflict* eta *improve* hitzak hazi-hitz moduan hartzen dituzte.

2- Ondoren, algoritmo batek hitz horien sinonimo eta antonimoak bilatzen ditu bai *Wordnet*en, bai sareko beste hiztegietan. Aurkitutako hitzak zerrendara gehituko dira. Beigman Klebanov *et al.*en (2012) lanean, algoritmoaren emaitzak dira bilatutako *anxiousness* (<*anxiety*), *battle* (<*conflict*) eta *amend* (<*improve*) hitzak.

3- Azkenik, iterazioa behin eta berriz hasten da hitza bilatzen. Ezin direnean hitz gehiago bilatu, iterazioa amaitzen da.

Hiztegian oinarrituta lexikoiak sortzea erraza eta azkarra da. Modu honetan sortutako lexikoiek akats batzuk eduki ditzakete, baina gero, eskuz konpon litezke. Modu honen desabantaila da sortutako lexikoiak oso loturik daudela domeinuari nahiz testuinguruari. Ondorioz, ez dago domeinu askotarako balio duen sentimenduen lexikoirik.

Lan hauek guk sortu dugun sentimendu lexikoiarekin bi ikuspegitatik dute lotura. Alde batetik, denak dira hiztegian oinarrituta sortuak. Beste aldetik, berriz, hitzen sentimenduen balentzia kalkulatu da hitzek elkarren artean duten lotura semantikoa aintzat hartuta.

Sentimenduen lexikoia sortzeko hiztegia erabiltzen duten eta aipatuko ditugun lanak 2.12 taulan laburbilduta daude.

Blair-Goldensohn *et al.*ek (2008) *bootstrapping* teknika konplexuagoa erabiltzen dute. Izan ere, lehenik eta behin hazi-hitz positibo, negatibo eta neutralak biltzen ditu. Ondoren, hitz horiei pisua esleitzen diete grafo semantikoa

Teknika	Lana
Wordnet + Bootstrapping	Hu eta Liu (2004) Blair-Goldensohn *et al.* (2008)
Grafoan oinarritutako ikasketa gainbegiratu gabea	Rao eta Ravichandran (2009)
Markoven ausazko ibilaldia (*Markov random walk*)	Hassan *et al.* (2014)
Elkarrekiko informazio puntuala (*Pointwise Mutual Information*, PMI)	Turney eta Littman (2003) Taboada *et al.* (2011)

2.12 taula: Hiztegian oinarrituz sentimendu lexikoia sortzeko zenbait teknika.

erabiliz. Grafo[22] semantiko horretan, ondoko adabegiak (*neighboring nodes*) *Wordnet*eko hitzen sinonimoak eta antonimoak dira eta ez daude hazi-hitz neutralen multzoan. Hazi-hitz neutralak sentimendu orientazioa hitz neutraletara ez zabaltzeko erabiltzen dira.

Rao eta Ravichandran (2009) lanak, berriz, hitzen polaritatea antzematea du helburu eta, horretarako, grafoan oinarritutako ikasketa erdi-gainbegiratua proposatzen dute. Beren ekarpena da sinonimia eta hiperonimia moduko erlazioak erabiltzea etiketen zabalkundearen emaitzak hobetzeko. Lanaren oinarrian Wordnet datu-base lexikala dago.

Hassan *et al.*en (2014) lanak, hitzen orientazio semantikoa kalkulatzeko asmoz, Markoven ausazko ibilaldia (*Markov random walk*) eredua erabiltzen du. Beren metodoak *Wordnet* datu-basean hedapena orientazio semantiko bera duten hitzen bidez egiten du lehenik eta, ondoren, hedapena orientazio semantiko ezberdineko hitzetara igarotzen da.

Bestalde, Turney eta Littmanen (2003) lanak elkarrekiko informazio puntuala (*Pointwise Mutual Information*, PMI) neurrian oinarritutako metodoa erabiltzen du. Hitz baten orientazio semantikoa kalkulatzeko, hitz batek hazi-hitz positiboekiko (*ona*, *bikaina*, *zuzena*) duen asoziazioaren indarra ken hazi-hitz negatiboekiko (*txarra*, *pobrea*, *okerra*) duen asoziazio semantikoa-

[22]Grafoa datu-egitura mota bat eta, bertan, erlazioak edo loturak adierazten dira.

ren indarra[23] kalkulatuz zenbatzen du. Elkarrekiko informazioa puntuala erabiliz neurtzen du hitz batek duen asoziazioaren indarra. Taboada *et al.*en (2011) SO-CAL tresnako lexikoia ere modu berean sortu da.

2.2.2.2. Corpusean oinarritutako sentimenduen lexikoia

Liuk (2012) azaltzen duenez, corpusa erabiliz sortutako sentimenduen lexikoiak bi arrazoi jakinengatik sortzen dira.

1- Ezagunak diren sentimenduen hitzak edukita, domeinu bakarreko corpusean sentimendua duten hitzak eta horien orientazio semantikoa ezagutzeko. Hau da, aldez aurretik lortutako sentimendu-hitzak domeinu bakar bateko corpusean aplikatzen dira eta corpusetik sentimendu-hitzak lortzen dira.

2- Helburu orokor baterako edo domeinu askotarako sortutako sentimenduen lexikoia beste domeinu berri batera moldatzeko, domeinu bakarreko corpusean oinarrituz. Kasu honetan, aurretik badagoen lexikoi bat domeinu batera murrizten da corpus baliatuz.

Bi arrazoiek antzekoak diruditen arren, badaude ezberdintasun batzuk. Lehen kasuan, sentimendu-hitzak corpusetik lortu dira. Bigarrengoan, aldiz, aldez aurretik daude sentimendu-hitzak, baina corpusaren iragazkia pasatu dute.

Corpusean oinarritutakoetan helburua lexikoia domeinu batera mugatzea bakarrik dela dirudien arren, hori egitea ez da uste bezain erraza. Izan ere, domeinuaren arabera, hitz batek bi aurkako esanahi izan ditzake eta hori tratatzea zaila da.

Corpusean oinarritutako lexikoiak erabili izan dira helburu orokorreko sentimenduen lexikoiak sortzeko, beti ere corpusa oso handia eta anitza bada. Hala ere, helburu horretarako hiztegian oinarritutako lexikoiak egokiagoak dira.

[23] Asoziazioa estatistikan erabiltzen den kontzeptua da eta bi aldagaien arteko erlazioa edo lotura adierazten du. Sentimenduen analisira mugatuz, bi hitzek elkarren artean duten lotura semantikoa adierazten du.

Corpusean oinarrituta sentimenduen lexikoiak sortzen dituzten lanak 2.13 taulan laburbilduta daude.

Teknika	**Lana**
Esaldi barnea: juntagailuak	Hatzivassiloglou eta McKeown (1997)
Esaldi barnea: juntagailuak Esaldi artean: aurkaritzako juntagailuak	Kanayama eta Nasukawa (2006)
Aspektua (iritziaren testuingurua)	Ding *et al.* (2008)
i) Elkarrekiko errefortzua ii) Domeinuen arteko zati amankomuna	Du *et al.* (2010)
Antzekotasun distribuzionala	Wiebe eta Mihalcea (2006)

2.13 taula: Corpusean oinarrituta sortutako zenbait sentimenduen lexikoi.

Corpusean oinarritutako hurbilpenak guk sortu dugun sentimenduen lexikoiarekin duen lotura da biak direla hein batean corpus batetik sortuak edo corpus batean oinarrituak. Gure kasuan, lexikoi bat gaztelaniatik euskaratu ostean, lexikoia corpuseko domeinuetara egokitu dugu. Hau da, 2.2.2.2 atalaren hasieran aipatu dugun moduan, helburu orokor baterako edo domeinu askotarako sortutako sentimenduen lexikoia beste domeinu berri batera moldatu dugu, domeinu bakarreko corpusean oinarrituz.

Positiboa	**Negatiboa**
adequate, central, clever, famous, intelligent, remarkable, reputed, sensitive, slender, thriving	contagious, drunken, ignorant, lanky, listless, primitive, strident, troublesome, unresolved, unsuspecting

2.14 taula: Hatzivassiloglou eta McKeownen (1997) laneko hazi-hitzak.

Hatzivassiloglou eta McKeownena (1997) da hurbilpen honetan lehenengoetariko lana. Bertan, corpusa eta hazi-hitzak (2.14 taula) erabili dira corpusean sentimendua duten adjektibo gehiago bilatzeko. Lehenik juntagailuz lotuak dauden adjektiboak lortzen dituzte zenbait erregela linguistiko erabiliz: *eta, edo, baina, edo... edo...* eta *ez... ez...* izan dira. Ondoren, adjektibo bakoitzaren orientazio semantikoa lortu eta orientazioaren arabe-

Positibo bezala sailkatua	Negatibo bezala sailkatua
bold, decisive, disturbing, generous, good, honest, important, large, mature, patient, peaceful, positive, proud, sound, stimulating, straightforward, strange, talented, vigorous, witty	ambiguous, cautious, cynical, evasive, harmful, hypocritical, inefficient, insecure, irrational, irresponsible, minor, outspoken, pleasant, reckless, risky, selfish, tedious, unsupported, vulnerable, wasteful

2.15 taula: Hatzivassiloglou eta McKeownen (1997) sentimenduen lexikoia zati bat.

ra multzokatzen dituzte. Azkenik, multzo bakoitzaren maiztasunak alderatu eta maiztasun handienekoak positibo moduan etiketatzen dira (2.15 taula).

(8) Koordinaziozkoa ("eta" esanahia gutxi gorabehera): *-te, -shi, -weni, -dakedenaku, - nominarazu.*

(9) Kausazkoa ("-(e)lako" esanahia gutxi gorabehera): *-tame, -kara, -node.*

(10) Aurkaritzakoa ("baina" esanahia gutxi gorabehera): *-ga, -kedo, -keredo, -monomo, -nodaga.*

(11) *shikashi* ("hala ere"), *demo* ("baina"), *sorenanomi* ("nahiz eta"), *tadashi* ("baldin eta"), *dakedo* ("baina"), *gyakuni* ("bestela"), *tohaie* ("arren"), *keredomo* ("hala ere"), *ippou* ("bestalde").

Lan horretatik abiatuz, Kanayama eta Nasukawaren (2006) lanak zenbait aurrerapen egin dituzte. Aurreko kontzeptua garatu dute. Esaldi barneko (*intra-sentential*) eta esaldi arteko (*inter-sentential*) kontzeptuak sortu dituzte. Lehena Hatzivassiloglou eta McKeown (1997) laneko kontzeptu bera da eta esaldi barnean kokatzen denez, erregela linguistikoak ere (*eta, edo...*) esaldi barnekoak dira ((8), (9) eta (10) adibideak). Esaldi artekoan, aldiz, jarraian dauden bi esaldiek ideia edo kontzeptu bera adierazten dute. Kasu honetan, bestelako erregela linguistikoak sortu behar izan dituzte: *baina, hala ere* ((11) adibidea). Esaldi artekoa izendatzeko sentimendu sendotasuna (*sentiment consistency*) kontzeptua erabili dute.

Ding *et al.*ek (2008) diotenez, goiko metodologia domenuari loturiko hitzak bilatzeko erabilgarria den arren, praktikan hori bakarrik ez da nahikoa. Izan ere, hitz batek aurkako bi orientazio semantiko izan ditzake bi testuinguru ezberdinetan. Esaterako, *bateriaren bizia luzea da* eta *fokuratzeak denbora luzea eskatzen du* esaldietan, *luzea* hitzak lehenengoan orientazio semantiko positiboa du eta bigarrengoan, berriz, negatiboa. Horregatik, domeinuari loturiko sentimenduen hitza eta haren balentzia ez direla nahikoa ohartzen dira eta horren aurrean, aspektua eta sentimendua duen hitzari garrantzi gehiago eman eta iritzi-testuingurua irudikatzeko bikoteak (aspektua, sentimendu_hitza egitura dutenak) proposatzen dituzte. Adibidez, goiko adibidearen bikotea (*bateriaren bizia*, *luzea*) izango litzateke.

Aurreko lanak ezagunak diren hitzetatik abiatuz domeinu bateko corpusean sentimendua duten hitz gehiago bilatu eta horiei orientazio semantikoa esleitzeko hurbilpen eta teknikak dira. Jarraian, berriz, jada sortuta dagoen sentimenduen lexikoi orokor bat domeinu zehatz batera moldatzeko teknikak eta hurbilpenak aipatuko ditugu.

Du *et al.* (2010) lanean domeinu bateko lexikoia beste domeinu batera egokitzen dute. Horretarako, domeinu barruko dokumentu etiketatutak, domeinu barruko sentimendu-hitzak eta domeinuz kanpoko dokumentuak erabiltzen ditu algoritmoak. Bi aspektu lantzen dituzte: dokumentu bat positiboa bada, bertan dauden hitz askok ere orientazio semantiko positiboa dute eta dokumentua negatiboa bada, bertako hitz askok orientazio semantiko negatiboa dute (honek elkarrekiko errefortzuarekin, *mutual reinforcement*ekin, du lotura) eta nahiz eta bi domeinuek distribuzio ezberdina izan, beren arteko zati amankomuna identifikatzea posible da.

Wiebe eta Mihalceak (2006) corpusean oinarrituta hitz adierei subjektibitate etiketak esleitzea proposatzen dute. Bi azterketa egiten dituzte. Lehenengoan, *Wordnet*eko adierei bi pertsonek "subjektiboa", "objektiboa", "biak" etiketak esleitzeko duten adostasuna neurtzen dute. Bigarrengoan, hitzen adierei etiketak automatikoki esleitzen dizkie antzekotasun distribuzionala metodoan[24] oinarrituz.

[24]Antzekotasun distribuzionala (ingelesez, *distributional similarity* teknika bat da. Teknika honen arabera, antzekoak diren objektu linguistikoek antzeko edukia (dokumentuetan

Gure sentimenduen lexikoia aurreko bi hurbilpenen uztarketa da. Alde batetik, hiztegian oinarritutakoa da, SO-CAL tresnaren gaztelaniazko lexikoian (Brooke *et al.*, 2009) oinarrituz egin baita. Beste aldetik, berriz, corpusean oinarritutakoa ere bada, itzulitako lexikoia corpuseko sei domeinuetara egokitu baitugu eta anbiguoak diren hitzetan, corpuseko adiera eta, horrekin batera, sentimenduen balentzia hartu baititugu aintzat.

eta esaldietan, esaterako) eta antzeko testuingurua (hitzen kasuan) duten.

2.3 Balentzia-aldatzaileak

Behin lexikoiak nola lortzen diren aipatuta, testuinguruko balentzia-aldatzaileek horien balioetan nola eragiten duten azalduko dugu. Testuinguruzko balentzia-aldatzaileen lanketan erreferentziazko lana Polanyi eta Zaenenena (2006) da. Lan horretan, aurretik egin diren zenbait lan osatu gabe daudela adierazten dute eta beren analisia proposatzen dute, jarraian azalduko dugun moduan.

Polanyi eta Zaenenek (2006) testuinguruko balentzia-aldatzaileak (ingelesez, *contextual valence shifters*) proposatzen dituzte. Beren arabera, testuak gertaerak eta egitateak deskribatzeaz gain, deskribatzen den gertaerarekiko idazleak edo parte-hartzaileek duten jarrera komunikatzen du. Horretarako, idazleak aukeraketa lexikala egiten du, baina testuaren antolakuntzak berak ere idazlearen jarrera adierazteko informazio kritikoa ematen du. Beste modu batera esanda, nahiz eta idazleek aukeratutako hitzen sentimendu balentzia jakin bat izan, testuaren antolakuntzak eta bertako elementuek hitzen sentimendu balentzia alda dezakete. Atal honetan aipatuko diren lanak 2.16 taulan laburbilduta daude.

2.3.1. Testuingurua kontuan hartzen ez duten lanak

Atal honetan, hitzen testuingurua kontuan hartzen ez dutelako osatugabetzat jo dituzten zenbait lan aipatu eta beren analisia deskribatuko dugu.

Edmonds eta Hirstek (2002) sinonimo hurkoak eta aukeraketa lexikala erabiltzen dituzte. Zehatzago esanda, sinonimo hurkoen esanahiak eta beren arteko ezberdintasunak errepresentatzeko modeloa aurkezten dute, baita egoera jakin batean sinonimo hurko horien artean aukeraketa lexikal egokiena egiteko prozesua ere. Sinonimo hurkoak ia sinonimoak dira, baina ez guztiz. Esanahiaren aldetik oso antzekoak dira, baina ez dira berdinak eta ezin dira elkartrukatu. Denotazioan, konnotazioan, inplikazioan, enfasian edo erregistroan ezberdintzen dira. 2.17 taulan, zenbait sinonimo hurko ageri dira beren arteko ezberdintasuna zer den aipatuz.

Errepresentazioaren granularitatea erabiliz[25] sinonimo hurkoak multzokatu

[25]Errepresentazioaren granularitatean (ingelesez, *representation granularity*), hitz ba-

Hitzen orientazio semantikoa testuingurua kontuan hartu gabe	**Lana**
Sinonimo hurkoak Denotazioa, konnotazioa, inplikazioa, enfasia, erregistroa	(Edmonds eta Hirst, 2002)
Juntagailuz elkartuta dauden adjektiboak	(Hatzivassiloglou eta McKeown, 1997)
Web bilaketa motorraren kontsulta Elkarrekiko informazio puntuala	(Turney eta Littman, 2002)
Maiztasun gutxiko hitzak, kolokazioa, adjektiboak, aditzak	(Wiebe *et al.*, 2004)
Berrikuntza: **testuinguruko balentzia-aldatzaileak**	
Esaldi mailako testuinguruko balentzia-aldatzaileak Diskurtso mailako testuinguruko balentzia-aldatzaileak	(Polanyi eta Zaenen, 2006)
Hitzen orientazio semantikoa testuingurua kontuan hartuta	
General Inquirer lexikoia, web corpusa, Asoziazio kalifikazioa	(Kennedy eta Inkpen, 2006)
SVM Balentzia-aldatzaileen unigramak eta bigramak	(Kennedy eta Inkpen, 2006)
Ezaugarrien konbinaketa Testuinguruko balentzia-aldatzaileak	(Morsy eta Rafea, 2012)
Termino kontaketa Testuinguruko balentzia-aldatzaileak	(Ngoc Phu eta Thi Tuoi, 2014)

2.16 taula: Hitzen orientazio semantikoa esleitzeko testuingurua kontuan hartzen eta hartzen ez duten lanak.

Bariazio mota	**Adibidea**
Abstrakzio dimentsioa	*seep : drip*
Enfasia	*enemy : foe*
Denotaziozkoa, zeharkakoa	*error : mistake*
Denotaziozkoa, lausoa	*woods : forest*
Estilistikoa, formaltasuna	*pissed : drunk : inebriated*
Estilistikoa, indarra	*ruin : annihilate*
Jarrera adierazpena	*skinny : thin : slim, slender*
Emotiboa	*daddy : dad : father*
Kolokaziozkoa	*task : job*
Aukeraketazkoa	*pass away : die*
Azpikategorizazioa	*give : donate*

2.17 taula: Sinonimo hurkoen sailkapena beren ezberdintasunetan oinarrituta (Edmonds eta Hirst, 2002, 5. orr.).

tean esanahia sortzeko; alde batetik, testuinguruz independienteak diren esanahiek dituz-

egiten dituzte. Ondoren, ezagutza lexikalaren eredu multzokatua garatzen dute ohiko eredu ontologikoetan oinarrituz. Multzo bakoitzean azpikontzeptuak daude eta sinonimo hurkoak adierazten dituzten estiloaren, aldeko edo aurkako jarreraren, inplikazioaren eta denotazioaren arabera daude multzokatuta. Prozesua bi mailetan aritzen da aukeraketa lexikala egiteko: multzoetan eta sinonimo hurkoen multzoetan.

Hatzivassiloglou eta McKeownek (1997) adjektiboen orientazio semantikoa iragartzeko lana aurkezten dute. Horretarako juntagailuz elkartuta dauden adjektiboak baliatzen dituzte, (12), (13) eta (14) adibideetan bezalakoak.

(12) The tax proposal was {simple and well-received} by the public.

(13) The tax proposal was {simplistic but well-received} by the public.

(14) The tax proposal was {*simplistic and well-received} by the public.[26]

Erregresio linealean, (14) adibideko murriztapenak bezalakoak baliatuz (*eta* juntagailuarekin batera ezin direla orientazio semantiko negatiboko eta positiboko bi hitz agertu), adjektiboen orientazioa semantikoa iragartzea lortzen dute.

Turney eta Littmanek (2002) adjektiboen, adberbioen, izenen eta aditzen orientazio semantikoa iragartzen dute baina, horretarako, beste hurbilpen bat (metodo gainbegiratua) darabilte. Oinarritzat ehun bilioi hitzeko corpus bat, web-bilaketa motorraren kontsulta eta elkarrekiko informazio puntuala darabiltzate.

Wiebe *et al.*ek (2004) hizkuntza-subjektiboa lantzen dute, hots, iritziak, ebaluazioak edo usteak adierazteko hizkuntzaren aspektuak. Corpus batzuetatik subjektibitatearen arrastoak sortzen eta probatzen dituzte. Arrasto horiek maiztasun gutxiko hitzak, kolokazioak eta antzekotasun distribuzionala baliatuz lortutako adjektiboak eta aditzak dira. Halaber, lan horrek aipaturiko arrastoak dentsitate handi batez edo askotan hitz baten ondoan agertzeak, hitz hori subjektibo bihurtzen duela ere adierazten dute.

ten testuinguruz dependienteak diren konbinazioak eta, beste aldetik, sinonimo hurkoen ezberdintasun esplizituen multzo bat baliatzen dira.

[26]Izartxoak (*) esaldia ez dela zuzena esan nahi du.

Laburbilduz, aipaturiko lanetan, esaldien edo testuen orientazio semantikoa edo sentimendu-balentzia kalkulatzeko, hitz jakin batzuen sentimendu-balentzia erabiltzen dute. Baina, hitz horien inguruan dauden eta hitzaren sentimendu-balentzia aldatu diezaioketen hizkuntza-fenomenoak ez dira kontuan hartzen eta, horregatik, lan osatugabetzat jo izan dira.

2.3.2. Berrikuntza: testuinguruko balentzia-aldatzaileak

Polanyi eta Zaenenen (2006) arabera, aurretik aipaturiko lanek hutsune bat dute eta hutsune hori orientazio semantikoa duten hitzek ondoan duten testuingurua kontuan ez hartzea da. Izan ere, Edmonds eta Hirsten (2002) kasuan, sinonimoen arteko ezberdintasuna hitz horien zenbait ezaugarri zehatzetan oinarrituz egiten da. Hitzetik kanpo dauden ezaugarriak, hots, hitzaren testuinguruari loturikoak, ez dira aintzat hartzen. Hatzivassiloglou eta McKeownen (1997) kasuan ere, corpusetik juntagailuz loturik dauden adjektiboak ateratzen dituzte, baina ez da haien testuingurua kontuan hartzen. Turney eta Littmanen (2002) eta Wiebe *et al.*en (2004) lanekin ere berdin gertatzen da, testuinguruarekin loturarik ez duten ezaugarriak bakarrik erabiltzen dira hitzaren orientazio semantikoa zehazteko.

Gure lanari dagokionez, lexikoian oinarritzen den dokumentu mailako sentimenduen sailkatzailea garatzeko bidean, jakitun gara lexikoiko sarrerek duten sentimenduen balentzia bere horretan aintzat hartuta ez dela nahikoa. Izan ere, Polanyi eta Zaenenek (2006) dioten moduan, hitz baten sentimenduen balentzia jakiteko beharrezkoa da bere testuingurua ere kontuan hartzea. Hori horrela izanik, gure lanaren eta beren lanaren arteko lotura da bietan esaldi eta diskurtso mailako balentzia-aldatzaileak identifikatzea dutela helburu. Hori dela eta, gure lanaren aurrekaria da jarraian deskribatuko dugun hurbilpena.

Osagai lexikalek jarrera positiboa edo negatiboa adierazten dute. Edozein klase irekik adieraz dezake jarrera positiboa edo negatiboa: izenak (*onarpen/katastrofe*), adjektiboak (*erraz/zail*), aditzak (*arindu/huts egin*) eta adberbioak (*azkar/mantso*) daude eta hitz anitzeko unitate lexikalak (*porrot egin*) ere izan daitezke. Horren erakusle dira (15), (16), (17) eta (18) adibideak.

(15) 25 urteko bat bizi zen hiriko partean oinez ibili zen[27].

(16) [Gizon gazte bat]$_{+}$ [bere auzoan zehar]]$_{+}$ [patxadaz paseatu]]$_{+}$ zuen.

(17) [Gazte ar bat]$_{-}$ [bere lurraldean zehar]$_{-}$ [harrokeriaz ibili]$_{-}$ zen.

(18) Filma [interesgarria]$_{+}$ eta [hunkigarria]$_{+}$ izan arren, ez zitzaidan [gustatu]$_{-}$.

(15), (16) eta (17) adibideek sentimendu-hitzen garrantzia erakusten dute. Hirurek gertaera bera azaltzen dute baina egin den aukeraketa lexikalak[28], hau da, aukeratutako sentimenduen balentziek ikuspegi guztiz kontrajarriak adierazten dituzte. Izan ere, (15) adibideak jarrera neutrala du, (16) adibideak positiboa eta (17) adibideak, aldiz, jarrera negatiboa. Gainera, jarrera jakin bateko hitz asko egoteak ez du esan nahi esaldiak eta testuak jarrera hori adierazten dutenik. Esaterako, (18) adibidean, jarrera positiboko hitz gehiago daude, baina esaldiak jarrera negatiboa du. Halaber, kontuan hartu behar da sentimenduen hitzek duten intentsitatea; izan ere, *ona* eta *bikain* ez baitira semantikoki berdinak. Baina Polanyi eta Zaenenek (2006) diotenez, sentimendu-hitzen ezaugarri horiek bakarrik kontuan hartzea ez da nahikoa, testuinguruko balentzia-aldatzaileek ere hiztunek bere jarrera adierazteko momentuan zerbait edo bere aurkako zerbait esatea eragin dezaketelako.

Esaldi mailan hainbat faktore (ezeztapena, aditzen tempusa, etab.) aipatzen dituzte sentimenduen balentzia alda dezaketenak.

(19) Luar [azkarra]$_{+2}$ da. Luar [ez da azkarra]$_{-2}$.

(20) Gertaera [susmagarria]$_{+2}$ da. Gertaera [oso susmagarria]$_{+3}$ da.

(21) Soluzioak [eraginkorra]$_{+2} \rightarrow_{0}$ izan beharko luke.

(22) [Ana pertsona txarra da]$_{-1}$. [Bere lagunak gaizki tratatzen ditu]$_{-1}$. [Ana pertsona txarra izango balitz, bere lagunak gaizki tratatuko lituzke]$_{0}$.

[27](15) adibide hau nahiz beste hauek (16, 17, 20, 24, 25, 26, 28 eta 30) Polanyi eta Zaenenetik (2006) hartuta eta euskarara itzulita daude.

[28]Psikolinguistikan, aukeraketa lexikala (ingelesez, *lexical selection*) hizkuntza ekoizterakoan, mezua identifikatu ondoren, erantzuteko mezua irudikatzeko hiztunak aukeratzen dituen item lexikalak dira eta, prozesu horretan, hitzari buruzko informazio semantikoa eta gramatikala aktibatzen da.

(23) [Azterketa gainditu du.]$_{+}$ [Azterketa nekez gainditu du]$_{-}$.

(24) [Antolakuntza [oso [distiratsuak]$_{+2}$]$_{+3}$ $\rightarrow_{-3}$ [huts egin]$_{-1}$ zuen [arazoa konpontzen]$_{+1}$ $\rightarrow_{0}$]$_{+5}$ $\rightarrow_{-4}$.

(19) adibidean, ezeztapenak *azkarra* adjektiboaren orientazio semantikoa aldarazten du, positibo izatetik negatibo izatera pasatzen baita. (20) adibidean, berriz, *susmagarria* adjektiboaren orientazio semantikoak bera izaten jarraitzen badu ere, balentzian[29] igoera bat egon da *oso* intentsifikatzailearen ondorioz. (46) adibidean, ordea, modalak esaldiaren orientazio semantikoa neutralizatu egiten du. Izan ere, deskribatzen den egoera ez da gertatu eta horrenbestez, *eraginkorra* adjektiboaren balentzia +2 izatetik 0 izatera pasatzen da. Antzeko zerbait gertatzen da (22) adibidean. Bertan, azken esaldian, tempusa alegiazkoa denez, aurreko esaldietan gertatzen den egoera zalantzan jartzen da eta ondorioz, esaldiaren orientazio semantikoa neutralizatzen da. (23) adibidean egoera ezberdina da, izan ere, *nekez* adberbioak, osagai aurresuposizionala[30] denak, orientazio semantiko positiboko esaldia negatibo bilakatzen du. Azkenik, (24) adibidean, ironia dago eta horrek esaldiko hitzen sentimenduen balentzian aldaketak sortarazten ditu. Aldaketen ondorioz, esaldia +5 sentimendu balentzia izatetik, −4 sentimendu balentzia izatera igarotzen da.

Diskurtso mailan ere hainbat testuinguruko balentzia-aldatzaile daudela adierazten dute. Besteak beste, jarraian ageri diren adibideak aipatzen dituzte.

(25) [Boris matematikan [aparta]$_{+2}$ $\rightarrow_{0}$ izan arren, bera oso irakasle [txarra]$_{-2}$ da]$_{0}$ $\rightarrow_{-2}$.

(26) John asko ibiltzen da oinez. Aurreko hilean, astearteetan 25 milia egiten zituen oinez.

(27) Etxea. Kokalekua. Prezioa. Hedadura.

[29]Sentimendu-balentzia hitz batek duen informazio subjektiboa adierazteko zenbakizko eredua da. Zenbakiak eskala batean kokatuta egon ohi dira eta negatiboak edo positiboak izan daitezke hitzak duen orientazio semantikoaren arabera.

[30]Pragmatikan, aurresuposizioa (ingelesez, *presupposition*) diskurtsoan ontzat jotzen den egia duen adierazpen batekin zerikusia duen sinesmenaren edo munduaren gaineko jarrera inplizitu bat da.

(28) Hura pertsona [zerria]$_{-1}$ da. Hark esan du hura pertsona [zerria]$_{-1}$ $\rightarrow_0$ dela.

(29) Bilbo hiriaren egoera. Kokapenari dagokionez, Bilbo leku bikainean kokatzen da. Itsasoa gertu du eta inguruko mendiak ikusgarriak dira. Hiriari dagokionez, Bilbo oso kutsatuta dago eta zikinkeria edonon topa daiteke.

(30) [Filmak [txundigarria]$_+$ $\rightarrow_0$ izan beharko luke.
Aktoreak [lehen mailakoak]$_+$ $\rightarrow_0$ dira. Stallone gizon [zoriontsua]$_+$ $\rightarrow_0$ eta [zoragarria]$_+$ $\rightarrow_0$ da. Bere emaztea [gozoa]$_+$ $\rightarrow_0$ [ederra]$_+$ $\rightarrow_0$ da eta bere gizona [gurtzen]$_+$ $\rightarrow_0$ du. Gizonak [opari]$_+$ $\rightarrow_0$ [txundigarria]$_+$ $\rightarrow_0$ du [bizitza]$_+$ $\rightarrow_0$ [guztiz]$_+$ $\rightarrow_0$ [bizitzeko]$_+$ $\rightarrow_0$.
Argumentuak sekulakoa$_{+1}$ $\rightarrow_{-1}$ dirudi, hala ere, filma [porrot]$_{-1}$ bat da]$_{-2}$.

(25) adibidean, *izan arren* kontzesio menderagailuaren ondorioz, *aparta* adjektiboaren +2 sentimendu balentzia neutralizatu egiten da. (26) adibideak, aldiz, diskurtso-egitura bat islatzen du. Zehatz esanda, bertako bi esaldiak ELABORAZIOA izeneko erlaziozko diskurtso-egituraren bidez lotuta daude. ELABORAZIOA eta LISTA[31] bezalako erlazioek aurretik esandakoari buruzko informazioa gehitzen dute. Ondorioz, (26) adibidean, bigarren esaldiak lehena indartzen du.

(27) adibideak, entitate askoko ebaluazio bat islatzen du. Hau da, esaldiko gai nagusia edo aspektua etxea da eta, ondoren, horri buruzko zenbait entitate daude: kokalekua, prezioa, hedadura. Bereizi egin behar dira entitateetako bakoitzak eta aspektuak berak duen orientazio semantikoa. Kasua ezberdina da (28) adibidean, bertako perpaus osagarria edo konpletiboa dago. Perpaus osagarria denez, jarreraren ebaluazioa ez du esaldia esan duenak egiten, baizik eta beste batek. Ondorioz, *zerria* hitzaren sentimendu-balentzia neutralizatu egiten da. (29) adibidean, berriz, esaldi batean hainbat azpigai daude Bilbori buruzkoak. Kokapen geografikoari dagokionez, iritziak positiboak dira, baina hiriari buruzkoak negatiboak dira. Honek lotura zuzena

[31]ELABORAZIOA erlaziozko diskurtso-egiturak aurretik aipatutako proposizio baten aspektuari buruz informazio xeheagoa ematen duen erlazioa da. LISTA erlazioan, aldiz, lotura duten proposizioak zerrendatu egiten dira.

du aspektu eta entitate mailako sentimendu-analisiarekin, izan ere, hemengo gaia eta azpigaiak hurbilpen horretako entitatea eta aspektuak dira.

(30) adibidea genero murriztapenaren adibidea da eta (27) adibidearen antzekoa da. Bertan, filmari buruzko iruzkina egiten da baina, aldi berean, filmeko pertsonaien ebaluazioa ere egiten da. Genero murriztapena dago testuan entitate bat (filma) eta bere zenbait aspektu (tartean pertsonaiak) ageri direlako. Ondorioz, filmaren iritzia positiboa edo negatiboa zehaztu nahi bada, beharrezkoa da filmeko pertsonaiei dagozkien hitzen sentimenduen balentziak kontutan hartzea.

2.3.3. Balentzia-aldatzaileetan oinarritzen diren lanak

Polanyi eta Zaenenen (2006) laneko hurbilpen teorikoa proposatu ondoren, hainbat lan daude sentimenduen sailkapena egiteko balentzia-aldatzaileak erabiltzen dituztenak.

Kennedy eta Inkpenek (2006), esaterako, filmen iruzkinen sentimenduen sailkapena egiten dute balentzia-aldatzaileak (ezeztapena, intentsifikatzaileak eta ahultzaileak) aintzat hartuz. Lehen metodoan, iruzkinak hitz-kopuru positiboak eta negatiboak zenbatuz egiten dituzte. *General Inquirer*[32] bidez hitz positibo eta negatiboak eta balentzia-aldatzaileak identifikatzen dituzte eta sentimenduen lexikoi bat eta web corpusa ere badarabilte. Hitzen sentimenduen balentzia kalkulatzeko asoziazio kalifikazioa erabiltzen dute. Bigarren metodoan, berriz, euskarri bektoredun makina metodoa erabiltzen dute, balentzia-aldatzaileen unigramak eta bigramak ezaugarriak erabiliz.

Bestalde, Morsy eta Rafeak (2012) ere filmetako eta produktuetako iruzkinen sentimenduak sailkatzeko balentzia-aldatzaileak erabiltzen dituzte. Zehazki, intentsifikatzaileak, ezeztapena eta balentzia zeinua aldatzen duten sentimenduen hitzak erabiltzen dituzte ezaugarri moduan. Esperimentuetan, ezaugarri ezberdinak konbinatzen dituzte eta balentzia-aldatzaileekin eragin positiboa dutela adierazten dute.

[32] *General Inquirer* (Stone *et al.*, 1966) testuko edukia ordenagailu bidez aztertzeko lexikoia da. Lexikoi honek hitzen informazio sintaktiko, semantiko eta pragmatikoa biltzen du.

Azkenik, Ngoc Phu eta Thi Tuoiek (2014) dokumentu mailako sentimenduen sailkapena egiteko ere balentzia-aldatzaileak erabiltzen dituzte. Lehenik eta behin, esaldia hitzetan segmentatzen dute. Ondoren, termino kontaketa egiten dute, hots, sentimendu-balentzia duten hitzak kontatu eta horren araberako eragiketa egiten dute. Azkenik, sentimendu hitzei balentzia-aldatzaileen faktorea gehitzen diete eta eragiketa egiten dute berriz. Aintzat hartzen dituzten balentzia-aldatzaileak ezeztapena, konparaziozkoak eta superlatibozkoak, intentsifikatzaileak eta ahultzaileak dira.

Kasu honetan, gure lanak harreman zuzena du Ngoc Phu eta Thi Tuoiren (2014) lanarekin. Izan ere, bietan dokumentu mailako sentimendu-sailkapena egin nahi da sentimendu lexikoi bat erabiliz. Horrez gain, bi lanetan testuinguruko balentzia-aldatzaileak ere erabili nahi dira.

2.4 Sentimenduen analisia eta euskara

Orain arte ikusi dugun moduan, sentimenduen analisia oso arlo zabala da ataza, teknika eta ikuspegi asko baitaude. Aurrekariek erakusten duten bezala, lan gehienak ingeles hizkuntzan egin dira eta, haren atzetik, beste zenbait hizkuntza handi (tartean, gaztelera, frantsesa eta txinera) daude. Egoera arrunt ezberdina da hizkuntza gutxituetan eta horren arrazoi nagusietako bat baliabideen falta da. Atal honetan, euskara aintzat hartuta sentimenduen analisia arloan egin diren lanak azalduko ditugu. Lan horiek 2.18 taulan laburbilduta daude.

Lana	**Ataza**	**Teknikak**
Chen eta Skiena (2014)	Sentimendu-lexikoia	Hainbat lexikoi + jakintza-grafoa
Cruz *et al.* (2014)	Sentimendu-lexikoia	SentiWordNet 3.0 eredua
Barnes *et al.* (2018b)	Sentimendu-lexikoia	Embedding elebidunak
Saralegi *et al.* (2013)	Sentimendu-lexikoia	i) Corpusean oinarrituta ii) Hizkuntza arteko proiekzioan oinarrituta
Vicente eta Saralegi (2016)	Sentimendu-lexikoia	i) Beste hizkuntza batean oinarrituta ii) Corpusean oinarrituta ii) Jakintza base lexikalean oinarrituta
Barnes *et al.* (2018a)	Corpusaren sorkuntza	Webguneetatik iritzi-testuak lortu Hainbat geruzatan etiketatu
Vilares *et al.* (2017)	Lexikoian oinarritutako sentimenduen sailkapena	i) Sentimendu-lexikoia (SO-CAL eta ML-Senticon) ii) Analizatzaile sintaktikoa
San Vicente *et al.* (2015)	Aspektuan oinarritutako sentimenduen sailkapena	IXA pipe tools + Weka
Rodriguez *et al.* (2017)	Robotika	Buruaren eta besoen mugimenduak Begien mugimenduak Ahots-intonazioa Sentimendua: EliXa (IXA pipe tools + Weka)

2.18 taula: Euskara eta sentimendua uztartzen dituzten zenbait lan.

Baliabideen sorkuntzan, zehazki, sentimenduen lexikoiaren sorkuntzan, Chen eta Skienak (2014) munduko 136 hizkuntza nagusietarako, tartean euskararako, sentimenduen lexikoiak sortzen dituzte. Hainbat baliabide linguistikotan oinarrituz, jakintza-grafo handi bat sortzen dute eta hazi-hitzak zabalduz, grafoak hizkuntza bakoitzerako sentimenduen lexikoia sortzen du.

Bestalde, Cruz *et al.*ek (2014) euskara, ingelesa, gaztelania, galiziera eta katalan hizkuntzetarako sentimenduen lexikoiak sortzen dituzte. Lema mailako sentimenduen lexikoia automatikoki sortzeko metodoa aurkezten da bertan.

Beren lexikoiek hainbat geruza[33] dituzte; modu horretan, sentimendu lexikoi horiek erabiltzen dituzten aplikazioek aukera dute doitasun bat edo beste aukeratzeko eta doitasunaren arabera, lexikoiak eskaintzen dituen hitz gehiago edo gutxiago aukeratzeko.

```
<?xml version="1.0" encoding="UTF-8" ?>
<layers lang="en">
  <layer level="0">
    <positive>
      <lemma pos="a" pol="1.0" std="0.0"> admirable </lemma>
      <lemma pos="a" pol="0.438" std="0.088"> amorous </lemma>
      <lemma pos="n" pol="0.594" std="0.12"> approval </lemma>
      <!-- it continues... -->
    </positive>
    <negative>
      <lemma pos="a" pol="-0.437" std="0.157"> afraid </lemma>
      <lemma pos="n" pol="-0.275" std="0.105"> aggression </lemma>
      <lemma pos="v" pol="-0.25" std="0.0"> alarm </lemma>
      <!-- it continues... -->
    </negative>
  </layer>
  <layer level="1">
    <positive>
      <lemma pos="a" pol="1.0" std="0.0"> adept </lemma>
      <lemma pos="r" pol="0.25" std="0.0"> admirably </lemma>
      <lemma pos="n" pol="0.667" std="0.315"> admiration </lemma>
      <!-- it continues... -->
    </positive>
    <negative>
      <lemma pos="v" pol="-0.875" std="0.0"> abase </lemma>
      <lemma pos="a" pol="-0.25" std="0.0"> abashed </lemma>
      <lemma pos="v" pol="-0.25" std="0.0"> abhor </lemma>
      <!-- it continues... -->
    </negative>
  </layer>
  <layer level="2">
      <!-- it continues... -->
  </layer>
</layers>
```

2.4 irudia: ML-SentiCon lexikoiaren ingelesezko bertsioaren zati bat (Cruz *et al.*, 2014).

2.4 irudian, lexikoia geruzetan nola antolatuta dagoen ikus daiteke. Bertan, hitzak hiru geruza-mailetan (*layer level*) banatuta daude eta, horietako maila bakoitzean, hitzak daude beren sentimendu-balentziarekin. Adibidez, *amorous* ("maitekor") hitza 0 geruza-mailan dago, adjektiboa da eta 0,088

[33]Lexikoiak guztira 8 geruza ditu. Lehen geruza da murriztapenik gehiena eta lema gutxi dituena eta, ondoren, geruzaz igo ahala murriztapenak gutxitu egiten dira eta lema kopurua igo egiten da. Horrela, esaterako, bigarren geruzak lehen geruzako lemak eta beste berri batzuk ditu.

sentimendu balentzia du. 0 geruza-mailan dagoenez, bere doitasuna ez da oso handia.

Lema mailako lexikoia sortu aurretik, ingeleserako *synset* mailako lexikoia sortzen dute, *SentiWordNet 3.0*ren antzekoa dena. Azken hau sentimenduen lexikoirik erabiliena da gaur egun. Euskarazko sentimenduen lexikoiak 26.392 lema ditu (22.879 izen, 57 adjektibo, 3.456 aditz eta 0 adberbio) eta *ML-SENTICON* izeneko baliabidean eskuragarri dago.

Barnes *et al.*ek (2018b), berriz, besteen artean, euskarazko sentimenduen lexikoia sortzen dute; baina aurreko lanekiko hurbilpena ezberdina da. Euskara moduko baliabide gutxiko hizkuntzetan, corpus etiketatu falta handia dago eredu onak sortzeko. Horregatik, sentimendu *embedding* elebidunak (*Bilingual Sentiment Embeddings*, BLSE) sortzen dituzte beste hizkuntza bat eta lexikoia sortu nahi den hizkuntza uztartuz. Horretarako, baliabide hauek erabiltzen dituzte: i) lexikoi elebidun txiki bat; ii) jatorri hizkuntzako corpusa sentimenduz etiketatuta; iii) hizkuntza bakoitzeko hitz *embedding* elebakarrak. Esperimentuak gaztelania, katalanera eta euskaraz konbinatuz egiten dituzte esaldi mailako hizkuntzen arteko sentimenduak sailkatzean.

Saralegi *et al.*ek (2013) ere sentimenduen lexikoia sortzen dute. Euskara du aztergaitzat eta bi hurbilpen ezberdin alderatzen eta aztertzen ditu euskarazko sentimenduen lexikoiak sortzeko. Bi hurbilpenak automatikoak izateaz gain, erraz eskuratzeko baliabideetan oinarritzen dira eta, beraz, baliabide gutxiko hizkuntzetarako egokia da. Hurbilpen horietako bat corpusean oinarritutakoa da eta bestea, berriz, hizkuntzen arteko proiekzioan. Vicente eta Saralegik (2016), berriz, euskarazko sentimendu lexikoia sortzeko hiru metodo alderatzen dituzte: i) beste hizkuntza batetik itzuliz, ii) corpusetik sentimendu kalifikazioa duten hitzak erauziz eta iii) jakintza-base lexikaletik (*Lexical Knowledge Bases*) sentimenduak etiketatuz. Metodo bakoitzak zenbat eskuzko lan eta lortzen diren emaitzekiko eskuzko lan horrek merezi duten ebaluatzen dute lanean.

Bestalde, Barnes *et al.*ek (2018a) sentimenduen analisirako corpus bat sortzen dute. Bertan, aspektu mailako sentimenduen sailkapena egiteko, katalaneraz eta euskaraz dauden hotelei buruzko iritzien corpus anotatua sortzen dute. Katalanerazko 568 eta euskarazko 343 iruzkin biltzen dituzte.

Iruzkinak aurreprozesatu ondoren, OpeNER proiektuko hurbilpenaz etiketatzen dituzte eta etiketatzeko *KafAnnotator* tresna erabiltzen dute. Anotatu beharreko elementuak iritzi-helburua edo jomuga, iritzi-adierazpenak, iritzi-eramailea eta polaritatea dira.

San Vicente *et al.*en (2015) *EliXa* sistema aspektuan oinarritutako sentimenduen analisia egiteko tresna da. Sistema honen abantaila da plataforma modularra dela; horrek aukera ematen du tresnari ezaugarriak kentzeko edo berriak gehitzeko. Sistema *IXA pipe tools* (Agerri *et al.*, 2014) eta Wekan oinarritzen diren hiru modulu independientez osatuta dago. Tresnari ezaugarri berriak gehitu ahal izateak egiten du posible sistema hau euskararako baliagarri izatea. *EliXa*k hizkuntza-baliabide zehatzak erabil ditzake, hala nola polaritate lexikoak eta testua normalizatzeko balio duten baliabideak. Momentuz, lau hizkuntzarako lau baliabide eskaintzen dituzte, tartean, euskararako. Euskararako baliabideak Elhuyarren sentimenduen lexikoia, maiztasun handiko hitzen zerrenda (ingelesez, *stopwords*) eta hiztegitik kanpoko hitzak (ingelesez, *Out-Of-Vocabulary*, OOV) biltzen ditu, besteak beste.

Vilares *et al.*ek (2017) sentimenduen lexikoia eta sintaxia uztartzen dituzte sentimenduen sailkapena egiteko. Iberiar penintsulako hizkuntzetarako polaritatearen sailkapena egitea dute helburu eta, horretarako, sentimenduen lexikoiaz eta erregela sintaktikoez[34] baliatzen dira. Hurbilpen hori analizatzaile sintaktikoan oinarritzen da. Gainera, erregelak dependenteak dira eta egoera eleanitzetan moldaketa bat behar dute. Euskararen kasuan ez ezik, baita katalaneran eta galizieran ere, aurretik dauden hainbat lexikoi bateratu egiten dituzte. Erabiltzen duten lexikoietako bat SO-CAL tresnakoa (Taboada *et al.*, 2011) da. Euskarazko bertsioa itzulpena eginez lortzen dute eta 4.066 sarrerako lexikoia lortzen dute. Erabilitako beste lexikoia *ML-Senticon* da. Lexikoi horren euskarazko bertsioak 1.662 sarrera ditu. Bi lexikoien bateraketa egin ondoren, lortzen duten lexikoiak 5.134 sarrera ditu. Sortzen duten baliabidearen beste atala gramatika-kategorien etiketatzailea eta dependentzia etiketatzaileak dira. Kasu hauetan, aurretik sortuta dauden baliabideak erabiltzen dituzte. Azken urratsean, konposizio-eragiketak egiten dituzte.

[34]Erregela sintaktikoak esaldien egitura finkatzen dituzten erregelen multzoa da.

Rodriguez *et al.*ek (2017), berriz, aurretik aipatu dugun *EliXa* sistema robotikan aplikatzen dute. Gizakiari izaera soziala ematen diona emozioen adierazpena da. Idatzizko eta ahozko moduez gain, pertsona baten egoera mentala adierazten duen moduetako bat da. Robot baten eta gizaki baten artean, hori islatzea beharrezkoa da, interakzioa ahalik eta naturalena izateko. Lan honetan, ahozko testuari eduki emozionala jartzen diote. Buru- eta beso-mugimenduekin, begien mugimenduarekin eta ahots intonazio ezberdinak erabiliz, robota gai da pozezko edo tristurako emozioak adierazteko, bai euskaran, baita ingeles eta gaztelanian ere. Robota, euskararen kasuan, emozio bat 0 eta +10 arteko eskalan sailkatzeko (bi muturrak tristura eta poztasuna dira) *EliXa* sisteman oinarritzen da.

Atal honetan aipatutako lanek badituzte zenbait berdintasun eta ezberdintasun guk garatutako tresna eta baliabideekin. Sentimenduen lexikoiaren sorkuntzan, gure hurbilpena izan da SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskaratzea eta ingelesekoarekin aberastea. Bestelako hurbilpenak erabili dituzte aipatu ditugun lanetan, izan ere, batzuetan *SentiWordnet* erabili dute (Cruz *et al.*, 2014), beste batzuetan, berriz, *embedding* bezalako testu-analisirako teknikak erabili dituzte (Barnes *et al.*, 2018b) eta azkenik, badira gure antzera corpusak erabili dituztenak ere (Saralegi *et al.*, 2013), nahiz eta ikuspegia ezberdina izan.

Corpusaren sorkuntzari dagokionez, Barnes *et al.*ek (2018a) gure teknika bera erabiltzen dute; hots, webguneetatik iritzi-testuak lortzen dituzte. Gero, hala ere, egindako etiketatzea ezberdina da. Azkenik, dokumentu mailako sentimenduen sailkapenean, gure lanak Vilares *et al.*enarekin (2017) antzekotasunak ditu, izan ere, bietan lexikoian oinarritutako hurbilpena eta sintaxia baliatzen dira, nahiz eta guk diskurtso maila ere lantzen dugun.

2.5 Laburpena

Kapitulu honetan, sentimenduen analisi arloan dauden ataza eta aurrekari ezberdinak aztertu ditugu. Ikusi dugunez, sentimenduen arloan hiru ataza nagusi daude: i) dokumentu mailako sentimenduen sailkapena, ii) subjektibitatearen sailkapena esaldi mailan eta iii) iritziaren jabearen erauzketa entitate eta aspektu mailan. Horietaz gain, beste ataza batzuk ere badaudela ikusi dugu eta hizkuntzaren prozesamenduaren beste arloekin lotura zuzena dutela, iritzi-laburpena edota *spam*ak diren iritzien detekzioa, adibidez.

Baliabide-sorkuntzari dagokionez, hainbat motatako iritzi-testuen corpusak daudela ikusi dugu. Testu-genero ezberdinak biltzen dituzte batzuek. Beste batzuk, berriz, corpus eleanitzak dira eta corpus gehienak etiketatuta daudela ere ikusi dugu. Sentimenduen lexikoiari dagokionez, horiek sortzeko biderik ohikoena corpus edo hiztegi bidezkoa da.

Balentzia-aldatzaileen identifikazioan egin diren lanak ere landu ditugu. Balentzia-aldatzailearen kontzeptua testuaren idazleak testuak eskaintzen dituen baliabideen bidez iritzia adierazteko moldeak identifikatzeko sortu dela azaldu dugu eta, ondoren, horren inguruko erreferentziazko lana den Polanyi eta Zaenenek (2006) zer-nolako balentzia-aldatzaileak deskribatzen dituen aztertu dugu. Azkenik, balentzia-aldatzaileak kontuan hartuz sentimenduen analisian egin diren zenbait lan aipatu ditugu.

Azken atalean, euskara eta sentimenduen analisia lantzen dituzten lanak deskribatu ditugu. Lan gehienek euskarazko sentimenduen lexikoia sortzea dute helburutzat, askotan, horretarako SentiWordNet datu-base lexikala erabiliz.

Gure aukerei dagokienez, corpusaren kasuan, gai ezberdinetako iritzi-testuez osatuta egotea nahi dugu, baita orientazio semantikoaren aldetik orekatua eta diskurtso-egitura RST hurbilpenaz etiketatuta egotea ere. Sentimenduen lexikoiari dagokionez, gure aukeraketa hiztegian eta corpusearen oinarritzen da. Guk sortu nahi dugun dokumentu maila sentimenduen sailkatzailea, berriz, lexikoian oinarritutakoa izatea hautatu dugu. Azkenik, balentzia-aldatzaileei dagokienez, Polanyi eta Zaenenek (2006) deskribatzen dituenen artean, sintaxi eta diskurtso mailakoak aztertzea erabaki dugu, fonologia eta morfologia mailekoez gain.

METODOLOGIA

3. KAPITULUA

Metodologiaren diseinua

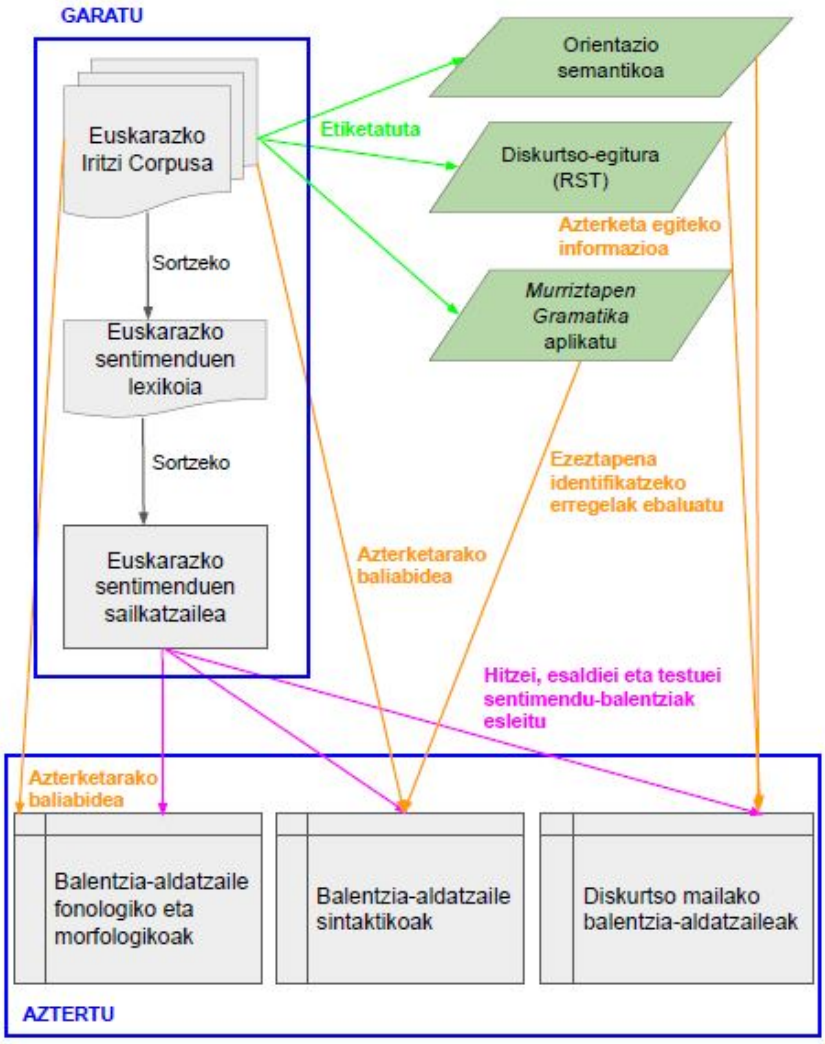
1. kapituluan zehaztutako helburuak betetzeko metodologia bat jarraitu eta zenbait baliabide eta tresna erabili ditugu. Kapitulu honetan horien inguruan arituko gara. 3.1 irudian, tesi-lan honen metodologiako urratsak laburbilduta ikus daitezke.

1- Ikerketa aurrera eramateko oinarrizko baliabideak sortu. Zehazki, Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016), euskarazko sentimenduen lexikoia eta euskatrazko sentimenduen sailkatzaile bat sortu ditugu.

   Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) 240 iritzi-testuz osatuta dago eta etiketatuta ere badago. Alde batetik, iritzi-testu osoaren orientazio semantikoa positiboa edo negatiboa den etiketatu dugu. Beste aldetik, corpuseko 70 iritzi-testuren diskurtso-egitura etiketatu dugu *Egitura Erretorikoaren Teoria* (RST) (Mann eta Thompson, 1988) erabiliz. Azkenik, corpuseko 48 iritzi-testuei *Murriztapen Gramatika* (MG) (Karlsson *et al.*, 1995) aplikatu diegu.

   Euskarazko sentimenduen lexikoia ere sortu dugu eta, horretarako, besteak beste, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarritu gara. Azkenik, euskarazko sentimenduen sailkatzailea ere garatu dugu eta bere oinarrian sortu dugun sentimenduen lexikoia dago.

2- Balentzia-aldatzaileak identifikatu. Bigarren urratsa, hizkuntza maila ezberdinetan, sentimendu-balentzia duten hitzetan eragin dezaketen fenomeno linguistikoak aztertu eta beren eragina neurtzea izan da.



3.1 irudia: Tesi-lanaren metodologia.

Balentzia-aldatzaile fonologikoak aztertu ditugu eta azterketa egiteko Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) eta euskarazko sentimenduen sailkatzailea erabili ditugu. Balentzia-aldatzaile sintaktikoetan ezeztapen-markak aztertu ditugu. Ezeztapen-marken adibideak Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) aurkitu ditugu eta, ondoren, ezeztapen-marka horiek identifikatzeko erregelak *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortu eta Euskarazko Iritzi Corpu-

sari(Alkorta *et al.*, 2016) *Murriztapen Gramatika* (Karlsson *et al.*, 1995) aplikatuta, erregela horiek ebaluatu ditugu.

Azkenik, diskurtso mailan, erlaziozko diskurtso-egitura eta unitate zentrala aztertu ditugu eta balentzia-aldaketan zer-nolako eragina duten zehaztu dugu. Hori egiteko Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) (diskurtso-egituraz eta orientazio semantikoz etiketatuta) eta euskarazko sentimenduen sailkatzailea erabili ditugu.

Jarraian, metodologiaren deskribapen zabalagoa egin dugu.

3.1 Sentimenduen analisirako baliabide eta tresnen sorkuntza

3.1.1. Euskarazko Iritzi Corpusa

Euskarazko Iritzi Corpusa sortzeko, lehenik eta behin, zer-nolako corpusa sortu nahi dugun erabaki dugu. Corpuseko testuek jarraian azaltzen diren ezaugarriak izatea erabaki dugu:

- Iritzi-testuak balorazio argi bat izatea. Testu batzuek iritzi argirik ez dutela oharturik, aldeko edo aurkako iritzi argia duten testuak hobetsi ditugu. Iritzi argia adierazten duten iritzi-testuak biltzen baditugu, tesi-lan honetan garatutako sentimenduen sailkatzailea hobeto ebaluatuko dugu.
- Iritzi-testuak egitura sintaktiko aberatsa eta balorazioa adierazteko baliabideak edukitzea. Iritzi-testuek egitura sintaktiko ezberdinak eta balorazioa egiteko molde ezberdinak edukitzea nahi izan dugu. Modu horretan, hizkuntza mailetako eta molde ezberdinetako balentzia-aldatzaileak identifikatzeko aukera izango dugu.

Irizpideak zehaztu ondoren, euskarazko iritzi-testuak dauden webguneak bilatu ditugu. Webgune horiek espezializatuak dira edota aldizkari eta egunkarienak dira. Labur azalduta, hiru iturri ezberdin erabili ditugu: i) egunkariak (hala nola, *Berria*, *Argia*, *Naiz* eta *El Correo*), ii) webgune espezializatuak (*Kritiken Hemeroteka*, *Entzun.com Zinea.eus* eta *Aizu*) eta, azkenik, iii) hainbat blog (horien artean, Joxe Landaren bloga, Baleikeko eguraldia atala eta EiTBko txirrindularitza bloga).

Euskarazko Iritzi Corpusa osatzeko webguneak aurkitu ostean, corpusak eduki beharreko ezaugarriak finkatu ditugu. Ezaugarriak hiru dira:

- Corpusak tamaina handia izango du. Corpusa 240 iritzi-testuz osatuta egongo da. Gure ustez, tamaina honetako corpusa nahikoa da gure helburuak betetzeko eta sentimenduen analisiko beste corpusek ere antzeko tamaina dute.

3.2 irudia: *Kritiken Hemerotekaren* webgunea.

- Corpusa domeinu anitzekoa izango da. Iritzi-testuak sei domeinutakoak izango dira: eguraldia, politika, kirola, zinema, musika eta literatura. Corpusa domeinu askotakoa izatea nahi dugu, modu horretan, domeinuari loturiko berezitasunak edo balentzia-aldatzaile zehatzak identifikatzeko aukera izango baitugu.

- Corpuseko testuek balorazio orekatua izango dute. Domeinu bakoitza testuek erakusten duten balorazioari dagokionez orekatua izango da. Hau da, domeinu bakoitzean balorazio positiboko 20 testu eta balorazio negatiboko beste 20 testu egongo dira. Corpusa orekatua izatea erabaki dugu eredutzat hartu dugun *SFU Review Corpus* (Taboada, 2008) halakoa delako.

Hurrengo urratsean, bildutako iritzi-testuekin datu-base bat sortu dugu. Datubase horretan, iritzi-testu bakoitzaren ezaugarri hauek zehaztu ditugu:

- Kodea. Iritzi-testu bakoitzari kode bat jarri diogu. Kode horrek beti *GAIA-zenbakia_balorazioa* egitura du.
  - Gaia: liburua (LIB), musika (MUS), pelikula (ZIN), politika (POL), kirola (KIR) eta eguraldia (EGU).

  - Zenbakiak: 1etik 40 arterainoko zenbakiak jarri dizkiegu iritzi-testuei, gaiko 40 iritzi-testu bildu baititugu.
  - Balorazioa: positiboa (POS) eta negatiboa (NEG).

- Izenburua. Testu bakoitzaren izenburua izan da zehaztu dugun beste ezaugarrietako bat.

- Helbidea. Iritzi-testua Interneten non dagoen ere zehaztu dugu, iturria onartzeko eta edozeinek lortu ahal izateko.

- Balorazioa. Datu-basean, iritzi-testuetako bakoitza positiboa edo negatiboan den ere zehaztu dugu.

- Tamaina. Iritzi-testuak duen hitz-kopurua ere zehaztu dugu. Hitz-kopurua kontatzeko, *Analhitza* tresna (Otegi *et al.*, 2017) erabili dugu.

Iritzi-testuetako bakoitzari loturiko informazioa datu-basean nola egituratuta dagoen ikus daiteke 3.1 taulan.

Domeinua	**Kodea**	**Izenburua**	**Iturria**	**Balorazioa**	**Hitz-kopurua**
Literatura	LIB29_pos	Zangotraba	Aizu	Positiboa	175
Kirola	KIR15_neg	Vigotik esku hutsik itzuli da Reala	Berria	Negatiboa	169

3.1 taula: Datu-basearen antolaketaren bi adibide.

Bestetik, iritzi-testuak formatu egokian ere jarri ditugu Hizkuntza Prozesamenduko tresnekin prozesatu ahal izateko. Testuak *txt* eta UTF-8 karaktere kodifikazio formatuan jarri ditugu.

Sortuko dugun corpusa ikerketarako egokia den ere neurtu nahi izan dugu eta horregatik, corpusa ingelesezko beste corpus subjektibo batekin (*SFU Review Corpus* (Taboada, 2008)) eta ingelesezko eta euskarazko bi corpus objektiboekin (corpus subjektiboen gai berak lantzen dituzten Wikipediako artikuluen multzoa) alderatu dugu. Honako aspektuak aztertu ditugu:

- Lehen pertsonaren presentzia. Ingelesean pertsona izenordainetan eta euskaran aditzetan, lehen pertsonak zenbateko presentzia duen neurtu dugu. Pertsonek beren esperientziak lehen pertsona singularrean edo pluralean azaltzen dituztelako aztertu dugu (Li *et al.*, 2011; Villarroel Ordenes *et al.*, 2017). Euskararen kasuan, 3.3 irudian agertzen diren instantzia guztietatik zenbat ageri diren lehen pertsona singularrean edo pluralean kontatu dugu.

```
Emaitza indibidualak
NI_ZURI 10
NR_GU   74
NR_HI   34
NK_NIK  185
NR_ZUEK 4
NR_HURA 4930
NR_HAIEK    1266
NI_ZUEI 3
NI_NIRI 77
NO  3
NI_GURI 149
NR_NI   73
NK_ZUK  22
NK_GUK  300
NI_HAIEI    77
NI_HARI 273
NK_HAIEK-K  635
NK_HIK-TO   1
NK_HIK-NO   32
NR_ZU   7
NK_HARK 2077
NK_ZUEK-K   12
NI_HIRI-NO  6
```

3.3 irudia: Euskarazko aditzetan dagoen lehen pertsonaren erabilearen neurketa.

- Adjektiboen presentzia. Gramatika-kategoria guztietatik, adjektiboek zer pisu duten neurtu dugu bina corpus objektibo eta subjektibotan. Pang *et al.*ek (2002) diotenez, adjektiboak izan dira gramatika-kategoriarik erabilienak sentimenduen sailkapeneko atazetan eta doitasun ona lortu dute. Horregatik aztertu dugu adjektiboen presentzia.

- Ezeztapen-markak. Zenbait lanen arabera (Dadvar *et al.*, 2011; Wiegand *et al.*, 2010; Jia *et al.*, 2009) ezeztapen-markek hitzen edota sin-

tagmen sentimenduen balentzian eragin dezaketelako eta hori euskaran aztertu nahi dugulako, euskarazko iritzi corpusean zenbat ezeztapen-marka dauden zenbatu dugu. Ezeztapen-marken kasuan ere, zenbaketa ordenagailuak eskaintzen dituen baliabideen bidez egin dugu.

Bukatzeko, Euskarazko Iritzi Corpusa diskurtso- eta sujektibitate-informazioz etiketatu dugu. Lehenik eta behin, corpuseko 240 testuetatik 70 iritzi-testu *Egitura Erretorikoaren Teoria* (RST) hurbilpena baliatuz etiketatu ditugu. Beste lanetan etiketatu den adina testu etiketatu dugu, eta, ondorioz, ez dugu corpus osoa etiketatu. Bestalde, RST erabili dugu euskaran gehien landu den hurbilpena delako. 3.2 taulan ikusten den moduan, guztira 70 testu (corpusaren % 29,16) etiketatu ditugu eta horietatik 19 testuek (etiketatutakoaren % 27,14) etiketatze bikoitza dute.

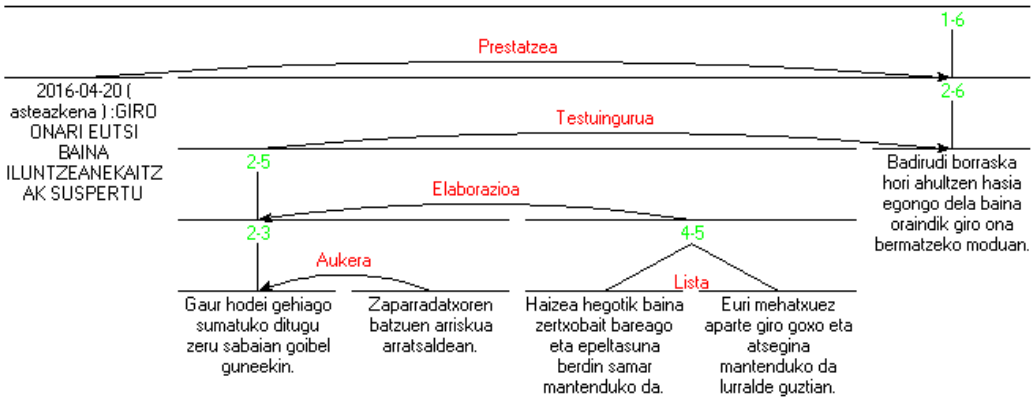
	A1	**A2**	**Etiketatuak guztira**	**Etiketatze bikoitza**
Zinema	30	9	30	9
Eguraldia	15	5	15	5
Literatura	5	25	25	5
Guztira	**50**	**39**	**70**	**19**

3.2 taula: Bi anotatzailek (A1 eta A2) anotatutako iritzi-testuen kopurua.

Anotazioa burutzeko, bi anotatzaileen Das eta Taboadaren (2018) irizpideei jarraitu behar izan diete. Domeinutik domeinura ezberdintasun batzuk egon dira etiketatzeko prozesuan. Esaterako, anotatzaileek eguraldiaren domeinuko testu bat anotatzeko 20 minutu inguru igaro behar izan dituzte eta literaturakoak anotatzeko, aldiz, ordu bete inguru. Horren arrazoiak testuen luzera eta konplexutasuna dira, eguraldikoak testu laburragoak eta sinpleagoak baitira, literaturakoen aldean. 3.4 irudian, eguraldiaren domeinuko iritzi-testu bat (EGU04) anotatuta ikus daiteke. Bi anotatzaileen arteko adostasunaren emaitzak 3.3 taulan daude. Zehazki, erlaziozko diskurtso-egitura ondo etiketatu den neurtu dugu.

Corpusaren etiketatzea bi modutan ebaluatu dugu. Alde batetik, eskuz unitateen esanahi erretorikoa ebaluatu dugu eta, bertan, azpiosagai zentralaren[1]

[1]Azpiosagai zentrala *span*eko unitaterik garrantzitsuena da (Egg eta Redeker, 2010).



3.4 irudia: EGU04 iritzi-testuaren diskurtso-egituraren anotazioa.

posizioa ez da aintzat hartu. Beste aldetik, ebaluazio kualitatiboa egin dugu. Kasu horretan, unitateen esanahi erretorikoaz gain, nuklearitatea, lotura eta osagaik hartu dira kontuan. Halaber, eskuzko ebaluazioan ez bezala, azpiosagai zentralaren posizioa aintzat hartu da ebaluazioa egiteko momentuan.

3.3 taulako emaitzek, eskuzko ebaluazioaren emaitzek, erakusten dutenez, erlaziozko diskurtso-egitura etiketatzean bi pertsonen arteko adostasuna % 9,81ekoa da. Ezbedintasunak daude domeinutik domeinura. Eguraldiaren domeinuan, adostasuna % 43,59ra igotzen da eta zinemaren domeinuan, aldiz, adostasuna % 37,73ra jaisten da. Literaturaren domeinuan, adostasuna % 41,67koa da.

Domeinua	**Adostasuna (%)**
Eguraldia	43,59 (17/39)
Literatura	41,67 (70/168)
Zinema	37,73 (83/220)
Guztira	**39,81 (170/427)**

3.3 taula: Bi etiketatzaileen arteko adostasun-maila iritzi-testuak RST hurbilpenaz etiketatzean (eskuzko ebaluazioa).

Bi pertsonen arteko adostasuna neurtu dugu ebaluazio kualitatiboa[2] egiten duen tresna erabiliz ere. Emaitzak 3.4 taulan daude ikusgai.

[2]Iruskieta *et al.*ek (2015) aipatzen duten moduan, ebaluazio mota honi kualitatiboa

	Osagaiak		Lotura		Nuklearitatea		Erlazioa	
Domeinua	**Bat egin**	**F1**	**Bat egin**	**F1**	**Bat egin**	**F1**	**Bat egin**	**F1**
Eguraldia	20/37	0,54	9/37	0,24	22/37	0,59	15/37	0,41
Literatura	84/155	0.54	67/155	0,43	105/155	0,68	48/155	0,31
Zinema	112/221	0,56	88/221	0.40	147/221	0,67	68/221	0,31
Guztira	**216/413**	**0,52**	**164/413**	**0,40**	**274/413**	**0,66**	**131/413**	**0,32**

3.4 taula: Etiketatzaileen arteko adostasunari buruz tresna automatikoak egindako ebaluazio kualitatiboa.

Eskuzko neurketan ez bezala, ebaluazio kualitatiboan[3] erlaziozko diskurtso-egitura motaz gain, beste zenbait aspektu neurtu ditugu. Diskurtsoen erlazio motari dagokionez, adostasuna 0,32koa da, eskuzkoa baino 0,08 puntu baxuagoa. Ezberdintasun horren atzean, eskuzko ebaluazioan ez bezala, tresnak azpiosagai zentralaren (ingelesez, *central subconstituent*) posizioa kontuan hartzen duela dago. Ebaluatu diren beste aspektuetan, adostasuna handiagoa da. Loturaren kasuan, adostasuna 0,40ekoa da eta osagai eta nuklearitate aspektuetan adostasuna 0,50etik gorakoa da, 0,52 eta 0,66koak baitira, hurrenez hurren.

Diskurtso-egitura etiketatu ondoren, corpus bera subjektibitatearen ikuspegitik etiketatu da. Zehazki esanda, testuetako eta erlaziozko diskurtso-egiturako zenbait osagairen orientazio semantikoa izan da etiketatu dena.

Bi anotatzaileek erlaziozko diskurtso-egituretako hiru osagai etiketatu behar izan dituzte: i) erlaziozko diskurtso-egitura bera, ii) haren nukleoa edo nukleoak (KONTRASTEA erlazioaren kasuan) eta iii) haren satelitea. Esleitu beharreko etiketak ere hiru dira: i) orientazio semantiko positiboa, ii) orientazio semantiko negatiboa eta iii) orientazio semantiko neutrala.

3.5 taulan, anotatzaile bat erlaziozko diskurtso-egitura etiketatzen ikus daiteke. Anotatzaileak lehen lerroan nukleoa den testu-zati bat du eta orientazio

deritzo hizkuntza ezberdinetan edo/eta pertsona ezberdinek garatutako diskurtso-egiturak alderatzeko aukera ematen duelako. Orain arteko ebaluatzeko moduak kuantitatiboak izan dira, baina honen ezberdintasuna da EDUak, testu-zatiak, nuklearitatea eta erlaziozko diskurtso-egiturak hartzen dituela kontuan.

[3]Ebaluazio kualitatiboak aspektu hauek aztertzen ditu: osagaia (oinarrizko diskurtso-unitateak), lotura (unitateek erlazioekin duten lotura), nukleartasuna (erlazioa N(ukleo)-S(atelite), S(atelite)-N(ukleo) edo N(ukleo)-N(ukleo) den) eta erlazioa (unitateen esanahi erretorikoa).

semantiko positiboa esleitu dio. Bigarren zutabean, satelitea den testu-zatia dago eta horri orientazio semantiko neutrala esleitu dio. Azkenik, hirugarren lerroan, erlaziozko diskurtso-egitura osoari orientazio semantiko positiboa esleitu dio.

Kodea	Erlazioa	Esaldia	Orientazio semantikoa	N-S
SENTARG01-A1.rs3	AHALBIDERATZEA	Aukera ederra da irakurlearentzat, Mirandek gaztetik hasi eta ia hil arte eutsi zion ohitura horri esker, modu kronologikoan emana gainera.	POS	N
		Froga bat egin liteke, herrien izena eta batez ere bere izena zein hizkuntzatan idazten duen begiratzea, une horretan bere herriarekiko sentimenduak eta harremanak nolakoak diren jakiteko.	NEU	S
		Aukera ederra da irakurlearentzat, Mirandek gaztetik hasi eta ia hil arte eutsi zion ohitura horri esker, modu kronologikoan emana gainera. Froga bat egin liteke, herrien izena eta batez ere bere izena zein hizkuntzatan idazten duen begiratzea, une horretan bere herriarekiko sentimenduak eta harremanak nolakoak diren jakiteko.	POS	N-S

3.5 taula: Erlaziozko diskurtso-egitura baten orientazio semantikoaren esleiketa.

Guztira, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden 240 iritzi-testuetatik 28 iritzi-testu erabili ditugu eta orientazio semantikoaren etiketatzea iritzi-testu horietako 384 erlaziozko diskurtso-egituretan egin dugu. 3.6 taulan ikusten den moduan, etiketatutako erlaziozko diskurtso-egiturek esanahi erretoriko ezberdinak dituzte[4]. Etiketatzaile batek 384 erlaziozko diskurtso-egitura etiketatu ditu eta beste batek horien % 40 etiketatu, etiketatzearen kalitatea neurtzeko. Hain zuzen ere, hori izan da etiketatzearen kalitatea neurtzeko erabili dena.

Horretarako, gidalerro batzuk sortu ditugu. Gidalerro horietan etiketatzaile bakoitzak egin beharrekoa azaltzen da. Kasu gatazkatsuenen berri ere ematen da, tartean esaldi metaforikoena, eta horiek nola anotatu behar diren ere esplikatzen da.

(31) (...) eta “egilearen intentzioa” $[\text{usna}]_+$ liteke. (SENTBER04)

[4]Orientazioa semantikoa etiketatuta duten iritzi-testuak *Basque RST Treebank* baliabidean (Iruskieta *et al.*, 2013) eskuragarri daude.

Erlaziozko diskurtso-egitura	**Instantziak**
AHALBIDERATZEA/MOTIBAZIOA	6
Baldintza taldea	18
KONTRASTEA	26
EBIDENTZIA/JUSTIFIKAZIOA	53
KONTZESIOA/ANTITESIA	70
Kausa taldea	71
EBALUAZIOA/INTERPRETAZIOA	140
Guztira	**384**

3.6 taula: Etiketatutako erlaziozko diskurtso-egiturak.

Esate baterako, (31) adibidean, *usnatu* aditza ageri da. Ez da berak duen berezko zentzuan ageri eta, adibide horretan, *usnatu* "egileak zer egin nahi duen antzematea" esatearen parekidea da. Horregatik, mota horretako adibideen orientazio semantikoa etiketatzea erabaki dugu. Etiketatzaileen etiketatze-adostasunaren emaitzek erakusten dutenez, Cohenen kapparen koefizientea 0,58 da. Beraz, *neurrizko adostasuna* (Landis eta Koch, 1977) da.

Bi anotatzaileen arteko desadostasunik handienak orientazio semantiko neutralarekin lotuta daude, 3.7 taulan ikus daitekeen moduan.

		A2			
		NEG	**NEU**	**POS**	**Guztira**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Guztira	92	120	181	393

3.7 taula: Erlaziozko diskurtso-egituren sentimendu balentzia anotazioari dagokion bi anotatzaileen arteko kontingentzia-taula.

Etiketatzaile batek orientazio semantiko neutrala esleitzen duenean, beste anotatzaileak ez du esleitzen eta alderantziz ere gertatzen da. Hainbat etiketatze-proba egin ostean eta gidalerroak hobetu ostean, etiketatzaile batek 384 erlaziozko diskurtso-egituren gainerako % 60aren orientazio semantikoa etiketatu du.

3.1.2. *Sentitegi* izeneko euskarazko sentimenduen lexikoiaren sorkuntza eta ebaluazioa

Euskarazko sentimendu lexikoia sortzerakoan, lehenik eta behin, horretarako zer-nolako baldintzak dauden aztertu dugu. Hiru aspektu nabarmentzen dira:

- Denbora. Sentimenduen lexikoia lanaren muina izan arren, ezin izan dugu denbora guztia horretan baliatu. Sentimenduen lexikoitik abiatuz, sentimenduen analisia atazaren hainbat alderdi (balentzia-aldatzaileak eta dokumentu mailako sentimenduen sailkatzailea, adibidez) lantzea dugu helburu. Horrek esan nahi du denbora banatu egin behar dugula sentimenduen lexikoia sortzearen eta sentimendu-analisiko hainbat atal lantzearen artean. Ondorioz, euskarazko lexikoia sortzeko denbora mugatua dugu.

- Tresnak eta baliabideak. Lexikoia sortzeko izan dugun beste mugetako bat eskuragarri dauden euskarazko tresna eta baliabideek dituzten ezaugarriak dira. Euskarazko hainbat sentimendu lexikoi sortu izan dira hurbilpen ezberdinak baliatuta (Chen eta Skiena, 2014; Cruz *et al.*, 2014; Barnes *et al.*, 2018b; Saralegi *et al.*, 2013; Vicente eta Saralegi, 2016), baina hitzei sentimendua esleitzeko moldea guretzat ez da egokia hainbat arrazoirengatik. Batzuetan, hitzari esleitutako sentimenduen balentzia ez dago eskala batean eta, horrenbestez, ezin da hitz batean gerta daitekeen intentsitatearen aldaketa neurtu balentzia-aldatzaileen eraginez. Beste batzuetan, berriz, lexikoian, hitzen sentimenduen balentziak bi eskaletan daude (positiboan eta negatiboan) hitz horiek duten polisemia tarteko eta, azkenik, eskala bera egokia ez den kasuak ere badaude. Hori horrela izanik, sentimendu lexikoi bat euskarara itzultzea erabaki dugu.

- Kalitatea. Gure helburua ahalik eta kalitate handieneko sentimenduen lexikoia sortzea da. Aurreko puntuan aipatu bezala, tresna eta baliabideek kalitatean eragin dezakete. Bestalde, ebaluatzeko modukoa den eta gerora hobetu daitekeen lexikoi bat sortu nahi dugu. Kalitatea eta bere ezaugarriak neurtzeko, aurretik dauden lexikoien antzeko

ezaugarriak dituzten lexikoi bat eratu nahi dugu. Ondorioz, sortutako lexikoiaren kalitatea neurtzeko, dagoen lexikoi baten antzekoa sortzeko beharra ikusi dugu.

Hurrengo urratsean, aurretik aipatutako baldintzak aintzat hartuta sentimenduen lexikoia nola sortu erabaki dugu. SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia jarraian azalduko dugun metodoari jarraituz itzultzeko erabakia hartu dugu. SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoia gaztelaniazko lexikoiko hitzari euskarazko ordaina emateko ere erabiltzea erabaki dugu. Honako abantailak ikusi ditugu, erabakia hartzerakoan:

- SO-CAL tresnako (Taboada *et al.*, 2011) lexikoien ezaugarriak. Hizkuntzako fenomeno ezberdinak kontuan hartzeko lexikoiko hitzen balentziak, sentimendua adierazten duten balioak, −5 eta +5 artean daude. Gure ustez, balentziak eta balentzien eskala egokiak dira guk sentimenduen analisian aztertu nahi ditugun alderdietarako, hain zuzen ere, balentzia-aldatzaileek eragiten dutena eskala horretan neur baitaiteke.
- Itzulpenak egiteko baliabideak. Lehen aipatu dugun bezala, sentimenduen lexikoia euskaraz sortzeko baliabideak lortzea zaila da eta daudenak dituzten ezaugarriengatik ez dira egokiak. Baina hiztegigintzan euskaran hainbat eta hainbat baliabide daude, eta hori dela eta, horiek erabiltzea egokia da. Besteen artean, *Elhuyar* hiztegia (Elhuyar, 2013) eta *Zehazki* hiztegia (Sarasola, 2005) sarean daude.
- Konparatzeko aukera eta ebaluazioa. Garatuko dugun sentimenduen lexikoiak beste sentimenduen lexikoien ezaugarriak baditu, elkarren artean konparatzeko aukera dago eta horrek lexikoia ebaluatzeko aukera ere ematen digu.

Ondoren, sentimenduen lexikoia garatzeko baliabideak eta tresnak bildu ditugu. Guztira bost baliabide edo tresna erabili ditugu: i) SO-CAL tresnaren gaztelaniako bertsioko lexikoia (Taboada *et al.*, 2011), ii) sarean dauden *Elhuyar* (Elhuyar, 2013) hiztegi eleanitza eta *Zehazki* (Sarasola, 2005) hiztegi

elebiduna, iii) SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko bertsioaren lexikoia, iv) Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) eta v) KWIC izeneko teknika:

- SO-CAL tresnaren gaztelaniako bertsioko lexikoia. Lexikoi hau izan da lanaren abiapuntua, itzuli den lexikoia baita. Guztira 4,880 sarrera ditu lexikoi honek, bost gramatika-kategorietan banatuta: izenak, adjektiboak, aditzak, adberbioak eta intentsifikatzaileak. 3.8 irudian, izenen zerrenda ikus daiteke eta bertako hitzek −5 eta +5 arteko sentimendu-balentzia dute.

Hitza	**Sentimendu-balentzia**
normalidad	2
imprudencia	-3
idiota	-3
envidia	-1
necesidad	-1
xenofobia	-4
hermosura	4
cumplimiento	3
violencia	-5
favorito	3

3.8 taula: SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko bertsioaren lexikoiaren zati bat.

- *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak. Gaztelaniazko SO-CAL lexikoia itzultzeko sarean dauden bi hiztegiak erabili ditugu: Elhuyar[5] eta Zehazki[6].

- Ingelesezko SO-CAL bertsioko lexikoia (Taboada *et al.*, 2011). Ingelesezko bertsioko lexikoiak 6.610 hitz ditu eta bost gramatika-kategorietakoak dira: izenak, adjektiboak, aditzak, adberbioak eta intentsifikatzaileak.

[5]Elhuyar hiztegiaren esteka: `https://hiztegiak.elhuyar.eus/eu_en`.
[6]Zehazki hiztegiaren esteka: `http://www.ehu.eus/ehg/zehazki/`.

3.9 irudian, lexikoiaren zati bat ikus daiteke. Euskarazko lehen bertsioko sarrerak hiztegi honetan agertzen diren ikusi dugu eta baita zer sentimendu-balentzia duten ere. Ondoren erabaki dugu hitzari zuen sentimenduaren balentzia mantendu edo ingelesekoa esleitu.

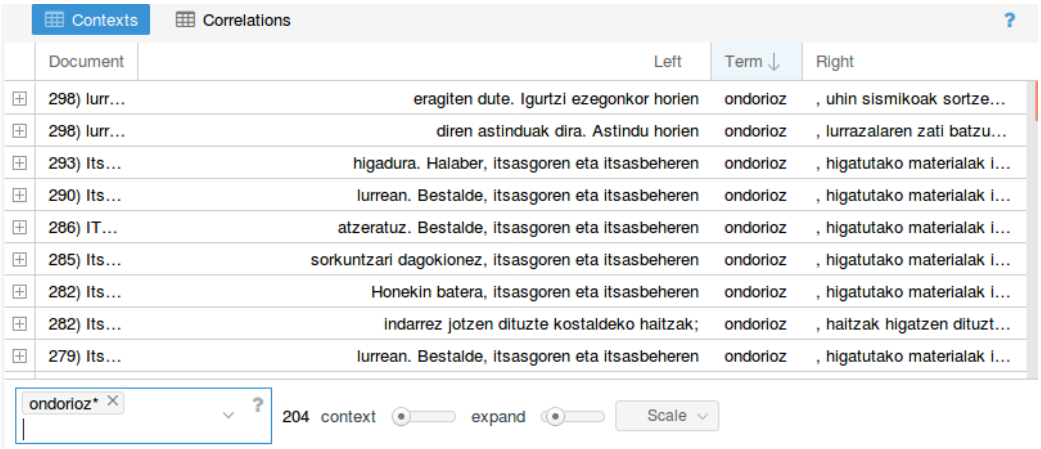
Izena	**Sentimendu-balentzia**
perfection	5
beauty	4
heroism	3
compassion	2
compromise	1
accommodation	-1
complication	-2
brutality	-3
atrocity	-4
monstrosity	-5

3.9 taula: SO-CAL tresnaren ingelesezko bertsioaren lexikoiaren zati bat.

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016). Itzulpen prozesuan erabili den beste baliabideetako bat izan da. Corpusak 240 iritzi-testu biltzen ditu sei domeinutakoak: kirola, eguraldia, politika, musika, zinema eta liburuak. Corpusa lexikoiko hitzen sentimenduen balentzia domeinuetara egokitzeko erabili da.

- KWIC teknika. Gaztelaniatik euskaratua izan den hitza corpusean bilatzeko eta hitzaren testuingurua ezagutzeko erabili da (ikus, 3.5 irudia). Modu horretan, gaztelaniatik euskarara itzulitako hitzari domeinuari loturiko sentimenduaren balentzia esleitu diogu.

Baliabideak eta tresnak bildu ostean, sentimenduen lexikoia itzuli eta sortu dugu. Itzulpen-prozesuan hiru urrats nagusi daude: i) gaztelaniazko lexikoia euskarara itzultzea, ii) euskarazko lexikoiaren garbiketa eta iii) euskarazko lexikoiaren ebaluazioa.



3.5 irudia: *Ondorioz* hitza eta bere testuinguruak.

Fenomenoa	GAZT	GAZT multzoka	EUS	ING	Azken balioa
F1	desacreditar	desacreditar −2	ospea_kendu −2 izena_kendu −2 sona_kendu −2	-	-
F2	atrofiar	atrofiar −1	atrofiatu −1	-	-
F3	amago	amago −1 cicatriz −2	seinale −1	-	-
F4	franquismo	franquismo −2	frankismo −2	-	−2
F5	correcto	acertado +3 correcto +3 decente +2	zuzen +3	right +1 correct +3	+3

3.10 taula: Itzulpen-prozesuaren azalpeneko adibideak.

Itzulpen-prozesuaren azalpena 3.10 taulako adibideetan oinarrituz egin dugu[7].

1- Itzulpena. Urrats nagusi honetan, SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskarara itzuli dugu.

   i) Itzulpen automatikoa gaztelaniatik euskarara. Lehenik eta behin, gaztelaniazko lexikoia automatikoki euskarara itzuli dugu *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak erabiliz. Gaztelaniazko hitz batek euskarazko ordain bat baino gehiago dituenean, ordain horietako bakoitzak hartu dugu aintzat, 3.10 tau-

[7]Euskarazko sentimenduen lexikoiaren sorkuntzari buruzko informazio osagarria 4.2.2 atalean dago.

lan ikus daitekeen moduan. Adibidez, F1 fenomenoan, *desacreditar* hitzaren euskarazko hiru ordain aintzat hartu ditugu: *ospea_kendu*, *izena_kendu* eta *sona_kendu*. Egoera horretan, gaztelaniazko hitzak duen balentzia; bere itzulpen-moduetako bakoitzak oinordekotzan hartu du.

ii) Iragazketa eta multzokatzea. Urrats honetan, euskaraz lortu diren ordainak iragazi eta euskarazko ordain berdinak multzokatu egin ditugu. Hau da, gaztelaniazko lexikoiko hitzei euskarazko ordaina ematean, errepikatuta dauden euskarazko ordainak batera jarri ditugu eta, beren ondoan, jatorrizko gaztelanizko hitzak eta gaztelaniazko hitz horien sentimendu-balentziak bildu ditugu.

Horren adibide dugu, 3.10 taulan, hirugarren zutabean, F3 fenomenoan, *seinale* hitza. *Seinale* ordainaren jatorrian gaztelaniako bi hitz daude: *amago* eta *cicatriz* eta horiek, euskarazko ordain bera dutenez (ordaina errepikatuta dagoenez), multzo berean jarri ditugu, baita bere jatorrizko gaztelaniazko hitzak eta gaztelaniazko hitzen sentimendu-balentziak ere ere. F5 fenomenoko *zuzen* itzulpenaren jatorrian, berriz, gaztelaniako *acertado*, *correcto* eta *decente* daude eta euskarazko ordainak eta beren jatorrizko gaztelaniazko hitzak bildu ditugu.

iii) *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrerak diren ordainak bakarrik aukeratu. Ordainak bananbanan aztertu ditugu eta aipaturiko bi hiztegietako sarrerak diren aztertu dugu. Lexikoia itzuli eta sortzerakoan, hiztegiko sarrera diren ordainak bakarrik hartuko ditugu aintzakotzat.

Esaterako, 3.10 taulan, F1 fenomenoan, *sona_kendu*, *izena_kendu* eta *ospea_kendu* ez dira hiztegi horietako sarrerak eta horrenbestez, ez ditugu aintzat hartu. *Atrofiatu*, aldiz, hiztegiko sarrera da eta kontuan hartu dugu.

iv) Sentimenduen balentziaren aukeraketa. Hiztegi bateko sarrerak ez diren ordainak kendu ostean, geratzen diren ordainetan, euskarazko ordainari dagokion gaztelaniazko hitza eta bere sentimendu-balentzia aukeratu dizkiogu.

Euskarazko ordainari dagokion sentimenduen balentzia eta gaztelaniako esanahia aukeratzerakoan, prozedura honi jarraitu diogu:

- Euskarazko ordainak itzulpen (eta balentzia) bakarra badu jatorrian gaztelaniaz, itzulpena eta hari dagokion balentzia hautatu dugu. Hori gertatu da 3.10 taulako F2 eta F4an.
- Euskarazko ordainak jatorrian hainbat gaztelaniazko hitz eta haien balentziak baditu, hitzari esleituko zaion sentimenduen balentzia (eta gaztelaniazko esanahia) Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarrituta egin dugu. Hori izan da 3.10 taulako F3 eta F5en kasua.
- Kasuak egon dira non lortu den ordaina ez den corpusean agertzen eta jatorrian gaztelaniazko hainbat dituen. Egoera horretan, ordainak duen esanahirik erabiliena edo ohikoena hobetsi da.

2- Garbiketa. Bigarren urrats honetan, sentimenduen lexikoiaren lehen bertsioa garbitu egin dugu eta bigarren bertsioa sortu dugu. Zehazki esanda, lexikoia domeinu jakin batzuetara moldatu dugu eta SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoia baliatuta aberastu eta zuzenketak egin ditugu.

v) Domeinu- eta corpus-moldaketa. Bigarren bertsioa sortzeko helburuz, lexikoiko hitzek Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) duten agerpena kontuan hartu dugu.

Adibidez, 3.10 taulan, F2 fenomenoan, *atrofiatu* Osasungintzako kontzeptu bat da eta domeinu hori ez dagoenez gure corpusean, kontzeptu hori lexikoitik ezabatu dugu.

vi) Lexikoiko sarrera bakoitza berraztertzea eta hobetzea. Lexikoiko euskarazko hitz bakoitzaren ingeleseko ordaina aurkitu dugu *Elhuyar* (Elhuyar, 2013) hiztegia erabilita eta, ondoren, hitz hori SO-CALen ingelesezko bertsioko lexikoian agertzen den aztertu dugu. Euskarazko sarrera ingelesezko lexikoian agertzen bada, euskarazko sarrerari ingelesezko sarreraren sentimendu-balentzia esleitu diogu. Euskarazko sarrera ingelesezko bertsioko lexikoian

ez bada agertzen, aldiz, sarrera hori kendu edo mantendu egin dugu hitzaren egokitasuna aintzat hartuz.

Esaterako, 3.10 taulan, F5 fenomenoan, euskarazko *zuzen* hitzaren baliokideak SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoian *correct* eta *right* dira eta sentimendu-balentzia ezberdinak dituzte.

3- Lexikoiaren ebaluazioa[8]. Azken urratsean, garatutako euskarazko sentimenduen lexikoia ebaluatu dugu. Horretarako, lehendabizi, bi pertsonen anotaziotik urre-patroi bat sortu da eta ataza batean urre-patroiak eta lexikoiak ematen dituzten emaitzak alderatu ditugu.

vii) Corpuseko maiztasunik handieneko hitzen urre-patroiaren anotazioa.

Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) maiztasunik handieneko 400 hitz erauzi ditugu (gramatika-kategoria bakoitzeko 100 hitz), *Analhitza* tresna (Otegi *et al.*, 2017) erabiliz, eta bi pertsonek hitz horiei −5 eta +5 arteko sentimenduen balentzia esleitu diete.

3.11 taulan ikusten den moduan, anotatzaile batek *ahul* hitzari −3 sentimendu balentzia esleitu dio eta *argi* hitzari +5. Ondoren, bien anotazioan oinarrituz, urre-patroia eratu ditugu.

viii) Urre-patroian erabilitako 400 hitzei euskarazko sentimenduen lexikoiaren bigarren bertsioa aplikatu zaie.

ix) Urre-patroiak eta lexikoiak emandako sentimenduen balentzia esleipenak alderatu ditugu Pearson korrelazioa (Benesty *et al.*, 2009) erabiliz. Pearson korrelazioa erabiliz, adostasun neurketa bi modutan egin da:

* Pearson 1. Neurketa honetan, 400 hitzeko zerrendan, bi pertsonek etiketatutako hitzak bakarrik hartu ditugu kontuan. Hitz bat pertsona bakar batek etiketatuta badago, hitz hori ez da kontuan hartu.

[8]Euskarazko sentimendu lexikoiaren ebaluazioari buruzko informazio gehiago 4.2.3 atalean dago.

Zenbakia	Adjektiboa	Orientazio semantikoa	Sentimendu-Balentzia
1	ageri		
2	ahul	Negatiboa	3
3	antisozial	Negatiboa	5
4	apur	Negatiboa	1
5	ar		
6	argi	Positiboa	3
7	aspergarri	Negatiboa	3
8	ausart	Positiboa	4
9	bakar	Positiboa	5
10	bakoitz		

3.11 taula: 400 hitzeko zerrendaren zati bat etiketatzaile batek bertako hitzei sentimendu-balentzia esleituta.

* Pearson 2. Neurketa honetan, zerrendako 400 hitzak erabili ditugu. Hitzen bat pertsona batek edo bik etiketatu gabe baldin badago 0 balentzia esleitu zaio, hitz hori kontuan hartu ahal izateko.

3.1.3. Lexikoian oinarritutako dokumentu mailako euskarazko sentimenduen sailkatzailea

Lan honen hirugarren urrats nagusia lexikoian oinarritutako dokumentu mailako euskarazko sentimenduen sailkapena garatzea izan da. Helburu horrekin, SO-CAL tresnaren[9] (Taboada *et al.*, 2011) euskarazko lehen bertsioa garatu nahi izan dugu, euskarazko sentimenduen sailkapena tresna horretan oinarritu baita. Hori dela eta, *Eustagger* tresna (Aduriz *et al.*, 2003) (3.1.3.1 atala) eta *SentiTegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) (3.1.3.2 atala) erabili ditugu euskarazko bertsioa sortzeko.

[9]SO-CAL tresnaren ingelesezko bertsioaren ezaugarriak 4.3.1 atalean deskribatuta daude.

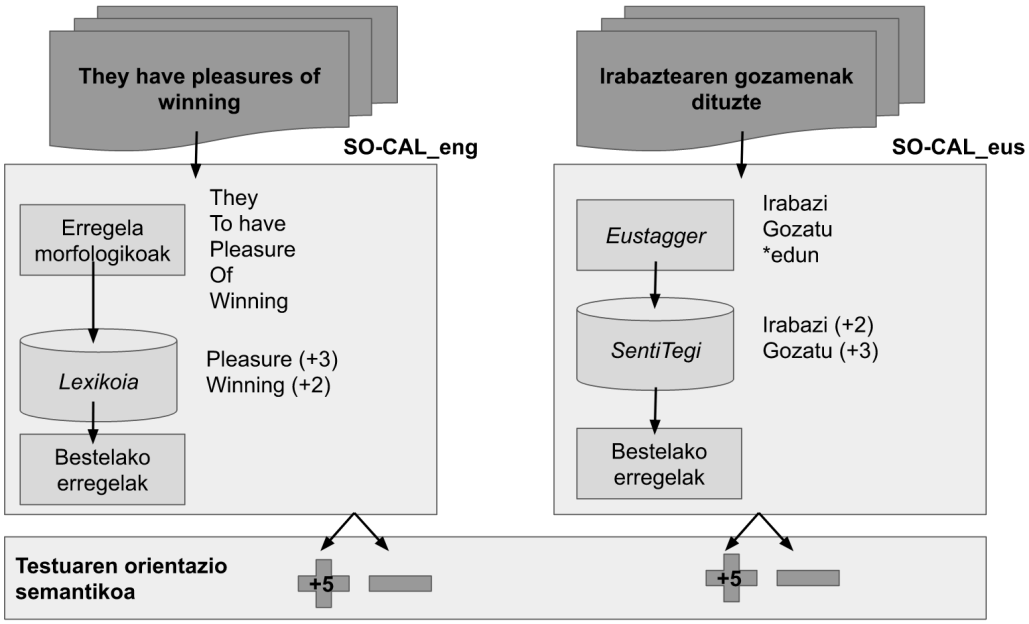
3.1.3.1. *Eustagger* tresna integratzea

Lexikoian oinarritutako sentimenduen sailkatzailea den SO-CALen (Taboada *et al.*, 2011) ingelesezko bertsioan, lematizatutako sentimendu-balentziadun hitzen lexikoia eta flexioari lotutako zenbait erregela daude. Baina euskaran, hori baliagarria izan arren, ez da modurik egokiena lexikoian oinarritutako sentimenduen sailkatzaile bat garatzeko; izan ere, tresnak testua prozesatzekotan flexioari lotutako erregela asko eduki beharko lituzke eta hori ez da bideragarria lan konplexua delako. Beste aukera bat flexioak eragindako hitz baten aldaera guztiak lexikoian sartzea izango litzateke baina, kasu horretan, lana konplexuagoa izango litzateke eta denbora gehiago eskatuko luke. Hori da euskarazko sentimenduen sailkatzailea sortzerakoan izan dugun zailtasuna.

Zailtasun horri aurre egiteko konponbidea, lexikoiak testuko hitzei sentimenduen balentzia esleitu aurretik, testuak lematizatzeko *Eustagger* tresna (Alegria *et al.*, 2002) integratzea erabaki dugu. Modu horretan, lehenik eta behin, tresnak testua lematizatuko du eta ondoren, tresnak lematizatutako hitz hori tresnaren lexikoietan badagoen aztertuko du eta baldin badago, testuko hitzari sentimenduen balentzia esleituko dio. Beraz, lematizatzaile bat tresnan integratuta, tresna gauza izango da lehenik eta behin, testua lematizatzeko; ondoren, lematizatutako hitza lexikoian dagoen aztertzeko eta, azkenik, baldin eta hitza lexikoian badago, hitzari sentimenduen balentzia esleitzeko.

Egindako aldaketa eta ingelesezko eta euskarazko SO-CAL tresnaren egituren alderaketa ageri dira, 3.6 irudian.

Ikusten den moduan, ingelesezkoan, erregela morfologikoak daude eta, modu horretan, testuetako hitzak lema bilakatzen dituzte. Euskarazkoan, aldiz, aberastasun morfologikoa tarteko, *Eustagger* lematizatzailea (Alegria *et al.*, 2002) integratu da helburu bera lortzeko asmoz. Ondoren, sentimenduen lexikoiak datoz eta lematizatutako hitzak lexikoian baldin badaude, tresnak hitz horiei sentimendu-balentzia esleitzen die. Azkenik, beste zenbait erregela daude ingeleserako, euskararako ere berak direnak (orientazio semantiko negatibodun hitzei pisu gehiago esleitzea, galde-perpausetako hitzei sentimendu balentzia ez esleitzea, etab. egiten dituztenak) eta bi hizkuntzetarako baliagarriak dira.



3.6 irudia: SO-CAL tresnaren ingelesezko eta euskarazko bertsioen egituren alderaketa.

3.1.3.2. *Sentitegi* sentimenduen lexikoia integratzea

Iritzi-testuak lematizatzeko *Eustagger* tresna (Alegria *et al.*, 2002) integratu ondoren, euskarazko sentimenduen lexikoia gauza da testuak prozesatzeko; ez, ordea, hitzei sentimendu-balentzia esleitzeko. Horretarako, tesi-lan honetan garatutako metodologiaren ondorioz sortutako *Sentitegi* euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018) integratu dugu tresna horretan.

Tresnak berak modulu bat du sentimenduen lexikoiei bideratuta; hala ingelesezko lexikoiak kendu eta euskarazkoak gehitu ditugu. Tresnak funtziona dezan, beharrezkoa da lexikoiak *txt* formatuan egotea eta bertan, lerro bakoitzeko sarrera bat eta bere sentimendu balentzia egotea. Halaber, gramatika-kategoria bakoitzak bere fitxategia eduki behar du. Izan ere, *Eustagger* lematizatzaileak (Alegria *et al.*, 2002) testuko hitza lematizatu eta bere gramatika-kategoria identifikatuko du eta ondoren, tresnak hitz hori bere gramatika-kategoriako lexikoian bilatuko du. Guk lau gramatika-kategorietako lexikoiak integratu ditugu: izenena, adjektiboena, adberbioena

eta aditzena. Egin dugun aldaketa 3.12 taulan ikus daiteke.

Ingelesa		**Euskara**	
Thriving	+3	Bikain	+5
Record-setting	+3	Maximo	+1
Leading	+3	Orokor	+3
Industrious	+3	Min	-2
Best-selling	+3	Polit	+4
Upset	+3	Gogor	-1
Clean	+2	Ahul	-2
Capacious	+2	Behar	-1
Cogent	+2	Txar	-3
Confident	+2	Zail	-2

3.12 taula: Ingelesezko eta euskarazko sentimenduen lexikoien zati bat.

Ezkerrean, ingelesezko lexikoia ageri da, hitz-zerrenda batekin eta beren sentimenduen balentziekin. Eskuinean, berriz, euskarazko lexikoiaren zati bat dago eta horrekin ordezkatu dugu ingelesezkoa. Bi lexikoiak *txt* formatuan dauden fitxategian daude eta *SentiTegi* lexikoia integratzeko, fitxategi bat bestearekin ordezkatu dugu.

3.1.3.3. Sentimenduen sailkatzailearen ebaluazioa

Sentimenduen sailkatzailean *Eustagger* lematizatzailea (Alegria *et al.*, 2002) eta *Sentitegi* sentimendu lexikoia (Alkorta *et al.*, 2018) integratu ondoren, sailkatzailea bera ebaluatu dugu[10].

Tresna ebaluatzeko Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 48 iritzi-testu erabili ditugu, sentimenduen analisiarekin lotutako azterketa eta lanketa ez delako iritzi-testu horietan oinarrituta egin. Metodologia atal honetan aipatu bezala, corpuseko iritzi-testuen orientazio semantikoa etiketatuta dago eta sentimenduen sailkatzaileak ematen dituen emaitzak horiekin konparatu ditugu, 3.13 taulan ikusten den moduan.

[10]Euskarazko sentimenduen sailkatzailearen ebaluazioari buruzko informazio osagarria 4.3.3 atalean dago.

Iritzi-testua	**Orientazio semantikoa**	**Sailkatzaileak esleitutako sentimendu-balentzia**
EGU32	Positiboa	0,80 (Positiboa)
LIB30	Positiboa	1,09 (Positiboa)
ZIN39	Negatiboa	0,57 (Positiboa)

3.13 taula: Euskarazko sentimenduen sailkatzailearen ebaluazioa.

Tresnak emaitzak zenbakiz ematen ditu eta ebaluatzeko orduan, zenbakizko emaitza horretan zeinua hartu dugu kontuan, hau da, emaitza horiek zeinu positiboa edo negatiboa duten.

3.2 Balentzia-aldatzaileen identifikazioa

Ikerketa aurrera eramateko oinarrizko balibideak eta tresnak garatu ondoren, euskaran dauden testuinguruko balentzia-aldatzaileak identifikatzen hasi gara. Urratsez urrats, hizkuntza mailetako balentzia-aldatzaileak zein diren eta nolako eragina duten aztertu dugu.

3.2.1. Fonologia eta morfologia: bustidura adierazkorra eta hizkiak

Lehenik eta behin, fonologia eta morfologia mailako balentzia-aldatzaileak identifikatzen hasi gara. Helburu horrekin, lehen urratsean, euskararen morfologia eta fonologia lantzen duten hainbat iturri bibliografikoren bilaketa egin dugu eta euskarak dituen hizkien (aurrizkiak, artizkiak eta atzizkiak) zerrenda bat osatu dugu. Zerrenda horretan, hizki bakoitza zer gramatika-kategorirekin agertzen den, bere esanahi semantikoa zein den eta adibide bat jarri dugu. 3.14 taulako zerrenda[11] osatzeko erabili ditugun iturri bibliografikoen artean Mujika (1982), Euskara Institutua (2011) eta Oñederra (1990) aipa ditzakegu.

Hizkia	**Izena**	**Adjektiboa**	**Aditza**	**Sailkapen semantikoa**	**Adibidea**
-zale	X			Zaletasuna	Ardozale
ez-	X	X	X	Ezezkotasuna	Ezberdin
-z- → -x-				Hurbiltasuna/ txikitasuna	Gazte → gaxte zahar → xahar

3.14 taula: Hizkiei buruz bildutako informazioa.

Iturri bibliografikoetatik bildutako hizki eta bustidura adierazkorren adibideak ageri dira 3.14 taulan. Lehenengo hizkia *-zale* atzizkia da, izenekin agertzen da eta zaletasuna adierazten du. Bigarrena, berriz, aurrizki bat da (*ez-*); izen, adjektibo nahiz aditzekin ager daiteke eta ezezkotasuna adierazten du. Bukatzeko, azken adibidean bustidura adierazkor[12] bat dago: [z] bilakatzea

[11] 3.14 taulako zerrenda lanaren zati bat da.

[12] Bustidura adierazkorra erabilera ez-estandarrean ohikoena den arren, bere zenbait kasu aurkitu ditugu corpuseko testu formaletan.

[x]. Bustidura adierazkor guztiak bezala, ez dago murriztapen gramatikalik eta hurbiltasuna edo txikitasuna adierazteko erabiltzen da. Prozedura bera jarraitu dugu beste hizki guztiekin.

Hurrengo urratsean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 192 iritzi-testu hartu ditugu eta *Analhitza* tresna (Otegi *et al.*, 2017) bidez aztertu ditugu. Tresna pasa ondoren, testuetako hitz guztiak maiztasunaren arabera zerrendatzea lortu ditugu. 3.15 Taulan, *Analhitza* tresnak (Otegi *et al.*, 2017) emandako hitz-zerrenda ageri da eta maiztasunean oinarritzen da. Ikusten denez, *-txo* hizkia daraman *zertxobait* hitza hamaika aldiz agertzen da corpusean.

Hitza	**Instantzia-kopurua**
zati	11
zertxobait	11
zerua	9

3.15 taula: Corpuseko hitzen zerrendaren zati bat maiztasunean oinarrituta.

Ondoren, euskarak dituen hizkiak eta bustidura adierazkorra egiteko moduak corpusean nola agertzen diren aztertu ditugu. Horretarako atal honetako lehen urratsean sortu dugun zerrenda erabili dugu eta, hori oinarritzat hartuz, corpusean agertzen diren hizki eta bustidura adierazkorra egiteko modu guztiekin zerrenda bat osatu dugu. Zerrendan, banan-banan, beren ezaugarriak deskribatu ditugu 3.16 taulan ikusten den moduan.

Adb.	**Hitza**	**Hizkia**	**Balentzia**	**Balentzian eragina**	**Semantika**	**Iz.**	**Adj.**	**Adi.**
(1)	istorio							
(2)	erasotxo	-txo	−2	Ahuldu	Txikitasuna	X	X	
(3)	kuttun	[t] → [tt]	+2	Indartu	Hurbiltasuna			

3.16 taula: Corpusean agertutako hizkien eta bustidura adierazkoren azterketaren lagin bat

3.16 taulan, (1) adibidean, bustidura adierazkorrik edo hizkirik ez da agertzen. (2) adibidean, *-txo* atzizkia ageri da eta honek *eraso* hitzaren sentimendu balentzia (−2) ahultzen du. (3) adibidean, azkenik, bustidura adierazkorra

dago [t] [tt] bilakatu baita eta aldaketa horrek adierazkortasuna indartu duenez, duen sentimenduaren balentzia are indartsuagoa bilakatzen da. Guztira 59 atzizki, 7 aurrizki eta 13 bustidura adierazkorraren instantzia aurkitu ditugu Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) eta horiekin guztiekin prozedura bera jarraitu dugu.

3.2.2. Sintaxi maila: ezeztapen-markak

Hitzen balentzian eragin dezakeen hizkuntzaren beste alderdi bat sintaxia da. Baina kasu honetan, balentzia aldatzearen ondorioa jasoko duena ez da hitz bat baizik eta hitz multzo batek osaturiko sintagma edo esaldi bat izan daiteke. Hau da, morfologiako eta fonologiako balentzia-aldatzaileek hitzari eragiten dioten moduan, sintaxiko balentzia-aldatzaileek hitzi ez ezik, sintagma edota esaldiari ere eragin diezaiokete.

Sintaxian, ezeztapen-markek esaldien sentimendu balentzia nola aldatzen duten aztertzeko asmoz, lehenik eta behin, euskararen ezeztapen-marken zerrenda osatu dugu eta zerrenda hori Altuna *et al.*en (2017) eta Altuna *et al.*en (1985) lanetan oinarrituta osatu dugu. Honako hauek izan dira ikerketarako erabili ditugun ezeztapen-markak: *ez*, *ezin*, *gabe*, *ezik*, *salbu*, *ezta* eta *ezean*.

Hurrengo urratsean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 96 testutan, ezeztapen-markaren bat duten esaldiak bilatu ditugu eta horiekin esaldi-zerrenda bat osatu dugu, 3.17 taulan ikus daitekeen moduan. Ez ditugu testu guztiak erabili, Altuna *et al.*en (2017) lanean ageri diren ezeztapen-marka gehienak lortu ditugulako eta testu gehiago aztertu arren, ezeztapen-marka berririk ez delako agertu. Guztira 359 ezeztapen-marken instantzia lortu dugu. *Analhitza* tresnak (Otegi *et al.*, 2017) emandako datuen arabera, 359 ezeztapen-marka horiek 320 esalditan banatuta daude. Horrenbestez, badaude esaldiak non ezeztapen-marka bat baino gehiago dauden. Bestalde, lortu dugun ezeztapen-marken azpicorpus honek 5.515 hitz ditu.

Ondoren, 320 esaldi horietako hitzei nahiz esaldi osoei sentimendu balentzia esleitu diegu. Hori egiteko, hitzei kategoria gramatikala edota marka lexikala esleitzen dien *Eustagger* tresnari (Aduriz *et al.*, 2003) *Sentitegi* (Alkorta *et al.*, 2018) sentimenduen hiztegiko hitzak gehitu dizkiogu; lexikoian hitzak

TESTUA	ITEMA
EGU23	Ez da espero euri asko egitea, baina giroa hezea eta freskoa izango da.
EGU23	Asko ez du egingo, baina zerbait bai.
EGU24	Zumaian ez du elurrik egingo baina elur-kota 400-500 metrora jaitsiko da, beraz kontuz errepideetan.
EGU24	Oraintxe bertan ematen du Frantzia iparraldera iritsiko dela eta hortik haize indartsua sortuko du gure lurraldean, baina ez da izango lehengokoa bezain indartsua.
EGU25	Abuztua hasi berri dugun honetan, ez dugu urte sasoi honetarako espero izan ohi dugun eguraldirik izango.
EGU25	Goizean berriz aterri mantenduko du eta ez du ia euririk egingo.

3.17 taula: Ezeztapen-markak aztertzeko sortutako azpicorpusaren zati bat.

kategoria gramatikalean oinarrituta banatuta baitaude.

(32) Pogostkinak ezin [hobeki]$_{+2}$ atera zituen. (MUS20)

(33) [Irabazi]$_{+2}$ ezinik jarraitzen du Eibarrek. (KIR17)

(34) Ikuspuntu [politikotik]$_{-1}$ ez ezik, [ekonomikotik]$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(32), (33) eta (34) adibideetan, etiketatzaileak eta hari gehitutako sentimendu lexikoiak lau hitzei esleitu die sentimendu balentzia: *hobeki* adjektiboari, *irabazi* aditzari, *politiko* adjektiboari eta *ekonomiko* adjektiboari. Lehen bi hitzei +2 balentzia esleitu die eta adjektiboei −1 eta −3, hurrenez hurren. Modu honetan, ezeztapen-markak eta sentimendu balentziadun hitzak edukita, analisi linguistikoa egiteko ingurunea prest utzi dugu. Eta horrenbestez, analisi linguistikoa egitea da hurrengo urratsa.

Ezeztapen-marken analisi linguistikoa egiterakoan, aintzat hartu duguna izan

da ezeztapen-markak eragiten duen aldaketa orientazio semantikoan eta zehazki, sentimendu-balentzian. Hots, ezeztapen-markak sentimendu-balentziadun hitz eta hitz-multzoetan, haren irismenean eta ondorioz, baita esaldiko sentimendu-balentzian ere, zer ondorio uzten duen aztertu nahi izan dugu. Azterketa eskuz egin dugu eta jarraian azaltzen den prozedura jarraitu dugu.

Azterketan, corpuseko esaldiak banan-banan aztertu ditugu eskuz. Esaldietako ezeztapen-markak identifikatu ondoren, haren irismena zein den adierazi dugu. Azkenik, irismen-eremuko sentimendu-balentzian nahiz esaldi osokoan utzi duen ondorioa aintzat hartuta, esaldia bera eta haren ezeztapen-marka multzokatu egin ditugu.

Azter ditzagun, berriz, aurretik aipaturiko hiru adibideak.

(35) Pogostkinak [ezin hobeki$_{+2}$] atera zituen. (MUS20)

(36) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)

(37) [Ikuspuntu politikotik$_{-1}$ ez ezik], [ekonomikotik$_{+3}$] ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(35) adibidearen kasuan, ezeztapen-marka *ezin* da eta bere irismena *hobeki* adjektiboraino iristen da. Irismen-eremuan hitz bakarra dagoenez eta balentziaduna denez, irismen-eremuaren balentzia +2 da, baita esaldiarena ere. Kasu honetan, ezeztapen-markaren eragina aztertuta, ezeztapen-markak *hobeki* balentziadun hitza indartu egiten duela ohartu gara, intentsitatez *ezin hobeki hobeki* baino indartsuagoa baita. Ondorioz, esaldia bera eta ezeztapen-marka balentzia indartzen duten taldean sailkatu ditugu.

(36) adibidean, aldiz, nahiz eta ezeztapen-marka bera den (*ezin*), ikusi dugu haren eragina ezberdina dela. Hau da, *ezin* ezeztapen-markaren irismena *irabazi* da, +2 balentzia duena eta honi indarra ahultzen dio. Hots, *irabazi ezinik irabazi* baino ahulagoa da sentimendu-balentziari dagokionez. Ondorioz, ezeztapen-marka eta esaldia bera beste multzo batean sartu ditugu, ahultze bat egon delako. Bi adibide hauekin ikusi den moduan, posible da ezeztapen-marka berak bi funtzio izatea.

Azkenik, (37) adibidean, *ez ezik* ezeztapen-markak −1 sentimendu-balentzia

duen *ikuspuntu politikotik* hitz multzoa ezeztatzen du, baina sentimendu-analisiaren ikuspegitik, hitz multzoaren sentimendu-balentzian ez da aldaketarik gertatzen. Izan ere, kasu honetan, ezeztapena ondotik informazioa gehitzeko erabiltzen da eta ez ezeztapenaren irismena den *ikuspuntu politikotik* ezeztatzeko. Beraz, *ez ezik* egiturak bere irismenaren parte diren hitzen sentimendu-balentzian ez du eragiten.

Erregeletan oinarritzen diren SO-CAL moduko (Taboada *et al.*, 2011) tresnetan ezeztapenari buruzko informazioa gehitu nahi bada, beharrezkoa da ezeztapen-markak eta haren irismena identifikatzea eta, horregatik, ezeztapen-markak eta beren irismena identifikatzeko erregelak sortu eta ebaluatu ditugu[13], 3.18 taulan ikus daitekeen moduan.

3.18 taulako, erregelak deskribatu aurretik, erregela horiek nola sortu diren azalduko dugu jarraian, (38) eta (39) adibideen bitartez. Bertan bosgarren erregelako bi aldaerak landuko ditugu. (38) adibidean, *ezin* ezeztapen-markaren ondoren, aditz laguntzailea (*da*) eta aditz nagusia (*baztertu*) ageri dira eta, azkenik, izen-sintagma bat den subjektua (*ekaitz zaparradaren bat izatea*). Beraz, esaldi honetarako lortutako egitura honelakoa izango litzateke: *ezin* + aditz laguntzailea + aditz nagusia + IS. Dena kako artean dago eta horrek ezeztapen-markaren irismena zein den adierazten du.

(38) (...) [ezin da baztertu$_{-1}$ ekaitz zaparradaren bat izatea.] (EGU35)

(39) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek (...) (KIR22)

(39) adibidea aztertzen badugu, ikusiko dugu *ezin* ezeztapen-markaren aurretik *irabazi* aditz nagusia agertzen dela eta bera dela ezeztapen-markaren irismena. Corpusean ageri diren *ezin* ezeztapen-markaren beste 36 instantziekin prozedura bera jarraitu ondoren, 3.18 irudiko (1) eta (5) adibideetan azaltzen diren erregelak lortu ditugu.

> Bildu den informazioetako bat erregela bakoitzak duen egitura sintaktikoa da. 3.18 taulako erregeletako egiturek honelako ezaugarriak di-

[13]Ezeztapen-markak eta beren irismena identifikatzeko *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako eta ebaluatutako erregelei buruzko informazio gehiago 5.2.3 atalean dago.

Adb.	Ezeztapen-marka	Erregelen egitura
(1)	ezin	PM *ezin* + [adjektiboa/adberbioa] (+ atzizki konp.) PM
(5)	ezin	PM [(IS) + aditza + *ezin*] PM PM [*ezin* (+ ad. lag.) (+ IS) + aditza (+ IS)] PM
(10)	ez	PM [IS] *ez ezik* PM

3.18 taula: Sentimendu-balentzian eragin ezberdinak dituzten ezeztapenak identifikatzeko proposatu diren erregeletako batzuk.

tuzte: kakoek [] ezeztapen-markaren irismena adierazten dute, parentesiek () sintagma horren agerpena hautazkoa dela adierazten dute. *Letra etzanak* ezeztapen-marka edo egitura lexikalizatuen osagaiak adierazten dituzte eta barrak /, azkenik, ezeztapen-markaren irismenean, aukeran aipaturiko osagai ezberdinek daudela adierazten du. Erregeletan ageri diren beste elementuak hitz-multzo eta gramatika-kategoriei dagozkie. ISk eta ASk izen- eta aditz-sintagma adierazten dute, hurrenez hurren. Erregeletan ageri diren beste elementuak dira adjektiboa, adberbioa, aditza, aditz laguntzailea eta sintagma. Bukatzeko, bada beste elementu bat erregeletan azaltzen dena eta hori puntuazio-markaren murriztapena (PM) da. Murriztapen honen helburua erregeletan ezeztapen-marken irismena esaldi barnean kokatzea eta beste esaldietako elementuak ez hartzea da. Izan ere, irismena ezeztapen-markaren aurrean nahiz atzean ager liteke eta aurreko edo ondorengo esaldietako hitzak ezeztapen-markaren irismeneko hitz moduan tratatzeko arriskua ikusten dugunez, erregelei puntuazio-marka murriztapena gehitzea erabaki dugu. Egitura lexikalizatuen kasuan, ez dago horrelako murriztapenik. Egitura lexikalizatuek egitura egonkorra dutenez (errepikatzen diren hitz-multzoak direnez) ez da beharrezkoa. Egitura lexikalizatuaren adibide bat "baino ez". Bi hitzak elkarrekin uste baino maiztasun handiagoz azaltzen dira eta bi hitzen baturak berezko esanahi bat sortzen du, "bakarrik" hitzak duen esanahiaren parekoa dena.

Aurreko urratsetan analisi linguistikoan bildutako informazioa modu egituratuan antolatu ondoren (3.18 taulan ikusten den moduan), ezeztapen-markak eta beren irismena identifikatzeko erregelak Karlsson *et al.*en (1995) *Mu-*

rriztapen Gramatika hurbilpenean sortu eta ebaluatu ditugu. 3.7 irudian erregelak nolakoak diren ikus daiteke[14]. Erregela horiek honela osatuta daude:

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT PUNT_GALD PUNT_ESKL
    PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = ``ez'';
LIST BESTERIK = ``beste'';

# (2)
# besterik ez egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)    (1C EZ);

(...)
```

3.7 irudia: Ezeztapen-markak eta beren irismen-eremua identifikatzeko erregelen adibide bat *Murriztapen Gramatika* (Karlsson *et al.*, 1995) ingurunean.

- LIST zerrendak. Bertan hitz jakinak edo bestelako elementuak, puntuazio-markak adibidez, zerrendatzen dira. 3.7 irudian, puntuazio-markak (LIST_PUNTUAZIOA) eta ezeztapen-markak (LIST_EZ, LIST_BESTERIK) zerrendatu ditugu.

- MAP komandoa (islapen-erregelak). Hauen bidez erregeletan adierazitako egiturak corpusean bilatzen dira eta aurkitzen badira, etiketa bat esleitzen zaio corpusean bilatu den egiturari. Ezaugarri hauetazz osatuta dago erregela:

  - Esleituko den etiketa. Etiketak aurretik ! ikurra du, 3.7 irudian.
  - Etiketa esleituko zaion hitza. Adibidez, *!besterikezHAS* etiketa determinatzaile bati (DET) esleituko zaio.
  - Erregelaren baldintzak. Esaterako, *!besterikezHAS* etiketa esleitzeko, 0 posizioan BESTERIK zerrendak, hau da, *beste* hitzak, (0C BESTERIK) egon behar du eta 1 posizioan EZ zerrendak (kasu honetan, *ez* ezeztapen-markak), erregelak funtziona dezan.

```
"<,>""<PUNT_KOMA>"
        PUNT_KOMA
"<batere>"
          "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG %SINT
"<harrotu>"   S:278/0
          "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG
             %ADIKAT S:278 !gabeAUR
"<gabe>"   S:141/0
          "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> %SINT S:141 !
             gabe
"<$.>""<PUNT_PUNT>"
        PUNT_PUNT
```

3.8 irudia: *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelek esleitutako ! etiketa.

*Murriztapen Gramatika*n (Karlsson *et al.*, 1995) idatzirik erregela bakoitzak bere etiketa uzten du corpuseko lerroan. 3.8 irudian, esaterako, Murriztapen Gramatikan sortu ditugun erregelek bi hitzi etiketa ezarri diete (*!gabeAUR* etiketa *harrotu* hitzean eta *!gabe* etiketa *gabe* hitzean).

Guk sorturiko erregelek corpusean ! ikurra duen etiketak esleitzen dituztela begiztatu ondoren, erregela horiek ebaluatu ditugu.

Ebaluazioa egiteko, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 48 testu erabili ditugu. 48 testu horiek prozesatu egin behar izan ditugu *Murriztapen Gramatika*ren ingurunera (Karlsson *et al.*, 1995) ekartzeko eta horretarako *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) oinarritzen den euskarazko desanbiguatzaile morfosintaktikoa erabili dugu (Aduriz *et al.*, 1997).

*Murriztapen Gramatika*ren (Karlsson *et al.*, 1995) ingurunean sorturiko 75 erregelak corpusaren test zatiko 144 testuetan aplikatu ondoren, erregela horiek ebaluatu egin ditugu. Lehenik eta behin, bi personek ebaluazioa egiteko duten adostasuna neurtu dugu corpusaren zati txiki bat erabiliz. Adostasuna neurtzeko kappa neurria (Cohen, 1960) erabili da. Emaitzek erakusten dutenez, kappa neurria 0,60koa da. Landis eta Kochen (1977) arabera, *neurrizko adostasuna* da. Etiketatzaileen arteko adostasuna neurtu ondoren, etiketatzaile batek ebaluazio osoa egin du.

[14]Ezeztapen-markak eta beren irismena identifikatzeko sortutako erregela guztiak A eranskinean daude zerrendatuta.

Ebaluatzeko, erregelek corpusean esleitutako etiketak ebaluatu ditu etiketatzaileak. Etiketatzaileak hiru etiketa erabili ditu: ETIK_ONDO, erregelaren etiketa zuzena denean; ETIK_FALTA, corpuseko hitzak etiketa behar zuenean eta erregelak jarri ez dionean eta, azkenik, ETIK_GAIZKI, erregelak corpuseko hitzari etiketa jarri dionean eta hitzak etiketa behar ez duenean.

Ebaluazioaren adibide bat ikus daiteke, 3.9 irudian. Bertan hiru hitz daude eta hirurek egon beharko lukete etiketatuta. Hiruretatik bi (*harrotu* eta *gabe*) etiketatuta daude eta ETIK_ONDO etiketa jarri zaie, lerroen hasieretan. Baina, hori ez da hala *batere* hitzaren kasuan. Nahiz eta ezeztapen-markaren irismen-eremuko parte izan, *harrotu* aditzari eragiten diolako, ez dago etiketatuta. Ondorioz, hitz horri ETIK_FALTA etiketa jarri zaio, kasu honetan ere, lerroaren hasieran. Aipaturiko prozedura hau jarraitu da bi pertsonen arteko adostasuna nahiz corpusa ebaluatzerakoan.

```
``<,>''``<PUNT_KOMA>''
        PUNT_KOMA
ETIK_FALTA ``<batere>''
           ``batere'' ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG %SINT
ETIK_ONDO ``<harrotu>''    S:278/0
           ``harrotu'' ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG
              %ADIKAT S:278 !gabeAUR
ETIK_ONDO ``<gabe>''    S:141/0
           ``gabe'' ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> %SINT S:141 !
              gabe
``<$.>''``<PUNT_PUNT>''
        PUNT_PUNT
```

3.9 irudia: Erregelak ebaluatzeko etiketatzaileak hitzei esleitutako etiketak.

3.2.3. Diskurtsoa: erlazio diskurtso-egiturak, beren osagaiak eta unitate zentrala

Diskurtso mailan, RST jarraituz[15] (Mann eta Thompson, 1988), bi alderdi landu ditugu. Alde batetik, balentzia-aldaketak erlaziozko diskurtso-egituretan aztertu ditugu. Beste aldetik, unitate zentraletan, hots, testuko

[15] *Egitura Erretorikoaren Teoria*ri (RST) buruzko azalpenak 5.3.2 azpiatalean daude.

diskurtso-unitate garrantzitsuenetan, orientazio semantikoa duten hitzek zer gramatika-kategoritakoak diren landu dugu.

3.2.3.1. Erlaziozko diskurtso-egiturak

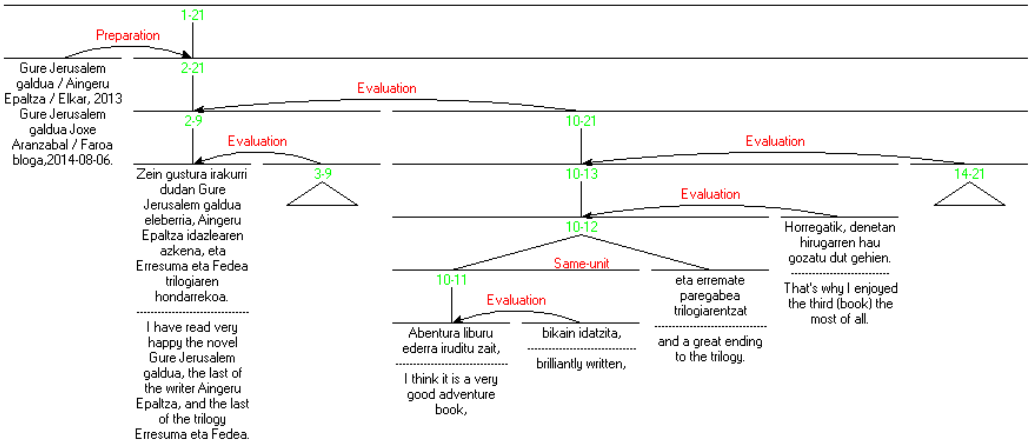
Erlazio erretorikoetan, lehenik eta behin, gure helburuak zeintzuk diren zehaztu ditugu. Guztira hiru helburu finkatu ditugu:

- Erlaziozko diskurtso-egituren eta bere osagaien arteko orientazio semantikoaren arteko adostasuna neurtu nahi dugu.
- Erlaziozko diskurtso-egiturek unitate zentraletik duten distantzia aintzat hartuz, erlazio erretorikoek eta testu osoak zer distantziatan duten orientazio semantikoaren adostasunik handiena neurtu nahi dugu.
- Iritzi-testuetan, esanahi erretoriko ezberdineko erlazioek testuetan beren kokagune zehatza baduten aztertu nahi dugu. Erlazio erretorikoen kokagunea unitate zentralarekiko distantzian oinarrituz finkatuko dugu.

Erlaziozko diskurtso-egituretan ditugun helburuak finkatu ostean, zenbait urrats bete ditugu. Jarraian, metodologia urratsez urrats deskribatzen da:

1- Diskurtso mailako osagai ezberdinei orientazio semantikoa esleitu diegu.
   - 240 iritzi-testuri.
   - Erlaziozko diskurtso-egitura hauei (29 iritzi-testuetatik hartutakoak): EBALUAZIOA/INTERPRETAZIOA (140), KAUSA taldea (71), KONTZETSIOA/ANTITESIA (70), EBIDENTZIA/-JUSTIFIKAZIOA (53), KONTRASTEA (26), BALDINTZA taldea (18), AHALBIDERATZEA/MOTIBAZIOA (6). Guztira, 384 instantziei esleitu zaie orientazio semantikoa.
   - Aurreko erlazio erretorikoen osagaiei: nukleoa eta sateliteari edota erlazioko lehen eta azken osagaiari esleitu zaie orientazio semantikoa.

2- Erlaziozko diskurtso-egituretako parametroen analisia egin dugu. Literaturaren domeinuko 28 testuen erlazio erretorikoak datu-base batean sartu ostean, jarraian zerrendatuta dauden parametroak aztertu ditugu. 3.10 irudiko EBALUAZIOA erlaziozko diskurtso-egitura (EDU 10-12) adibide moduan erabiliko dugu parametroak esplikatzeko:



3.10 irudia: SENTFAR-01 testuaren RST-zuhaitza.

- Nuklearitatea. Erlaziozko diskurtso-egitura nukleo bakarrekoa edo nukleo anitzekoa izan daiteke. 3.10 irudian, EBALUAZIOA erlaziozko diskurtso-egitura N(ukleo)-S(atelitea) motakoa dela zehaztu dugu.
- Erlaziozko diskurtso-egituren eta EDUen orientazio semantikoa. Erlaziozko diskurtso-egiturari nahiz EDUei hiru motako orientazio semantikoa esleitu diegu: positiboa, negatiboa eta neutrala. Lehenik eta behin, EDUei esleitu diegu orientazio semantikoa eta ondoren, erlazio erretoriko osoari. 3.10 irudiko EBALUAZIOA erlazioan (10-11 eta 12), bi EDUei eta erlazio osoari orientazio semantiko positiboa esleitu diegu.
- Unitate zentralarekiko distantzia. Erlaziozko diskurtso-egituren eta unitate zentralaren arteko distantzia bien artean dauden erlaziozko diskurtso-egituren kopurua zenbatuz neurtu dugu. Aztertzen

ari garen EBALUAZIOA erlazioren kasuan (EDU 10-12), distantzia +2 dela adierazi dugu, unitate zentrala 3.10 irudian ezkerretik hasita bigarren EDUa baita eta tartean bi erlaziozko diskurtso-egitura baitaude.

– Erlaziozko diskurtso-egituren esanahi erretorikoa. Erlazioen esanahi erretorikoa izan da aztertutako azken parametroa. Adibidetzat hartu dugun erlazioa EBALUAZIOA motakoa da. Horrek 10-11 EDUek egoera bat aurkeztu eta 12 EDUak egoera horri buruzko ebaluaziozko aipamen bat egiten du.

3- Erlaziozko diskurtso-egituretan eta beren osagaiei orientazio semantiko etiketatzaileen artean dagoen adostasuna neurtu dugu. Eskuzko ebaluazioa egin dugu, F-neurria erabiliz.

3.2.3.2. Unitate zentrala

Bestetik, unitate zentralean zer gramatika-kategoritako hitzak diren ohikoenak jakiteko eta orientazio semantikoa duten hitzek nolako banaketa duten aztertzeko, jarraian zerrendatuta dauden urratsak burutu ditugu:

1- Corpusetik unitate zentralak atera. Lehenik eta behin, pertsona batek Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 192 testu hartu eta testu bakoitzaren unitate zentrala hautatu du. Test zatiko testuak ez dira erabili. Unitate zentrala hautatu aurretik, testuak segmentatu egin dira RST hurbilpeneko Das eta Taboadaren (2018) gidalerroei jarraituz. Bestalde, testuen segmentazioa egin ahal izateko, *RSTTools* (O'Donnell, 2000) programa erabili dugu.

3.11 irudian, EGU01 testua segmentatuta ikus daiteke XML formatuan. Segmentu horietako bakoitza oinarrizko diskurtso-unitate bat da eta anotatzailearen zeregina, hurrengo urratsarekin lotura eginez, diskurtso-unitate horietako zein den unitate zentrala, hots, iritzi-testuko diskurtso-unitaterik garrantzitsuena, identifikatzea da.

2- Unitate zentralaren hautaketako adostasuna neurtu. Pertsona horrek unitate zentrala ondo hautatu duen frogatzeko, beste pertsona batek

```
<rst>
  <header>
    <relations>
    </relations>
  </header>
  <body>
    <segment id=''1''>Gaurko eguraldia. HEGO EKIALDEKO HAIZEA ETA HODEI
        BATZUREKIN EPELDU EGINGO DA.</segment>
    <segment id=''2''> Giro ona tokatuko zaigu gaur ere hego haize
        epelarekin.</segment>
    <segment id=''3''> Borraska pixka bat hurbiltzeak haizea apur bat
        mugituko du</segment>
    <segment id=''4''> baina </segment>
    <segment id=''5''>tenperatura igoaz, </segment>
    <segment id=''6''>giroa goxo mantenduko da leku gehienetan.</segment>
    <segment id=''7''> Hodei zirrinta mehe batzuk agertuko dira zeru sabaian
        </segment>
    <segment id=''8''> baina eguzkia apur bat lausotu arren,</segment>
    <segment id=''9''> astro handia bistan edukiko dugu</segment>
    <segment id=''10''> eta dotoreziak ez du bat ere galduko.</segment>
    <segment id=''11''> Baditeke iluntze aldera ekaitz hodei batzuk garatzea
         eta euri zarrastaren batzuk botatzea agian. </segment>
    <segment id=''12''>Akats txiki horiek gora behera eguraldi bikaina oro
        har.</segment>
  </body>
</rst>
```

3.11 irudia: Bi pertsonek egin beharreko unitate zentralaren aukeraketa EGU01 iritzi-testuan.

ere testu horien unitate zentralak hautatu ditu. Bi pertsonen arteko adostasuna neurtzeko ez dira unitate zentral guztiak erabili; izan ere, guztira 78 iritzi-testuren (corpus osoaren % 32,5) unitate zentrala hautatu behar izan dute bi pertsonek. Adostasuna kalkulatzeko F-neurria erabili dugu.

Adostasunaren neurketak 3.19 taulan erakusten dira, 78 iritzi-testuetatik 51etan, bi pertsonek diskurtso-unitate bera jo dute unitate zentraltzat. Beste 44 iritzi-testuetan, aldiz, diskurtso-unitate ezberdinak lotu dituzte unitate zentralarekin. Beraz, 78 iritzi-testuren unitate zentrala aukeratzean egon den adostasuna 0,65ekoa izan da.

3- Unitate zentraletako hitzen gramatika-kategoriaren azterketa. Hurrengo urratsean, unitate zentral bakoitzean zer hitz-mota agertzen den aztertu nahi izan dugu, baita hitzen bat edo batzuk maiztasun erre-

Neurketa	Instantziak	F-neurria
Adostasuna	51	0,65
Desadostasuna	27	0,35

3.19 taula: Bi pertsonen arteko adostasun-maila 78 iritzi-testuetan unitate zentrala aukeratzerakoan.

gular batez azaltzen diren ere. Horretarako, *Analhitza* tresna (Otegi *et al.*, 2017) erabili dugu. Aipaturiko *Analhitza* tresna horretara, hautatuak izan diren unitate zentral guztiak igo ditugu. Unitate zentral horiek domeinuka aztertu ditugu tresnaren bidez. Unitate zentralak aztertzeko, honako alderdiak hartu ditugu kontuan:

i) Hitzen maiztasuna. Gramatika-kategoria bakoitzeko, gehienez maiztasun handieneko 30 hitz aztertu ditugu.

ii) Hitzen gramatika-kategoria. Aintzat hartutako hitzen besten ezaugarri bat beren gramatika-kategoria izan da. *Analhitza*k (Otegi *et al.*, 2017) zuzenean hitzak gramatika-kategoria aintzat hartuz sailkatzen ditu; horrenbestez, lana erraztu digu.

3.20 taulan, *Analhitza* tresnak Otegi *et al.* (2017) eguraldiaren domeinuko hitzak zein gramatika-kategoriatan sailkatu dituen ikus daiteke, Izen-kategoriako hitzak dira gehienak (% 32,65).

Gramatika-kategoria	Kopurua	%
Izenak	80	32,65
Adjektiboak	25	10,20
Aditzak	69	28,16
Adberbioak	19	7,76
Determinatzaileak	10	4,08
Juntagailuak	35	14,29
Preposizioak	0	0,00

3.20 taula: *Analhitza* tresnak (Otegi *et al.*, 2017) eguraldiaren domeinuko hitzei egindako gramatika-kategoriaren sailkapena.

Gainera, *Analhitza* tresnak (Otegi *et al.*, 2017) gramatika-kategoria

bakoitzeko hitzak zerrendatuta ematen ditu, 3.21 taulan ikusi daitekeen moduan.

Adjektiboak
epel
eder
dotore
goxo
eguzkitsu
atsegin
txar
oker

3.21 taula: Eguraldiaren domeinuko unitate zentraleko hitzen zerrenda.

iii) Hitzen domeinua. Maiztasunik handieneko hitz horiek domeinuari oso lotutakoak diren edo ez ere aztertu dugu. Domeinu bakoitzean, hitzak bi multzotan sailkatu ditugu: hitzak semantikoki domeinukoak diren edo ez. Eguraldikoaren kasuan, hitzak hiru multzotan sailkatu ditugu: i) eguraldi domeinuari loturiko hitza, ii) eguraldiaren domeinuarekin zerikusirik ez duen, iii) denbora-adierazpen bat den. Eguraldia domeinuan, denborak garrantzia duela uste dugulako egin ditugu hitzen sailkapenean hiru multzo.

Esaterako, eguraldiaren domeinuko maiztasunik handieneko hitzetan, *negu* hitza domeinuko moduan sailkatu dugu eta *itxura*, berriz, ez; ez dagoelako zuzenean domeinu horri lotuta. Azkenik, *asteburu* hitzak denbora adierazten duenez, denborazkoetan sailkatu dugu.

4- Unitate zentralei eta horien hitzei sentimenduen balentzia esleipena. Aztertu dugun beste ezaugarrietako bat unitate zentraletako gramatika-kategorien sentimendu balentzia izan da. Sentimendu-balentziaren esleipena egiteko sortu dugun euskarazko sentimenduen sailkatzailea erabili dugu.

(40) adibidean, euskarazko sentimenduen sailkatzaileak eguraldiaren domeinuko unitate zentral bati sentimendu-balentzia nola esleitu dion ikus daiteke. *Ezegonkor* adjektiboari esleitu dio sentimendu-balentzia eta gramatika-kategoria bakoitzean zenbat hitzei esleitu zaien sentimendu-balentzia zenbatu dugu.

(40) giro $[ezegonkor]_{-1,5}$ nagusi izan izan etorri egun. (EGU01)

3.22 taulan, eguraldiaren domeinuko gramatika-kategorietan, zenbat hitzek duten sentimendu-balentzia zenbatzen da. Adjektibo-kategoriako hitzen dute sentimendu-balentzia gehien.

EGURALDIA	**Sentimendu-balentziadun hitzen kopurua**
Izena	7
Aditza	12
Adjektiboa	22
Adberbioa	3
Guztira	44

3.22 taula: Eguraldiaren domeinuan, gramatika-kategoria bakoitzeko sentimendu-balentziadun hitzen kopurua.

3.3 Laburpena

Kapitulu honetan, tesi-lan honen metodologia deskribatu dugu. Lehenik eta behin, balentzia-aldatzaileak identifikatzeko eta beren eragina neurtzeko baliabide eta tresnak sortu ditugu: Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) (diskurtso-egituraren eta orientazio semantikoaren aldetik etiketatuta) eta *SentiTegi* (Alkorta *et al.*, 2018) izeneko euskarazko sentimendu lexikoia.

Bigarren urratsean, balentzia-aldatzaileak identifikatu eta euren eragina neurtu dugu. Balentzia-aldatzaile fonologikoak eta morfologikoen instantziak corpus zati batetik eta sentimendu-balentzia esleitu diegu. Gero, instantzia horiek hitzean eragiten duten neurtu dugu.

Ezeztapen-markak ere antzeko modu batean landu ditugu. Ezeztapen-markak corpus zati batean bilatu eta ezeztapen-marka duten esaldiei sentimendu-balentzia esleitu diegu. Beren eragina neurtu ondoren, ezeztapen-markak eta beren irismena identifikatzeko erregelak garatu eta, azkenik, *Murriztapen Gramatika* ingurunera egokitu ditugu horiek ebaluatzeko asmoz.

Diskurtso-mailan, RST hurbilpenaz etiketatutako corpus zatitik esanahi erretoriko ezberdinetako erlazio lortu ditugu. Ondoren, eskuz nahiz *SentiTegi* (Alkorta *et al.*, 2018) erabiliz orientazio semantikoa eta sentimendu balentzia esleitu diegu eta erlazio erretorikoak sentimenduen analisiaren ikuspegitik landu ditugu. Bestalde, unitate zentrala ere landu dugu.

4. KAPITULUA

Sentimenduen analisirako baliabideak

Kapitulu honetan, tesi-lan honen ondorioz garaturiko baliabideak aurkeztuko ditugu. Guztira hiru baliabide garatu ditugu: Euskarazko Iritzi Corpusa (4.1 atala), *Sentitegi* izeneko euskarazko sentimenduen lexikoia (4.2 atala) eta, azkenik, dokumentu mailako euskarazko sentimenduen sailkatzailea (4.3 atala).

4.1 Euskarazko Iritzi Corpusa

Atal honetan, euskarazko egunkari eta webgune espezializatuetik bildutako 240 iritzi-testuez osaturiko corpusa azalduko dugu. Lehenik eta behin, 3.1.1 atalean jarraituriko metodologiaz sortu dugun iritzi-testuen corpusaren ezaugarriak aipatuko ditugu. Ondoren, berriz, metodologiaren garapenean izandako zailtasunen berri emango dugu.

4.1.1. Euskarazko Iritzi Corpusaren ezaugarriak

Sortu dugun Euskarazko Iritzi Corpusak guztira 240 iritzi-testu ditu sei domeinutakoak: kirola, politika, musika, zinema, liburuak eta eguraldia.

Corpusaren beste ezaugarrietako bat bertako iritzi-testuetan agertzen den balorazioaren oreka da. Domeinuko 40 testu daude eta erdiek balorazio positiboa dute eta beste hainbestek, berriz, negatiboa.

Analhitza tresnak (Otegi *et al.*, 2017) ematen dituen emaitzen arabera, corpusak 52.092 token eta 3.711 esaldi ditu. Corpusa beste hizkuntzetako iritzi-

testuen corpusekin alderatuz, tamainaz zertxobait txikiagoa dela esan daiteke, 4.1 taulan agertzen den moduan. Beste hizkuntzetan, oro har, iritzi-testu kopurua handiagoa da, baina, tokenen kasuan, euskarazko corpusaren eta beste corpusen arteko ezberdintasuna are handiagoa da.

Corpusa	**Testuak**	**Hizkuntza**	**Tokenak**
Boldrini *et al.* (2010)	300		
Hu eta Liu (2004)	113		81.855
Rushdi-Saleh *et al.* (2011)	500	Arabiera	215.948
Wilson (2008)	535	Ingelesa	265.000
Taboada (2008)	400	Ingelesa	289.270
Quan eta Ren (2009)	1.487	Txinera	878.164

4.1 taula: Sentimenduen analisirako eskura dauden iritzi-testuen corpusak.

4.1.2. Euskarazko Iritzi Corpusaren garapena

Euskarazko Iritzi Corpusa sortzerakoan, bi zailtasun izan ditugu. Alde batetik, guk nahi genituen moduko iritzi-testuak bilatzea zaila suertatu da (4.1.2.1 atala). Beste aldetik, domeinu batzuetan, orientazio semantiko jakin batzuetako iritzi-testuak biltzea ere asko kostatu da (4.1.2.2 atala). Azkenik, 4.1.3.2 atalean, corpusaren baliagarritasuna neurtzeko hautatutako irizpideak zergatik hautatu ditugun azalduko dugu.

4.1.2.1. Ezaugarri jakin batzuetako iritzi-testuak bilatzeko zailtasunak

Gure lana eta lanaren emaitza konparatu ahal izateko, badagoen baliabide (kasu honetan, corpusa) baten antzeko bat egin nahi izan dugu. Hau da, hitz edo sintagmen balentzia aldatzen duten fenomeno linguistikoak aztertzeko moduko iritzi-testuak bildu nahi ditugu. Horregatik, Euskarazko Iritzi Corpusa sortzerakoan, *SFU Review Corpus*[1] (Taboada, 2008) hartu dugu erreferentziatzat.

[1]Corpusa eskuragarri dago `https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html` webgunean.

*SFU Review Corpus*ak (Taboada, 2008) zortzi gaietako 400 iritzi-testu biltzen ditu. Gai bakoitzeko 50 testu daude, erdiak iritzi positibodunak dira eta beste erdiak, berriz, negatibodunak. Iritzi-testuek lantzen dituzten gaiak hauek dira: liburuak, autoak, ordenagailuak, sukaldeko tresneria, pelikulak, musikak eta telefonoak.

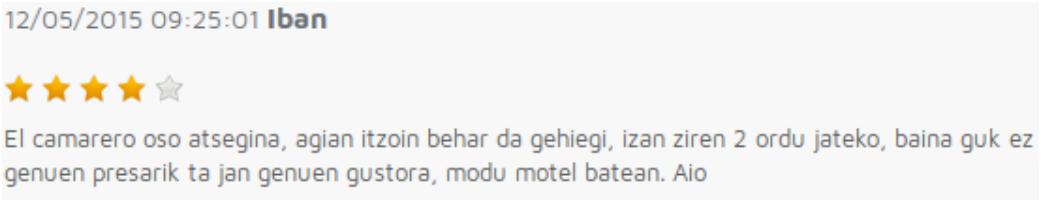
Gure asmoa gai horiei buruzko iritzi-testuak biltzea izan da, baina euskara hizkuntza gutxitua denez, eta horrek zailtasunak areagotu egiten dituenez, asmoa birplanteatu egin behar izan dugu. Izan ere, liburu, musika eta pelikulei buruzko iritzi-testuak aurkitu ditugu, baina besteen kasuan, kalitate baxukoak edo desegokiak diren iritzi-testuak aurkitu ditugu edo ez ditugu aurkitu.

4.1 irudian, *Tripadvisor* webgunetik hartutako iritzi-testu bat ikus daiteke. Hiru hizkuntzetan idatzita dago iritzi-testu bakarra eta hiruretan esaten dena ez da guztiz berdina. Hizkuntza bat baino gehiago agertzen diren iritzi-testuak aurkitzea oso ohikoa izan da gure iritzi-testuen bilaketan. Badirudi, iritzi-emaile euskaldunek beren iritziek ahalik eta irismen handiena izatea nahi dutela eta, horregatik, euskaraz gain beste hizkuntza bat erabiltzen dutela.

4.1 irudia: Iritzi-testu hirueledun baten adibidea.

Halaber, aurkitu ditugun euskarazko iritzi-testu gehienak labur samarrak dira eta horrek gure ikerketa-lana mugatu dezake, aztertu nahi ditugun hizkun-

tza-gertakariak maiztasun txikiagoz agertu direlako. Euskarazko iritzi-testu labur bat ageri da 4.2 irudian.



4.2 irudia: Euskarazko iritzi-testu labur baten adibidea.

Ikusten denez, ortografia zuzena ez izateaz gain, egitura sintaktikoa ez da aberatsa eta bertatik diskurtsoaren egitura lantzea ere zailagoa bilakatzen da, testuaren tamaina ez delako handia. Egoera horren ondorioz, zenbait aldizkari eta komunikabide aztertu ditugu eta aurkitzen errazagoak diren beste gai batzuk dituzten iritzi-testuak biltzea pentsatu dugu. Hala, liburuei, musikari eta pelikulei buruzko iritziak biltzeaz gain, eguraldiari, politikari eta kirolari buruzko iritzi-testuak biltzea erabaki dugu. Gainera, ahalik eta kalitate hobereneko iritzi-testuak biltzeko, komunikabide eta aldizkari espezializatuetara jotzea erabaki dugu.

Nahiz eta iritzi-testuen gaiak batzuetan *SFU Review Corpus*ekoetatik (Taboada, 2008) urrundu diren, beste ezaugarrietan antzekoak izan daitezen saiatu gara. Horregatik, erabaki dugu Euskarazko Iritzi Corpusak domeinu bakoitzeko 40 testu izatea eta domeinu bakoitzeko, iritzi-testuen erdiak positiboak izatea eta besteak negatiboak izatea. Beraz, gai bakoitzeko 20 testu positibo eta beste hainbeste negatibo bildu ditugu.

4.1.2.2. Orientazio semantiko jakin bateko iritzi-testuak bilatzeko zailtasunak

Iritzi-testuak balorazioen aldetik orekatuak izateak beste zailtasun eta arazo batzuk ekarri dizkigu. Esaterako, literatura-liburuen kasuan, zailtasunak izan ditugu balorazio txarreko iritziak aurkitzeko. Zailtasun horren jatorriaren berri Egañak (2013) ematen du.

Egañak (2013) 1975tik 2005ra bitarteko 2.300 euskal literatura kritika aztertu ditu eta haren ikerketaren arabera, literatura kritiken % 84k iruzkin positiboa

edo oso positiboa dute. *Berria* egunkarian egindako elkarrizketan dioenez[2], "euskal literaturak berak ere bere burua legitimatzeko arazoak" ditu eta horren ondorio da iruzkin positiboen kopuru handia. Halaber, "Euskal literaturan oso jende gutxi da kritikari profesionala" eta "subjektuak eta legitimatua den objektuak elkarrengandik urrun samar" ez daudenez, legitimatzeko arazoak sortzen dira. Hots, euskal literatura komunitate txikia denez, eta denen arteko loturak beste literatura-komunitateekin alderatuta estuagoak direnez, iruzkin negatiboak gutxiago egiten dira. Egoera hori aintzat hartuta, balorazio negatiboak bereziki atzerriko lanei buruzko iritzi-testuetan aurkitu ditugu. Politikako iritzi-testuetan, berriz, aurkakoa gertatu da. 2008tik aurrera dagoen krisi ekonomikoa tarteko, iritzi gehienak iruzkin negatibokoa izan da eta iruzkin positibokoak aurkitzea zaila gertatu zaigu.

4.1.3. Euskarazko Iritzi Corpusaren baliagarritasunaren neurketa

Euskarazko Iritzi Corpusa osatu ondoren, corpus horrek subjektibotasunik baduen eta ezeztapen-marken instantzia ugari badauden eta, horrenbestez, sentimendu-analisirako baliagarria den neurtu dugu. Halaber, gure zenbait ikerketa-lerrotarako (hizkuntza maila ezberdinetako balentzia-aldatzaileak identifikatzeko eta euskarazko sentimenduen lexikoi bat sortzeko) beharrezkoak diren hizkuntza-fenomenoak ere agertzen ote diren aztertu dugu. Aztertzea erabaki ditugun aspektuak (lehen pertsonaren erabilera, adjektiboen agerpena eta ezeztapena) iturri bibliografikoetan oinarrituta hautatu ditugu. Jarraian, i) corpusaren baliagarritasuna neurtzeko erabili ditugun baliabideak eta irizpideak eta ii) neurketen emaitzak azalduko ditugu.

4.1.3.1. Baliagarritasuna neurtzeko baliabideak eta irizpideak

Euskarazko Iritzi Corpusaren subjektibotasuna neurteko lehen pertsonaren erabilera eta adjektiboen agerpena neurtu dugu eta, horretarako baliabideak 4.2 taulako corpusak izan dira. Corpus horiek alderatu egin ditugu.

[2]Elkarrizketa esteka honetan dago eskuragarri: `www.berria.eus/paperekoa/1744/024/001/2015-02-04/literatura_batek_sendotzeko_gatazka_behar_du_bere_barruan.htm`, *Berria*, 2015-02-14.

	Subjektiboa	**Objektiboa**
Euskara	Euskarazko Iritzi Corpusa	Euskarazko Wikipediako artikuluak
Ingelesa	SFU Review Corpus	Ingeleseko Wikipediako artikuluak

4.2 taula: Lehen pertsonaren eta adjektiboaren agerpena neurtzeko erabilitako corpusak.

- *SFU Review Corpus* (Taboada, 2008). Gure erreferentziazko corpusa da. Biek hainbat domeinutako iritzi-testuak biltzen dituzte eta, horrenbestez, eduki subjektiboa dute. 16.705 esaldi ditu corpusak.
- Ingelesezko Wikipediako zenbait artikulu. *SFU Review Corpus*aren (Taboada, 2008) domeinuekin lotura duten artikuluak hartu ditugu. Guztira 8 artikulu bildu ditugu eta corpus horrek 64.258 hitz ditu.
- Euskarazko Wikipediako zenbait testu. Ingeleseko Wikipedian bezala, kasu horretan ere, Euskarazko Iritzi Corpuseko domeinuekin bat datozen artikuluak erabili ditugu. Euskarazko Wikipediaren kasuan, 28 artikulu bildu ditugu eta corpus horrek 37.964 hitz ditu.

Wikipedia entziklopedia bat izanik, artikuluek ikuspegi neutroa edukitzea eskatzen da[3]. Testurik objektiboenak horiek direla uste dugulako, ingelesezko eta euskarazko corpus objektiboak Wikipedian oinarrituta sortu ditugu.

Lehen pertsonaren eta adjektiboaren agerpenaren neurketa aipaturiko corpusetan horrela eta arrazoi horiengatik gauzatu dugu:

1- Lehen pertsonaren agerpena. Iritziak, normalean, lehen pertsonan adierazi ohi dira (Halliday *et al.*, 2014) lehen pertsona zenbateko maiztasunez agertzen den aztertu dugu gure corpusean. Gure corpuseko emaitzak konparatzeko 4.2 taulako beste hiru corpusak erabili ditugu.

   Lehen pertsonaren agerpena neurtzerakoan, zailtasun batekin egin dugu topo. Euskaran, aditz laguntzailean agertzen dira aditzari lotutako marka morfologikoak eta, horrenbestez, baita pertsona eta nume-

[3]Esteka honetan azaltzen dira Wikipediaren ezaugarriak: `https://eu.wikipedia.org/wiki/Wikipedia`.

roari buruzko informazioa ere. Ingelesean, aldiz, hirugarren pertsonaren kasuan izan ezik, aditz-forma guztiak berdinak dira eta horrek lehen pertsona neurtzea eragozten du. Konponbide moduan, ingelesean pertsona-izenordainen presentzia neurtzea erabaki dugu. Beraz, lehen pertsonaren agerpena neurtzeko, euskaran aditz laguntzaileak erabili ditugu eta ingelesean, ordea, pertsona-izenordainak. Neurketa egitean, lehen pertsona hori singularra edo plurala den kontuan hartu dugu.

2- Adjektiboen agerpena. (Liu, 2010, 32 orr.) lana gramatika-kategoriez (*Part-Of-Speech*, POS) ari denean, adjektiboak iritzien adierazle garrantzitsuak direla adierazten du. Horregatik, adjektiboek gramatika-kategoria guztien artean duten presentzia neurtu dugu.

Euskarazko Iritzi Corpusean ezeztapen-marka asko ageri diren zenbatu dugu; izan ere, ezeztapena nolako balentzia-aldatzailea den landutako hizkuntza-fenomenoetako bat izan da. Kasu horretan, bere agerpen asko dauden interesatzen zaigu eta ez beste corpusekin alderatzea; izan ere, ezeztapena lehen pertsona eta adjektiboen erabilera ez bezala ez dira subjektibitatearen adierazle.

3- Ezeztapena. Ezeztapena garrantzitsua da esaldien eta testuen sentimendu-balentzia aldatzeari dagokionez, Polanyi eta Zaenenen arabera (2006) testuinguruko balentzia-aldatzailea da. Hori dela eta, corpusean agertzen diren ezeztapen-markak zenbatzea erabaki dugu; instantzia kopuru minimo bat behar baitugu azterketa ondo gauzatzeko.

   Laburbilduz, lehen pertsonaren eta adjektiboen agerpen kopurua beste corpusetako (euskarazko eta ingelesezko Wikipediako corpusak eta *SFU Review Corpus* (Taboada, 2008)) agerpen kopuruarekin alderatu dugu. Agerpen kopuruaren alderaketa eginez Euskarazko Iritzi Corpusa tesi-lanerako baliagarria zaigun zehaztu dugu. Ezeztapenaren kasuan, aldiz, beren agerpena Euskarazko Iritzi Corpusean zenbatu egin dugu, aztertu dugun balentzia-aldatzaileetako bat baita.

4.1.3.2. Neurketen emaitzak

Corpusaren baliagarritasuna neurtzeko aztertu diren ezaugarriak hauek izan dira: lehen pertsonaren agerpena (neurketa eta alderaketa), adjektiboaren agerpena (neurketa eta alderaketa) eta ezeztapena (kontaketa).

- Lehen pertsonaren agerpena. Corpus horietan agertzen den lehen pertsona kopurua 4.2 taulako baliabideetakoekin alderatu dugu. Euskaran, lehen pertsonaren erabilera (singularra eta plurala) aditz laguntzailean neurtu dugu, eta ingelesean, pertsona-izenordainetan.

Corpusa	Lehen sg.	Lehen pl.	Guztira
Euskarazko Wikipedia	% 0,12	% 1,09	**% 1,21**
Ingelesezko Wikipedia	% 0,03	% 0,09	**% 0,12**
Euskarazko Iritzi Corpusa	% 3,27	% 5,10	**% 8,37**
SFU Review Corpus	% 10,10	% 1,70	**% 11,80**

4.3 taula: Lehen pertsonaren agerpena lau corpus ezberdinetan.

4.3 taulan agertzen den moduan, euskarazko nahiz ingelesezko Wikipedietatik hartutako testuekin osatutako corpusetan, lehen pertsonaren agerpena oso baxua da. Ingelesezko Wikipediaren kasuan, azaltzen diren pertsona-izenordain guztien % 0,12 lehen pertsonari dagokio. Joera bera erakusten du euskarazko Wikipediak, non bertako aditzetako marka morfologikoetan, % 1,21a lehen pertsonakoak diren.

Corpus subjektiboetan, aldiz, lehen pertsonaren agerpena handiagoa da. Euskarazko Iritzi Corpusean, aditzetako marka morfologikoetan, lehen pertsonaren agerpena % 8,37koa da. *SFU Review Corpus*ean (Taboada, 2008), lehen pertsonaren agerpena handiagoa da; pertsona-izenordainen % 11,80 baitago lehen pertsonan. Bestalde, aipagarria da corpus subjektiboetan, lehen pertsona plurala dela nagusi euskararen kasuan eta lehen pertsona singularra, berriz, ingelesaren kasuan.

- Adjektibo-kopurua. Corpusaren baliagarritasuna neurtzeko aintzat hartu dugun beste alderdi bat adjektiboen kopurua izan da. Adjektiboak oinarrizko elementuak dira maila lexikoan subjektibotasuna erakusterakoan, beraz, corpus subjektiboetan adjektiboen agerpen handiagoa

espero liteke corpus objektiboetan baino. Hala ere, 4.4 taulako emaitzek ez dute hori erakusten.

Corpusa	Guztira
Euskarazko Wikipedia	**% 8,50**
Ingelesezko Wikipedia	**% 9,09**
Euskarazko Iritzi Corpusa	**% 9,82**
SFU Review Corpus	**% 8,35**

4.4 taula: Adjektiboen agerpena corpus objekibo eta subjektiboetan.

Emaitzen arabera, adjektiboen agerpena antzekoa da corpus guztietan, bai objektiboa edo subjektiboa izanda, baita euskarazkoa edo ingelesezkoa izanda ere. *Analhitza*k (Otegi *et al.*, 2017) emandako datuen arabera, Euskarazko Wikipedian eta ingelesezko Wikipedian, adjektiboak gramatika-kategoria guztietatik % 8,50 eta % 9,09 dira, hurrenez hurren. Euskarazko Iritzi Corpusean adjektiboak gramatika-kategoria guztien % 9,82 dira eta *SFU Review Corpus*ean (Taboada, 2008) % 8,35. Ondorioz, badirudi adjektiboen agerpen handiagoa ez dela corpus subjektiboen ezaugarri bat. Baliteke horren ordez, adjektibo-mota izatea testu objektibo eta subjektiboen arteko ezberdintasuna.

- Ezeztapena. Corpusean neurtu dugun azken ezaugarria ezeztapena eta zehazki, ezeztapen-markak izan dira. Hiru ezeztapen-marka moten kontaketa egin dugu.

  Alde batetik, perpaus osoari (adibidez, *gaur dendak ez daude zabalik*) eragiten dion *ez* ezeztapen-marka 718 aldiz agertzen da. Izen sintagmari (adibidez, *arrakastarik gabe*) eragiten dino *gabe* ezeztapen-marka 107 aldiz agertzen da eta azkenik, menpeko perpausei eragiten dien *ezean* lau aldiz zenbatu dugu. Kasu horietan, ezeztapen-marka kopuru handia aurkitu dugu corpusean eta horrek horiek ikertzeko aukera eman digu.

4.2 Euskarazko sentimenduen lexikoia

Atal honetan, hasteko, euskarazko sentimenduen lexikoiaren bi bertsioek dituzten ezaugarriak aurkeztuko ditugu. Ondoren, euskarazko lexikoiaren garapeneko zenbait alderdi (ordainen kasuistika eta hartutako erabakiak, besteak beste) aipatuko ditugu eta, azkenik, lexikoiaren ebaluazioa azalduko dugu.

4.2.1. Euskarazko sentimenduen lexikoiaren ezaugarriak

Gaztelaniazko SO-CAL lexikoia (Brooke *et al.*, 2009) itzultzeko jarraitu dugun prozesuaren ondorioz, euskarazko sentimenduen lexikoi beraren bi bertsio lortu ditugu. 4.5 taulak sentimenduen lexikoiaren bi bertsioen emaitzak erakusten ditu.

	V1.0		V2.0	
Gramatika-kategoria	**Sarrerak**	**%**	**Sarrerak**	**%**
Izenak	2.282	28,06	461	37,27
Adjektiboak	3.162	38,85	446	36,05
Aditzondoak	652	7,98	54	4,36
Aditzak	1.657	20,36	276	22,32
Intentsifikatzaileak	387	4,75		
Guztira	**8.140**	100	**1.237**	100

4.5 taula: Euskarazko sentimenduen lexikoiaren bi bertsioen ezaugarriak.

Sentimenduen lexikoiaren lehen bertsioa (V1.0) bigarren bertsioa (V2.0) baino handiagoa da. Horren atzean zenbait arrazoi daude. Hasteko, lehen bertsioan *Zehazki* (Sarasola, 2005) eta *Elhuyar* (Elhuyar, 2013) hiztegien bidez lorturiko euskarazko ordain guztiak hartu dira kontuan. Bigarren bertsioan, aldiz, bi hiztegi horietako sarrerak diren euskarazko ordainak bakarrik hartu dira kontuan. Horrenbestez, itzulpen bidez lortu diren lexikoiaren sarrerek aniztasun handia erakusten dute: batzuek izenlagun hizkia daramate (adibidez, *garrantzizko*), beste batzuk kolokazioak[4] dira, etab. Halaber, intentsifi-

[4]Kolokazioak uste baino maiztasun handiagoaz batera agertzen diren hitzak dira. *Adarra jo* horren adibidea da.

katzaileak ere lexikoiaren lehen bertsioan badaude, bigarrengoan ez bezala.

Lexikoi horien arteko berdintasunak eta ezberdintasunak aintzat hartuta, lexikoiaren lehen bertsioak 8.140 sarrera ditu, bigarren bertsioak 1.237 dituen bitartean. Batetik bestera asko jaisten da sarrera kopurua; izan ere, bigarrengo bertsioa domeinuari lotuta dago. Bi bertsioetan izenak eta adjektiboak dira gramatika-kategoria nagusienak. Aditzak zerbait gutxiago dira eta aditzondoak, berriz, are gutxiago. Bigarren bertsioan, intentsifikatzailerik ez dago arrazoi praktikoengatik. Gure ustez, intentsifikatzaileek beste hitzei eragiten dietenez, gertuago daude sintaxi mailatik lexiko mailatik baino. Gainera, kontuan hartu behar da intentsifikatzaileek % −100 eta % +50 arteko balioa dutela biderketa eragiketan, beste gramatika-kategoriek −5 eta +5 arteko balioak (gehiketa eta kenketa eragiketak) dituzten bitartean.

Lexikoiaren bi bertsioen beste ezaugarri bat lexikoi paraleloak direla da. Hots, lexikoietako sarrerekin datu-base bat sortuta dago eta euskarazko sarrerarekin gaztelaniazko eta ingelesezko ordainak (bigarren bertsioan) paraleloki eta modu ordenatuan daude jarrita. Gaztelaniazko eta ingelesezko ordainek ere beren sentimendu-balentzia badute, 4.6 taulan ikus daitekeenez. Bertan, lexikoia paraleloa dela erakusten duten lau adibide daude. Adibideak adjektiboak dira eta beren sentimendu-balentziak eskalan daude.

Hitza lexikoian	**Balentzia**	**SPA**	**Balentzia**	**ENG**	**Balentzia**
bikain	+5	excepcional	+5	excellent	+5
on	+2	buen	+2	-	-
eskas	−1	escaso	−2	insufficient	−1
txar	−3	adverso	−3	bad	−3

4.6 taula: Lexikoi paraleloaren adibide batzuk.

Ikusten den moduan, batzuetan zerrendetako sentimendu-balentziak ez datoz bat (*eskas* (−1), *escaso* (−2) eta *insufficient* (−1), esaterako), gaztelaniazko eta ingelesezko lexikoiak ez direlako modu bere-berean sortu. Baina, euskarazko lexikoiko sarreretako sentimendu-balentziak beti ingeleseko edota gaztelaniako sarreretako sentimendu-balentzia batekin bat egin behar du.

Behin euskarazko lexikoiaren euskarazko lehen bertsioa euskarazko sentimenduen sailkatzailean integratu ondoren, sorturiko euskarazko sentimen-

duen lexikoia gauza da hitzei nahiz esaldiei sentimendu-balentzia automatikoki esleitzeko. Sentimenduen lexikoiak hitzei esleitzen die sentimendu-balentzia, eta ondoren, sentimenduen sailkatzaileak eragiketa egiten du esaldiaren sentimendu-balentzia kalkulatzeko. Hurrengo adibideek hori erakusten dute.

(41) Halere, pentsa litekeenaren aurka, gaien urritasunak eta diskurtso [errepikakorrak]$_{-6}$ ez dakarte ñabardura aberastasunik, are gutxiago argumentu-mailako sakontasunik.[–6] (LIB18)

(42) Arazo [nagusia]$_{+2}$, nire ustez, gaien [emankortasun]$_{+4}$ zalantzazkoan eta ekintzaren bilakaera [eskasean]$_{-3}$ datza.[+3] (LIB18)

(43) (...) Emaitza [ezustekorik]$_{-1.5}$ gabeko istorio bat da, irakurlea [epel]$_{-1.5}$ uzteko arrisku dezente duen tonu arras moderatu batean emana.[–3] (LIB18)

Goiko adibideetan, hitzek eskuinetara sentimendu-balentzia bat dute, euskarazko sentimenduen sailkatzailean inplementatu dugun sentimenduen lexikoiarenak. Esaldien amaieran, esaldi osoaren sentimendu-balentzia ageri da eta esaldian dauden sentimendu-balentzi guztien batura da. Lehen adibidean, esaterako, lexikoiaren arabera hitz batek, *errepikakorrak* hitzak, bakarrik du sentimendu-balentzia –6 eta ondorioz, esaldiarena ere halaxe da[5]. Hurrengo bi adibideetan ere, lexikoiak funtzionamendu bera du euskarazko sentimenduen sailkatzailearen baitan.

4.2.2. Lexikoiaren sorkuntza

Euskarazko sentimenduen lexikoia sortzeko metodologian, hartu ditugun erabakiak azalduko ditugu (4.2.2.1 atala); ondoren, gaztelaniazko lexikoia euskaratzerakoan agertu zaigun kasuistikaren berri emango dugu (4.2.2.2 atala);

[5]Gogoan hartu behar da lexikoiko sarreren sentimendu-balentzia –5 eta +5 artekoa dela, baina SO-CAL tresnan badaude zenbait eragiketa hizkuntzarekin zerikusia dutenak eta hizkuntza guztietan aplikatzen direnak eta horiek hitzaren sentimendu-balentzia indar edo ahul dezakete. Kasu horretan, *errepikakorrak* hitzaren sentimendu-balentzia –6raino indartu da.

jarraian, gaztelaniazko lexikoi horren euskarazko ordainak ematen sentimenduen balentziak aukeratzeko zein irizpide hartu ditugun kontuan azalduko dugu (4.2.2.3 atala).

4.2.2.1. Metodologian hartutako erabakiak

Sentimenduen lexikoiko hitzek eta beren sentimendu-balentziek elkarren artean koherentzia izan dezaten, metodologian zehar erabaki hauek hartu ditugu:

- Nahiz eta gaztelaniazko lexikoia euskaratu dugun, kasu gehienetan, euskarazko ordainei ingelesezko lexikoiaren sentimendu-balentziak esleitu dizkiegu. Lan honetan gaztelania edota ingelesa izan zitezkeen hastapeneko hizkuntzak. Guk gaztelania hobetsi dugu hastapeneko hizkuntza moduan arrazoi hauengatik: alde batetik, sentimenduen lexikoiaren gaztelaniako bertsioaren batez besteko doitasuna % 71,81ekoa da, ingelesekoarena % 76,65 den bitartean. Beste aldetik, gaztelaniako hitzei euskarazko ordaina emateko baliabide gehiago daude ingelesetik baino eta, gaztelaniatik ordaina ematea errazagoa izango litzateke.

  Halaber, beste arrazoi bat badago eta hori hizkuntzari lotuta dauden ezaugarri soziokulturalak dira. Euskarak eta gaztelaniak ezaugarri soziokultural gehiago partekatzen dituzte, ingelesak eta euskarak baino. Horren adibide dugu 4.2.2.2 azpiataleko 4. Fenomenoa, *frankismo* hitzaren kasua, hain zuzen ere. Euskarak eta gaztelaniak partekatzen duten ale hitz bat da. Horrelako kasu gehiago egon dira euskarazko ordainak ematean. Hizkuntzen artean gertatzen diren itzulpeneko kasu hauek Meng *et al.*en (2012) lanean aipatzen dira eta Mohammad *et al.*ek (2016) itzulpenetan gertatzen diren sentimenduen aldaketak ere aipatzen ditu.

  Bestalde, ingelesezko lexikoiak gaztelaniazko lexikoia baliatuta euskaraz sortu dugun lexikoian esleitu diren balentziak egokiak diren aztertzeko eta ez badira zuzentzeko balio izan digu. 3.10 taulako 3. Fenomenoan (*seinale*) gertatu den moduan, euskarazko ordain batek aukeran gaztelaniazko lexikoiko ale hitzaren (*cicatriz*) eta ingelesezko

lexikoiko ale hitzaren (*signal*) sentimedu-balentziak dituenean, ingeleseko sentimendu-balentziaren alde egin dugu, batez besteko doitasuna handiagoa duelako (% 76,65). Euskarazko sarrerak ingeleseko lexikoian ordainik ez duenean, aldiz, kasu batzuetan euskarazko sarrera hori lexikoitik kendu egin dugu, gaztelaniako lexikoiaren doitasun baxuagoak (% 71,81) euskarazko lexikoiaren kalitatea jaitsi dezakeelakoan. Ondorioz, 576 gaztelanizko hitzei, hasiera batean, euskarazko ordaina eta sentimendu-balentzia esleitu diegun arren, horiek euskarazko sentimenduen lexikoitik kendu egin ditugu, ingelesezko lexikoian agertzen ez direlako.

Beraz, batetik, lexikoiaren gaztelaniazko bertsioaren estaldura eta ezaugarri soziokulturalak eta, bestetik, lexikoaren ingelesezko bertsioaren doitasuna (hitzen balentzia) uztartu ditugu.

- Gaztelaniazko lexikoiko hitz baten euskarazko hainbat ordain kontuan hartzea. Gaztelaniazko lexikoiko hitzei euskarazko ordaina ematean izan dugun beste zalantzetako bat euskarazko ordain guztiak kontuan hartu behar diren izan da. Izan ere, gaztelaniazko hitzak batzuetan euskarazko hainbat ordain izan ditzke esanahi bera mantenduz. Esaterako, gaztelaniazko hitz batek euskarazko hainbat izan ditzake euskaran, 4.7 taulan ikusten den moduan. Halaber, hitzak polisemikoak ere izan daitezke, eta horrek ere gaztelaniazko hitz bati euskarazko hainbat ordain ematea eragiten du; nahiz eta esanahia kasu horretan ezberdina den. Egoera horren aurrean, hau da, gaztelaniazko hitz bati euskarazko hainbat ordain ahal zaizkionean, euskarazko ordain posible guztiak aintzat hartzea erabaki dugu sortuko den lexikoiak ahalik eta estaldura handiena izan dezan.

  4.7 taulan bi adibide daude non ordain posible guztiak kontuan hartu direla erakusten den. Gaztelaniazko lexikoian *enfermedad* eta *horror* hitzak daude eta horiek *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak erabiliz euskarazko ordainak lortu dira eta hitz horietako bakoitzak euskaran hainbat ordain posible ditu, hirugarren zutabean agertzen den moduan.

- Hitz polisemikoen domeinuaren egokitzapena. Gaztelaniako lexikoiko

Gaztelaniaz jatorrian	Balentzia	Ordain posible guztiak euskaraz
enfermedad (izena)	−1	gaixotasun, gaitz, eritasun, afekzio, ezontsa
horror (izena)	−4	laztura, izu, lazgarrikeria, izugarrikeria

4.7 taula: Euskarazko ordain posible bat baino gehiago dituzten adibideak.

hitzei euskarazko ordaina ematean eta sentimendu-balentzia esleitzean, testuinguruaren arabera esanahi ezberdinak eta zeinu ezberdinetako (+ edo −) sentimendu-balentziak dituzten ordainak agertu dira, 4.8 taulako modukoak. Egokiena euskarazko ordain guztiak (eta sentimendu-balentziak) kontuan hartzea izango litzateke, nahiz eta euskarazko ordainek esanahi ezberdinak izan, baina horrek lexikoiaren konplexutasuna asko handituko luke. Horretaz gain, lexikoia inplementatuta egongo litzakeen sistemak hitzen adiera-desanbiguaziorako gaitasuna eduki beharko luke.

Hori konpontzeko, esanahi ezberdinak eta zeinu ezberdinetako sentimendu-balentziak dituzten ordainetan esanahi bat bakarrik hartzea erabaki dugu. Hautaketa egitean, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen diren testuak eta domeinuak erabili ditugu.

4.8 taulan, hitz polisemikoetan balentzia hautaketa nola egin den esplikatzen da. *Deigarri* eta *lehiatsu* adjektiboek esanahi ezbedinak (*deigarri*ren kasuan, *aparatoso* −3, *llamativo* +3 eta *lehiatsu*ren kasuan, *ansioso* −3, *apasionado* +4 eta *competitivo* +3) dituzten hitzen ordainak dira eta, metodologiako laugarren urratsean, euskarazko ordain horiei sentimendu-balentzia esleitu behar diegu. Aukeran sentimendu-balentzia bat baino gehiago daudenez (*deigarri*ren kasuan, −3 eta +3 eta *lehiatsu*ren kasuan, −3, +4 eta +3), euskarazko ordain horien testuingurua aztertu dugu Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) gako-hitzak testuinguruan (*Key Word In Context*, KWIC) erabiliz eta ordainen corpuseko testuinguruekin hobekien datorren sentimendu-balentzia aukeratu dugu. *Deigarri* eta *lehiatsu* ordainetan, +3 sentimendu-balentzia izan da hautatu duguna.

- Koherentziaren sendotasuna. Euskarazko ordaina lortu ondoren, ordai-

Euskal ordaina	Jatorriz gaztelaniaz eta balentzia	Bal. aukeratua
deigarri (adj.)	aparatoso −3 , llamativo +3	+3
lehiatsu (adj.)	ansioso −3 , apasionado +4 , competitivo +3	+3

4.8 taula: Hitz polisemikoen tratamendua.

nei sentimendu-balentzia esleitu diegu, balentziak bat zetozen kasuetan, balentzietan koherentzia mantentzen saiatu gara.

Irizpidea	Euskara	Balentzia	Euskara	Balentzia
A	errukigabe	−4	errukigabeko	−4
B	tonto	−3	tuntun	−3
C	arduradun	+2	arduragabe	−2

4.9 taula: Itzulpen-prozesuaren koherentzia erakusten duten adibideak.

4.9 taulan ageri dira zenbait adibide. Bertan agertzen diren kasuak euskarazko ordaina ematerakoan laugarren urratsekoak dira, balentzia hautatu behar den mementokoak. Guztira hiru kasu ezberdin antzeman ditugu:

A- Izenlagunarekin agertzen diren euskarazko ordainak. 4.9 taulan, A kasuan, *errukigabe* lema ageri da. *Errukigabe* eta *errukigabeko* izenondoak dira, baina, bigarrenak izenlagunaren *-ko* atzizkia ere badu. Bi hitzei balentzia hautatzeko mementoan, balentzia bera eman diegu; gramatika-kategoria gorabehera, esanahi bera baitute. Aipagarria da horrelako kasu asko egon direla balentzia hautaketan zehar. Izenlaguna daraman euskarazko ordainaren beste kasu bat *berehala/berehalako* hitz parea izan da eta biei +2 balentzia esleitu diegu. Kontuan hartu behar da, kasu hori euskarazko lexikoiaren lehen bertsioa sortzerakoan gertatu dela, lexikoiaren bigarren bertsioan hiztegietako sarrerak diren hitzak bakarrik hartu baititugu kontuan.

B- Esanahi bera, baina genero ezberdina. Kasu gutxi batzuk egon dira non bi hitz ezberdin erabiltzen diren generoarengatik, bai-

na esanahi bera dutenak. Horrelako kasuetan ere bi hitzetan sentimendu-balentzia bera jartzen ahalegindu gara; aukeran zeuden sentimendu-balentziek uzten zuten heinean. Hala, 4.9 taulako B kasuan, gizonezkoentzat *tonto* eta emakumezkoentzat *tuntun* hitzak ditugu, aipaturiko egoera islatzen dutenak. Ezberdintasun bakarra hitzek erreferentzia egiten dioten pertsonen generoa denez, biei sentimendu-balentzia bera (−3) esleitu diegu.

C- Hizkiak dituzten hitzak. Gaztelaniako lexikoiko hitzei euskarazko ordaina ematean agertu den beste hitz-multzo mota da hizkiak dituztenena. Hizkiak erabiltzearengatik, aurkakotasuna adierazten duten hitzen bikoteak agertu dira ordaina ematean. Hitz bat aurkakotasuna adierazten duen hizkiarekin eta hori gabe agertzea nahiko ohikoa izan da. Aurkakotasuna adierazteko gehien agertu diren hizkiak ezeztapenarekin lotutako hauek izan dira: *des-*, *ez-*, *-ezin* eta *-gabe*. Hitz bera aurkakotasun hizkiarekin eta gabe agertu izan denean, bi egoeretako intentsitate bereko sentimendu-balentzia bera jarri diegu, baina zeinu ezberdinarekin. 4.9 taulan, C kasuan, hitz bera ageri da (*ardura*), baina batean *-dun* atzizkia du, jabetza adierazten duena eta bestean, berriz, *-gabe* atzizkia du, gabezia adierazten duena.

- Ordain *ez-zuzenak*. Euskarazko ordainak ematean agertu den beste fenomenoetako bat ordain *ez-zuzen*ena da. Ordain horiek berez zuzenak dira, baina gure ikerketa-lanerako ez dira baliagarriak testuinguru zehatz batzuetan erabiltzen direlako edota Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) domeinuekin ez dutelako loturarik.

  Adibidez, gaztelaniazko lexikoian agertzen den *provinciano* (−1) hitza, *atzeratu* moduan itzulita zuzena eta erabilgarria izango litzateke guretzat; izan ere, mespretxuzko esanahi bat da eta horiek erabilgarriak dira sentimenduen analisian (Sorby, 2008). Hala ere, *Elhuyar* hiztegia (Elhuyar, 2013) erabilita *Araba, Bizkaia edo Gipuzkoako herritar (gipuzkoarra bereziki)* izan da lortu den ordainetako bat. Ordain hori gure lanerako ez da erabilgarria. Horrenbestez, ordain hori ez dugu aintzakotzat hartu.

- Hizkuntza figuratiboa. Aurkitu dugun beste ordain-mota bat hizkuntza figuratiboa da. Bertan, hitzek edo hitz-multzoek ez dute berez duten esanahia. Hiztunak esanahia moldatu egiten du, bigarren mailako esanahia emanez (Ghosh *et al.*, 2015). Bigarren mailako esanahi horiek zuzenak dira, baina testuinguru jakin eta mugatu batean erabiltzen direnez, horrelako ordainak kontuan ez hartzea erabaki dugu.

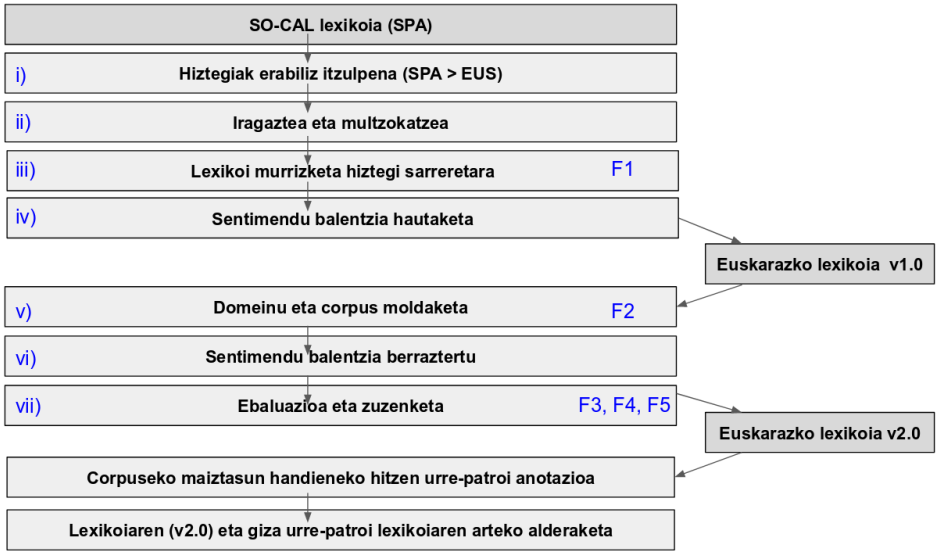
  Adibidez, gaztelaniazko lexikoiko hitzari euskarazko ordaina ematean *beltza* (−2) azaldu da. Hitz horrek bi esanahi ditu: i) beltza, kolorea eta ii) tristea, zoritxarrezkoa. Azken hori esanahi figuratiboa da, bigarren mailako esanahia. Hitzari sentimendu-balentzia aukeratzeko urratsean, euskarazko ordainak gaztelaniazko hitzaren −2 balentzia du jaso du; gaztelaniazko hitzak esanahi figuratiboari egiten diolako erreferentzia. Nahiz eta, hori egokia izan arren, gure lexikoian ez sartzea erabaki dugu: alde batetik, hitzaren erabilera hori ez delako ohiko esanahiarena baino handiagoa eta, beste aldetik, gure corpusera begira ere ez da egokia.

4.2.2.2. Gaztelaniazko lexikoiko hitzari euskarazko ordaina ematean egondako kasuistikak

SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskaratzerakoan, 3.1.2 ataleko metodologiari jarraitu dugu. Hala ere, gaztelaniazko hitzei euskarazko ordaina ematean, gaztelaniazko hitz guztiek ez dute amaiera bera izan. Hau da, euskarazko ordaina eman beharreko hitzaren ezaugarrien arabera, urrats gehiago edo gutxiago bete behar izan dira gaztelaniazko hitzei ordaina emateko.

4.3 irudian azaltzen dira SO-CALen gaztelaniako bertsioko lexikoiko hitzei euskarazko ordainak emateko urratsak[6] eta euskarazko ordaina ematean agertu diren bost fenomenoak. "Fenomeno"[7] hitza aipatzean, gaztelaniazko lexikoiko hitzek urratsei dagokienean hasiera bera, baina amaiera ezberdina izan dutela adierazi nahi dugu. F1-F5 bidez dago adierazita euskarazko ordainetako bakoitzak non bukatzen duen bere ibilbidea.

[6] 4.3 irudian, urratsak zenbaki erromatarren bidez adierazita daude.
[7] 4.3 irudian, fenomenoak Fren bidez adierazita daude.



4.3 irudia: Gaztelaniazko lexikoiko hitzei euskarazko ordaina emateko gauzatutako urratsak.

Jarraian, bost fenomeno horietako bakoitza esplikatuko dugu egitura honi jarraituz: lehenik eta behin, fenomenoa zertan den azalduko dugu; ondoren, fenomeno horren atzean dagoen arrazoia aditzera emango dugu eta, azkenik, fenomeno horren adibide bat emango dugu.

- Hiztegietako sarrerak ez diren euskarazko ordainak (F1).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu da, baina ordaina ez denez *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat ez dugu aintzakotzat hartu.
  - Arrazoia. Sortzen ari garen lexikoiko sarrerek ezaugarri berak izatea nahi dugu eta, horregatik, lexikoian zer sartu erabakitzerakoan muga bat jarri behar izan dugu eta muga hori euskarazko ordaina hiztegiko sarrera bat izatea da.
  - Adibidea. Gaztelaniako *desacreditar* hitzari hiru euskarazko ordain (*ospea_kendu*, *izena_kendu* eta *sona_kendu*) eman dizkiogu eta

ordainetako bakoitzak jatorrizko hitzaren balentzia −2 oinordekotzan hartu du. Baina, ordain horiek ez dira hiztegietako sarrerak eta, horrenbestez, lexikoitik kendu egin ditugu.

- Corpusean agertzen ez diren euskarazko ordainak (F2).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, euskarazko ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, baina ordaina ez da Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen. Horrenbestez, hitz hori ez dugu kontuan hartu euskarazko lexikoiaren bigarren bertsioan[8].
  - Arrazoia. Corpuseko domeinuekin loturarik ez duten hitzak lexikoitik kendu egin ditugu. Modu horretan, lexikoiak corpusarekiko koherentzia izatea nahi dugu. Erabaki horren ondorioz, lexikoia lehen bertsioan 8.140 sarrera izatetik bigarren bertsioan 1.813 sarrera izatera igaro da.
  - Adibidea. Esaterako, *atrofiatu* hitzak −1 balentzia du, baina hitz hori ez da Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) azaltzen. Izan ere, aztertzen ari garen hitza ez da corpusean dauden domeinuetan ohikoa. Hitza corpusean agertzen ez denez, lexikoitik kendu egin dugu; (corpuseko) domenuei loturiko lexikoia eratu nahi baitugu.
- Balentziaren egokitasuna zalantzazkoa duten euskarazko ordainak (F3).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzei euskarazko ordaina eman diegu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da eta ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016)

[8] 4.3 irudian, euskarazko lexikoiaren lehen bertsioa iv) eta v) urratsen artean ageri da eta bigarren bertsioa vii) urratsaren ondoren. Euskarazko lexikoiaren lehen bertsioan, gaztelaniako lexikoiko hitz guztiei eman diegu euskarazko ordaina, baina euskarazko lexikoiaren bigarren bertsioan, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen ez diren euskarazko ordainak kendu ditugu.

agertzen da. Baina, ordainak duen balentziaren egokitasuna zalantzazkoa da eta zalantza argitzeko SO-CAL tresnaren (Taboada *et al.*, 2011) ingeleseko bertsioaren lexikoian agertzen den egiaztatu da eta ez da azaltzen. Ondorioz, lexikoiaren bigarren bertsioan, hitz hori kendu egin dugu, baina lehen bertsioan mantentzen dugu.

- Arrazoia. Euskarazko sarreraren baliokiderik ez dago SO-CALen (Taboada *et al.*, 2011) ingelesezko lexikoian eta gaztelaniazko bertsiotik lortutako balentzia ez da egokitzat jotzen, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden domeinuak kontuan hartuta. Hori dela eta, euskarazko sarrera lexikoitik kentzea erabaki dugu.
- Adibidea. Fenomeno horren adibidea da *seinale* hitza. SO-CALen gaztelaniako lexikoitik (Brooke *et al.*, 2009) −1 balioa hartu du zaio, baina ez dugu egokitzat jo, nahiz eta balentzia intentsitate txikikoa izan. Arrazoi horregatik, hitza lexikoitik kentzea erabaki dugu.

- Arrazoi soziokulturalengatik aintzat hartutako euskarazko ordainak (F4).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen da, baina ordain hori ez da SO-CALen (Taboada *et al.*, 2011) ingeleseko lexikoiko hitz bat. Kalitatea zalantzazkoa ez denez, nahiz eta ingeleseko lexikoian ez agertu, lexikoiaren lehen bertsioan nahiz bigarrenean jaso dugu.
  - Arrazoia. Euskarazko sarreraren baliokiderik ez dago SO-CALen (Taboada *et al.*, 2011) ingeleseko lexikoian eta gaztelaniazko bertsiotik lortutako balentzia egokitzat jotzen da, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden domeinuak aintzakotzat hartuz. Hau da, hitz hori ingelesean ez da hain ohikoa arrazoi soziokulturalengatik, baina bai gaztelanian. Horregatik eta kali-

tatea zalantzazkoa ez denez, euskarazko ordaina aintzakotzat hartzen da. Ondorioz, hitz bera aurreko fenomenoaren egoera berean egon arren, emaitza ezberdina da eta hitza lexikoian mantentzea erabaki dugu.

- Adibidea. Esaterako, euskarazko lexikoiko sarrera bat *frankismo* da eta −2 balioa hartu du gaztelaniazko bertsiotik. Hitz hori espainiar politikari lotutakoa da eta, horregatik, gaztelaniazko bertsioan agertzen da, baina ingelesezkoan ez. Hitzaren ezaugarri hori kontuan hartu dugu eta, horregatik, euskarazko lexikoian mantentzea erabaki dugu.

- Gaztelaniazko eta ingelesezko lexikoietatik eta corpusetik elikatutako euskarazko ordainak (F5).

  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen da, baita SO-CALen (Taboada *et al.*, 2011) ingelesezko lexikoian ere. Egoera horretan, hitza euskarazko lexikoiaren bi bertsioetan dago.
  - Arrazoia. Euskarazko sarreraren baliokidea badago ingelesezko bertsioan ez ezik, gaztelaniazkoan ere badago, eta euskarazkoari balentzia esleitzerakoan ingelesezko bertsiokoa hobesten da.
  - Adibidea. 3.10 taulako kasuan, *zuzen* hitzak gaztelaniazkotik +3 balentzia oinordekotzan hartu du, eta ingelesekoan bi aukera daude: *right* +1 balentziarekin eta *correct* +3 balentziarekin. Gaztelaniazko hitzak (*correcto*) eta ingelesezko hitzak (*correct*) balentzia bera dutenez (+3), *zuzen* hitzak gaztelaniatik oinordekotzan hartutako balentzia mantentzea jo dugu egokitzat.

4.2.2.3. Euskarazko ordainari balentzia aukeratzeko irizpideak

Euskarazko ordainari dagokion sentimendu-balentzia eta gaztelaniako esanahia aukeratzerakoan, jarraian azaltzen den prozedura jarraitu dugu.

- Prozedura 1.
  - Azalpena. Baldin eta jatorrizko gaztelaniazko hitzak euskarazko ordain bakarra eta egokia badu, euskarazko ordainak haren sentimendu-balentzia hartuko du.
  - Adibidea. Hori gertatu da 2. eta 4. Fenomenoetan. Fenomeno horietan, ordain bakarra dago eta ordain bakar hori egokia da, ez delako ez esanahi figuratiboa, ez Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) domeinuekin bat ez datorren ordaina. Hori dela eta, *atrofiatu* hitzari −1 balentzia esleitu diogu eta *frankismo* hitzari −2 balentzia.
- Prozedura 2.
  - Azalpena. Baldin eta euskarazko ordainak jatorrian hainbat gaztelaniazko hitz eta haien balentziak baditu, hitzari esleituko diogun sentimendu-balentzia (eta gaztelaniazko esanahia) Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarrituta egingo dugu. Lan hori corpusean oinarrituta dagoenez, gaztelaniazko esanahien (eta sentimendu-balentzien) artetik aukeraketa egin beharko dugu eta aukeratutakoak bat etorri edo egokia izan beharko Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) hitz hori agertzen den testuinguruarekin. Hori gauzatu ahal izateko hitz-gakoa testuinguruan (*Key Word in Context*, KWIC) erabili dugu.
  - Adibidea. Prozedura hori jarraitu dugu *erakargarri* hitzaren kasuan. Lexikoiko gaztelaniako hitzei euskarazko ordaina eman ondoren, *erakargarri* hitzaren jatorrian hitz eta sentimendu-balentziak hauek daude: *apasionante* +5, *apetecible* +2, *apetitoso* +2 , *atractivo* +2, *fascinante* +5, *glamuroso* +4, *goloso* +3, *interesante* +4, *irresistible* +4, *seductor* +1 eta *tentador* +2. KWIC teknikaren bidez, *erakargarri* hitzaren testuingurua corpusean behatu ondoren, +2 balentzia esleitu diogu, corpuseko testuinguruarekin bat gehien berak egiten duelako.
- Prozedura 3.

– Azalpena. Euskarazko lexikoaren lehen bertsioa euskaraz sortzerakoan, kasuak egon dira non gaztelaniazko hainbat hitzek euskaraz ordain bera duten, eta euskarazko ordain hori corpusean ez den agertzen. Egoera horretan, euskarazko ordainak esanahirik zabalena duen gaztelaniazko hitzaren sentimendu-balentzia esleitu diogu.

– Adibidea. *Bat_ez_datorrena* hitzaren kasua prozedura horren adibidea da. Egitura hori ez da corpusean agertzen; eta, beraz, bere aukeretan (*descontento* −3, *discordante* −2, *insatisfecho* −2) esanahirik erabilienaren (*discordante*) balentzia (−2) hautatu dugu.

4.2.3. Euskarazko sentimenduen lexikoiaren ebaluazioa

Hasteko, 4.2.3.1 azpiatalean, ebaluazioa egiteko prozedura aurkeztuko dugu. Ondoren, 4.2.3.2 azpiatalean, ebaluazio horren emaitzak azalduko ditugu.

4.2.3.1. Ebaluazioa egiteko prozedura

Euskarazko sentimenduen lexikoia ebaluatzeko honako baliabideak erabili ditugu:

- 400 hitzeko zerrenda. *Analhitza* (Otegi *et al.*, 2017) tresna baliatuz, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) gramatika-kategoria bakoitzeko (izenak, adjektiboak, adberbioak eta aditzak) maiztasun handieneko 100 hitz lortu ditugu.
- Urre-patroia. Horretarako, bi etiketatzailek 400 hitzei banan-banan −5 eta +5 arteko sentimendu-balentzia esleitu die.

Euskarazko sentimendu lexikoiak zerrendako 400 hitzi sentimendu-balentzia esleitu die automatikoki eta, ondoren, sentimendu-balentzia horiek urre-patroiko sentimendu-balentziekin alderatu dira. Bi horien arteko adostasuna neurtzeko Pearson korrelazioa (Benesty *et al.*, 2009) erabiltzea erabaki dugu. Etiketatzaile edo lexikoi batek hitz bati −3 balentzia esleitzen dion bitartean, beste etiketatzaile edo lexikoi −1 edo +1 balentzia esleitzen badio, Pearson korrelazioan oinarrituta aintzat hartuko dugu −1 balentziak +1 balentzia baino

hurbiltasun handiagoa duela eta, horrenbestez, bi etiketatzeren arteko adostasuna handiagoa dela. Gainera, kalitatearen neurketa bi modutan egitea erabaki dugu jarraian azaltzen diren bi arrazoiengatik:

- Pearson 1: bi etiketatzaileek etiketatutako hitzak bakarrik kontuan hartuta egindako korrelazio neurketa da. Beraz, 400 hitz horietatik ez dira guztiak erabili neurketa egiteko mementoan.

  4.10 taulan, Pearson 1 deitu dugun korrelazioaren neurketa nola egin den ikus daiteke. Taulan 10 hitz daude, baina horietako hiru (*ageri*, *ar* eta *bakoitza*) ez daude bi etiketatzaileek etiketatuta. Horrenbestez, Pearson korrelazioaren koefizientea kalkulatzean, hiru hitz horiek ez dira kontuan hartu.

Adjektiboak	**Eti1**	**Etik2**	**Etik1**	**Etik2**
ageri	POS		1	
ahul	NEG	NEG	3	3
antisozial	NEG	NEG	1	5
apur	NEG	NEG	1	1
ar				
argi	POS	POS	2	3
aspergarri	POS	NEG	3	3
ausart	POS	POS	3	4
bakar	NEG	POS	1	5
bakoitz				

4.10 taula: Bi etiketatzaileren arteko korrelazioaren kalkulua, etiketatu gabeko hitzak aintzat hartu gabe (Pearson 1).

- Pearson 2: corpusetik ateratako hitz guztiak, 400 hitz guztira, erabili dira. Batzuetan, bi anotatzaileek hitzei sentimendu-balentzia jarri diete. Beste kasu batzuetan, aldiz, anotatzaileak edo anotatzaileek ez dio(te) hitzari sentimendu-balentzia jarri, haien ustez ez duelako sentimendu-balentziarik behar. Horrelako kasuetan, anotatu gabeko hitzei 0 sentimendu-balentzia esleitu diegu. Hori da Pearson 1ekiko ezberdintasun bakarra, anotatu gabeko hitzei 0 sentimendu-balentzia esleitzea. Modu horretan, hitz horiek ere neurketan erabili ditugu.

4.11 taulan, Pearson 2 deitu dugun korrelazioaren neurketa nola egin den azaltzen da. 4.10 taulan bezala, anotatzaileek hiru hitzi (*ageri*, *ar* eta *bakoitza*) ez diete sentimendu-balentzia esleitu. Baina horiek ere neurketan kontuan hartu nahi ditugu eta, horregatik, 0 sentimendu-balentzia esleitu diegu.

Adjektiboak	**Etik1**	**Etik2**	**Etik1**	**Etik2**
ageri	POS	0	1	0
ahul	NEG	NEG	3	3
antisozial	NEG	NEG	1	5
apur	NEG	NEG	1	1
ar	0	0	0	0
argi	POS	POS	2	3
aspergarri	POS	NEG	3	3
ausart	POS	POS	3	4
bakar	NEG	POS	1	5
bakoitz	0	0	0	0

4.11 taula: Bi etiketatzaileren arteko korrelazioaren kalkulua, etiketatu gabeko hitzak ere aintzat hartuz (Pearson 2).

4.2.3.2. Ebaluazioaren emaitza

Jarraian, egin ditugun bi ebaluazioak aurkeztuko dugu. Lehenik eta behin, bi etiketatzaileen arteko korrelazioa azalduko dugu. Ondoren, urre-patroiaren eta euskarazko sentimenduen lexikoiaren arteko korrelazioaz arituko gara. Zerrendako hitzei sentimendu-balentzia esleitzerakoan dagoen bi pertsonen arteko korrelazioa hurrengo 4.12 taulan ikus daiteke.

Gramatika-kategoria	**Pearson 1**	**Pearson 2**
Izena	0,87	0,59
Adjektiboak	0,71	0,60
Aditzondoak	0,93	0,82
Aditzak	0,87	0,76
Guztira	**0,79**	**0,73**

4.12 taula: Bi anotatzaileen arteko Pearson korrelazioaren neurketa.

Pearson 1 balioak erakusten korrelazio-koefizientea altua dela (0,79). Horrek bi anotatzaileek zerrendako hitz askotan hitzei esleitu dieten sentimendu-balentzia nahiko antzekoa dela esan nahi du. Gramatika-kategorien artean badaude ezberdintasun batzuk, korrelaziorik altuena 0,93koa delako eta baxuena, aldiz, 0,71koa. Korrelaziorik altuena aditzondoei dagokie eta baxuena adjektiboei. Pearson 2ri dagokionez, emaitzek erakusten dutenez, korrelazio-koefizienteak altua izaten jarraitzen du, nahiz eta Pearson 1ekin alderatuta zerbait baxuagoak diren. Pearson 2ren kasuan, gramatika-kategorietan koefizienteak 0,82 eta 0,60 artean kokatzen dira.

Kontigentzia-taulak emaitzak interpretatzeko informazio osagarria ematen digu 4.13 taulan. Kontigentzia-taula hori Pearson 2rena da; izan ere, zutabe eta zerrendetan 0 agertzen da, eta horrek adierazten du anotatzaileek hitzari ez diotela sentimendu-balentzia esleitu. Bertan ikusten denez, bi anotatzaileen arteko ezberdintasuna nagusiki egoera batean gertatzen da: anotatzaile batek hitzari balentzia esleitzen dionean besteak ez dio esleitzen, eta alderantziz. Egoera hori desadostasun guztien % 90,19ren jatorrian dago (ezberdintasuna dagoen 102 instantzietatik 92etan, hain zuzen ere).

Guztira kategoriak			
	0	**Negatiboa**	**Positiboa**
0	187	12	27
Negatiboa	14	42	5
Positiboa	39	5	69

4.13 taula: Bi anotatzaileen arteko kontigentzia-taula.

4.14 taulan, berriz, sentimenduen lexikoiaren eta urre-patroiaren arteko korrelazio-koefizienteak ageri dira. Urre-patroiaren eta sentimenduen lexikoiaren arteko korrelazio-koefizienteen neurketak ezberdintasun batzuk erakusten ditu aurretik egindako bi anotatzaileen arteko korrelazio koefizientearekiko.

Pearson 1i dagokionez, hau da, sentimendu-balentzia esleitu zaien hitzak bakarrik kontuan hartuta, korrelazio-koefizientea aurreko neurketatik gertu dago. Kasu horretan, koefizientea 0,76 da eta aurreko neurketan 0,79. Koefizientean ezberdintasun handiak daude gramatika-kategorien artean. Koefizienterik altuena izenetan dago (0,96) eta baxuena, berriz, aditzetan (0,69).

Gramatika-kategoria	Pearson 1	Pearson 2
Izena	0,96	0,59
Adjektiboak	0,78	0,56
Aditzondoak	0,75	0,47
Aditzak	0,69	0,54
Guztira	**0,76**	**0,54**

4.14 taula: Euskarazko sentimenduen lexikoa (V2.0) eta urre-patroiaren arteko Pearson korrelazioaren neurketa.

Baina, ezberdintasun esangurantsua Pearson 2n gertatzen da, bi anotatzaileen arteko korrelazio koefizientearekin alderatzen denean. Pearson 2an koefizientea 0,54 izan da; aurreko neurketan 0,73 izan denean (0,19ko ezberdintasuna). Koefizienterik altuena izenenak izaten jarraitzen du (0,59) eta baxuena, berriz, aditzondoena da (0,47).

Datu hauen gure interpretazioa hauxe da: lexikoiaren eta urre-patroiaren arteko Pearson 1 koefizienteak altua izaten jarraitzen du zerrendako hitzei esleitu zaien balentzia antzekoa izan delako bi kasuetan, batean, bi anotatzaileen artean eta bestean, urre-patroia eta lexikoiaren artean. Baina, ezberdintasuak daude balentzia esleitu zaien zerrendako hitzetan. Bi anotatzaileen kasuan, balentzia esleitu zaien hitzak kasik berak izan dira. Hemen, urre-patroiaren eta lexikoiaren arteko emaitzak kontuan hartuz, balentzia esleitu zaien hitzak beti ez dira berak izan.

Kontigentzia-taulak lexikoiaren eta urre-patroiaren arteko ezberdintasunak argitara ematen ditu 4.15 taulan. Bi etiketatzaileen artean ezberdintasunak balentzia esleitu beharreko hitzaren inguruan daude. Batek hitz bati balentzia esleitzen dionean, besteak ez egitea, eta alderantzizko kasua. Lexikoiak edo urre-patroiak hitz bati balentzia esleitzean besteak balentzia ez esleitzea ezberdintasun guztien % 89,83 (118tik 106 kasutan) izan da. Bi anotatzaileek hitzei sentimendu-balentzia esleitzean ere gertatu da anotatzaile batek hitzari balentzia esleitzean beste anotatzaileak ez esleitzea, baina askoz intentsitate baxuago batean, kasu horretan ez bezala. Baina, bi pertsonen arteko emaitzak alderatuta badago beste ezberdintasun bat. Bi pertsonen artean, batek balentzia esleitzean besteak ez egitea, eta alderantziz, nahiko

Guztira kategoriak			
	0	**Negatiboa**	**Positiboa**
0	195	2	15
Negatiboa	30	34	8
Positiboa	59	4	53

4.15 taula: Euskarazko sentimenduen lexikoa (V2.0) eta urre-patroiaren arteko kontigentzia-taula.

orekatua izan da. Hemen, ordea, joera bat nagusitzen da: kasu gehienetan, batek (urre-patroiak) zerrendako hitzari sentimendu-balentzia esleitzen dionean, besteak (lexikoiak) ez du egiten. Horrek lexikoia urre-patroiaren aldean kontserbadoreagoa eta zorrotzagoa dela aditzera ematen du. Lexikoiak askozaz ere hitz gutxiagori esleitzen dien balentzia, Pearson 2ko korrelazio koefizientea baxuagoa da.

Laburbilduz, Pearson 1 korrelazio-koefizientea altua eta antzekoa izan da bai bi anotatzaileen artean, bai lexikoiaren eta urre-patroiaren artean. Horrek hitzei esleitu zaien sentimendu-balentzia (−5etik +5era) antzekoa (0,79 vs. 0,76) izan dela esan nahi du. Pearson 2 korrelazio koefizientea, aldiz, altua izan da bi anotatzaileen artean, baina jaitsiera bat egon da lexikoiaren eta urre-patroiaren artean (0,73 vs 0,54). Lexikoiak urre-patroiarekin (eta bi anotzaileekin) alderatuz, hitz gutxiagori esleitu die sentimendu-balentzia eta, horregatik, da koefizientea baxuagoa.

1 Ikerketa hipotesia

1.2 ataleko ikerketa hipotesiari erantzunez, emaitzek erakusten dute posible dela gaztelaniatik euskara sentimenduen lexikoi bat itzultzea eta bera baliagarria izatea.
Pearson korrelazioek erakusten dutenez, itzulpenaren ondoren, urre-patroiaren eta sentimenduen lexikoiaren adostasuna 0,76 koefizientekoa hitzei sentimenduen balentzia esleitzekoan. Hala ere, badaude hobetzeko alderdi batzuk, besteak beste, garaturiko sentimenduen lexikoian sarrera gehiago gehitzea falta delako.

4.3 Euskarazko sentimenduen sailkatzailea

Lehenik eta behin, guk sortu dugun sentimenduen sailkatzailearen oinarrian dagoen SO-CAL tresnaren ezaugarriak deskribatuko ditugu (4.3.1 atala). Ondoren, euskarazko sentimenduen sailkatzailea sortzeko SO-CALen egindako moldaketak azalduko ditugu (4.3.2 atala). Azkenik, euskarazko sentimenduen sailkatzailearen ebaluazioaren emaitzak aurkeztuko ditugu (4.3.3 atala).

4.3.1. SO-CAL izeneko sentimenduen sailkatzailearen ezaugarriak

Guk sortu dugun euskarazko sentimenduen sailkatzailearen oinarrian SO-CAL tresna (Taboada *et al.*, 2011) dago. Tresna hori hasiera batean ingeleserako sortu bazen ere; gaur egun beste hizkuntzetako bertsioak ere badaude, esaterako txinerarena (Miao *et al.*, 2013).

Tresna horren oinarrian sentimenduen lexikoia dago. Sentimenduen lexikoiak lau gramatika-kategorietako hitzak (izenak, adjektiboak, aditzak eta adberbioak) biltzen ditu. Hitz horiek −5 eta +5 arte sentimendu-balentzia dute, 4.16 taulan ikusten den moduan.

Gramatika-kategoria horretako hitzez gain, intentsifikatzaileak ere badaude eta horiek hitzen sentimendu-balentzian aldaketak eragiten dituzte. Intentsifikatzaileen hitzen sentimendu-balentzia % +100 eta % −50 artean indartu edo ahuldu dezakete 4.17 taulan, intentsifikatzaile horietako batzuk ageri dira.

Intentsifikatzaileen eta sentimendu-balentziadun hitzen artean (4.17) adibidean bezalako eragiketako gertatzen dira. *Sleazy* ("zikin") hitzak −3 sentimendu-balentzia du baina *somewhat* ("zerbait") intentsifikatzailearen ondorioz, bere sentimendu-balentzia −3tik −2, 1era jaisten da.

(44) Sleazy: (−3), somewhat sleazy: −3 x (%100 − %30)= −2, 1. (Taboada *et al.*, 2011, 275 orr.)

SO-CAL tresnak (Taboada *et al.*, 2011) ezeztapena ere tratatzen du eta, ho-

Hitza	Sentimendu-balentzia
monstrosity	-5
hate (izena eta aditza)	-4
disgust	-3
sham	-3
fabricate	-2
delay (izena eta aditza)	-1
determination	+1
inspire	+2
inspiration	+2
endear	+3
relish (aditza)	+4
masterpiece	+5

4.16 taula: SO-CAL tresnako lexikoiaren zenbait sarrera (Taboada *et al.*, 2011).

Intentsifikatzailea	Modifikatzailea (%)
slightly	-50
somewhat	-30
pretty	-10
really	+15
very	+25
extraordinarily	+50
(the) most	+100

4.17 taula: SO-CAL tresnako zenbait intentsifikatzaile (Taboada *et al.*, 2011).

rretarako, 5.2.2 atalean azaldutako desplazamendu-ezeztapena izeneko hurbilpena erabiltzen du. (45) adibidean, tresnak ezeztapena nola laguntzen duen ikus daiteke. Ezeztapenak −4 balentzia du esaldiak orientazio semantiko positiboa duenean eta −4 esaldiak orientazio semantiko negatiboa duenean. (45) adibideko lehen zatian *terrific* (“‘bikaina”) hizak −5 sentimendu-balentzia du baina ezeztapenak sentimendu-balentzia hori +1era aldatzen du. Adibide bereko bigarren zatian, aldiz, *terrible* (“beldurgarria”) hitzak −5 sentimendu-balentzia du baina ezeztapenak balentzia hori +1ean bihur-

tzen du.

(45) He's not terrific (5 − 4 = 1) but not terrible (−5 + 4 = −1) either. (Taboada *et al.*, 2011, 277 orr.)

Sentimendu lexikoiaren beste ezaugarrietako bat da sentimenduen analisirako baliagarriak ez diren esaldiak identifikatzea eta horiei sentimendu-balentzia ez esleitzea. Sentimenduen analisirako baliagarriak ez diren esaldiei *irrealis* edo testuinguru ez-faktualak deritze. Ingelesean, testuinguru ez-faktuala hitzen hurrenkeraren, aditz modalaren edota aginteraren bidez agertzen da. SO-CAL tresnak (Taboada *et al.*, 2011) horrelako elementuak aurkitzen dituenean, esaldi horietako hitzen sentimendu-balentzia ez du kontuan hartzen. Komatxoen artean agertzen diren hitzen sentimendu-balentziak ere ez ditu kontuan hartzen.

Esaterako, (46) adibidean, *great* ("sekulako") hitzak +3 sentimendu-balentzia du baina esaldiak testuinguru ez-faktuala duenez, bertako hitzen sentimendu-balentzia ez aintzat hartzen.

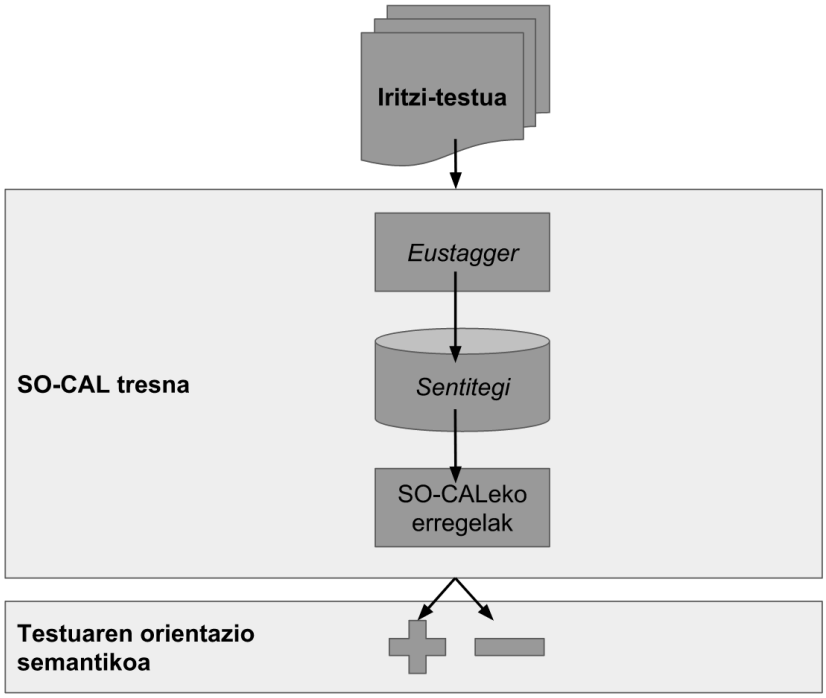
(46) This should have been a great movie. (+3 = 0). (Taboada *et al.*, 2011, 279 orr.)

Kennedy eta Inkpenek (2006) lexikoian oinarritutako sentimenduen sailkatzaileek orientazio semantiko positiboa esleitzeko joera antzeman zuten eta, horregatik, tresna honetan sentimendu-balentzia negatiboa duten hitzei pisua esleitu zaie. Hain zuzen, % +50eko pisua esleitzen zaie hitz horiei. Gainera, sentimendu-balentziadun hitz bat testu batean hainbat aldiz errepikatzen bada, errepikapen horietako bakoitzean sentimendu-balentziadun hitz horrek bere balentzia galtzen du ($1/n$ eragiketa, n= errepikapen-kopurua). Erregela horren adibidea euskaran 4.7 irudian ikus daiteke.

Azkenik, SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko bertsioak zenbait erregela morfologiko ditu, testuetako hitzei flexio-morfemak kendu eta lema moduan uzten dituetan. Hitza lematizatuta dagoenez, tresnako lexikoiak hitzari sentimendu-balentzia esleitu diezaioke, baldin eta hitza lexikoian badago.

4.3.2. Sentimenduen sailkatzailaren arkitektura

SO-CAL izeneko sentimenduen sailkatzailearen (Taboada *et al.*, 2011) euskarazko bertsioa garatzeko zenbait aldaketa egin behar izan ditugu. Aldaketa horiek 4.4 irudian daude ikusgai.



4.4 irudia: SO-CAL sentimenduen sailkatzailearen euskarazko bertsioa.

Euskarazko bertsioa sortzeko egin dugun lehen aldaketa *Eustagger* (Aduriz *et al.*, 2003) integratzea izan da. Ingelesak, hizkuntza analitikoa denez, erregela morfologiko gutxi behar ditu flexio-morfemak kendu eta hitzak lematizatuta uzteko. Euskara, ordea, morfologikoki aberatsa da eta, horregatik, testuetako hitzak lematizatzeko *Eustagger* (Aduriz *et al.*, 2003) integratu dugu.

Gainera, ingelesezko sentimenduen lexikoiaren ordez, guk sortu dugun *Sentitegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) integratu dugu. Gure sentimenduen lexikoiak izen, adjektibo, aditz eta adberbio gramatika-

kategorietakoak biltzen ditu.

Azkenik, SO-CAL tresnak (Taboada *et al.*, 2011) berak dituen bestelako erregelak bere horretan utzi ditugu. Erregela horiek sentimendu-balentzia negatiboa duen hitzei pisua esleitzekoa, testu batean hainbat aldiz errepikatzen diren hitzei balioa kentzekoa edota komatxoen artean dauden hitzen sentimendu-balentziak kontuan ez hartzekoak dira.

Euskarazko sentimenduen lexikoia garatzeko aldaketa horiek egin ostean, jarraian azalduko dugun moduan ibiltzen da euskarazko sentimenduen lexikoia. 4.5 irudian, euskarazko sentimenduen sailkatzaileak aztertuko duen iritzi-testua ageri da. Ikusten den moduan, testua gordina da.

```
Asteburu polita orokorrean.
Ez dugu Penintsula hegoaldean dagoen beroa eta ezta nahi ere. Tenperatura
    atseginekin doa uda eta horrelaxe jarraituko dugu asteburuan ere.
    Gainera, oro har eguraldia ondo portatuko da. Hala ere, larunbata
    hodeiekin hasiko dugu eta baliteke zirimiri pixka bat egitea ere,
    bereziki goizaldean. Arratsalderako ordea, argituko du eta eguzkia
    ikusiko dugu. Iparraldeko haizea ibiliko da eta tenperatura maximoa 21 C
    -koa izango da, atsegina.
Igandean eguzkia nagusituko da berriro ere, baina udan askotan gertatzen den
     bezala, goizaldean hodei baxuak sortuko dira seguruenik eta hodei
    hoiekin esnatzea posible da, baina goizean zehar eta larunbatean baino
    azkarrago, ostarteak zabaltzen hasiko dira, eguraldi eguzkitsua geratuz
    eguerditik aurrera. Tenperatura pixka bat igo egingo da, maximoa 24 C-ra
     iritsiz.
Esandakoa gure betiko uda naturalarekin jarraituko dugu: hodei baxuak,
    eguzkia eta tenperatura atseginak, berorik gabe.
```

4.5 irudia: EGU40 iritzi-testua.

Euskarazko sentimenduen sailkatzaileari 4.5 irudiko iritzi-testua ematean, tresnak lehendabizi testua lematizatzen du, eta ondoren, bertako hitzei sentimendu-balentzia esleitzen die. 4.6 irudian, iritzi-testua lematizatuta dago eta esaldien eskuinetara esaldietako hitzei esleitu dien sentimendu-balentzia agertzen da.

Hurrengo urratsean, sentimendu-balentzia esleitu dien hitzei zenbait erregela aplikatzen dizkie, hala nola sentimendu-balentzia negatiboa duten hitzei pisua esleitzekoa, hainbat aldiz errepikatzen diren hitzei balioa kentzekoa eta komatxoen artean dauden sentimendu-balentziadun hitzak kontuan ez hartzekoa.

```
asteburu polit orokor . 7.0
tenperatura atsegin joan uda eta horrelaxe jarraitu *edun asteburu ere . 3.3
gain , oro har eguraldi ondo portatu izan . -1.5
hala ere , larunbat hodei hasi *edun eta *edin zirimiri pixka bat egin ere ,
     bereziki goizalde . 0
arratsalde ordea , argitu *edun eta eguzki ikusi *edun . 2.0
iparraldeko haize ibili izan eta tenperatura maximo 21 C-koa izan izan ,
    atsegin . 4.0
igande eguzki nagusitu izan berriro ere , baina uda asko gertatu izan bezala
     , goizalde hodei baxu sortu izan seguru eta hodei hoiekin esnatu
    posible izan , baina goiz zehar eta larunbat baino azkar , ostarte
    zabaldu hasi izan , eguraldi eguzkitsu geratu eguerdi aurrera . 5.8
tenperatura pixka bat igo egin izan , maximo 24 C-ra iritsi . 0.5
esan gu be uda natural jarraitu *edun : hodei baxu , eguzki eta tenperatu
    atsegin , bero gabe . 1.75
```

4.6 irudia: EGU40 iritzi-testua lematizatuta eta hitzei sentimendu-balentziak esleituta.

4.7 irudian, euskarazko sentimenduen sailkatzaileak iritzi-testuko sentimendu-balentziadun hitzei aplikatutak erregelak ikus daitezke. Ikusten denez, zenbait hitzek sentimendu-balentzia negatiboa dute eta horiei pisua esleitu die (NEGATIVE erregela). Ondorioz, esaterako, adjektiboetan *baxu* hitzaren sentimendu-balentzia -3tik $-4,5$era pasa da. Beste erregela batek, testu berean hainbat aldiz errepikatzen diren hitzei balioa kentzen die. *Atsegin* horren adibidea da. Lehen aldiz agertu denez, tresnak *atsegin* hitzari $+1$ sentimendu-balentzia esleitu dio, baina bigarren aldiz agertu denean, hitzari pisua kendu dio (REPEATED erregela) eta bere sentimendu-balentzia $+1$etik $+0,5$era igaro da.

Sentimendu-balentziadun hitzei erregela horiek aplikatu ostean, tresnak iritzi-testuari sentimendu-balentzia esleitzen dio, 4.8 irudian ikusten den moduan. Kasu honetan, iritzi-testuaren sentimendu-balentzia $+1,04$ da; beraz, iritzi-testuak balorazio positiboa egiten du.

4.3.3. Sentimenduen sailkatzailearen ebaluazioa

Atal honetan, sentimenduen sailkatzailaren ebaluazioaren emaitzen berri emango dugu. Ebaluazioaren emaitzak hainbat modutara emango ditugu: asmatutako eta ez asmatutako testuen orientazio semantikoak, orientazio semantiko

```
Text Length: 168
---------
Nouns:
-----
har -1.0  X 1.5 (NEGATIVE) = -1.5
-----
Average SO: -1.5
-----
Verbs:
-----
jarraitu *edun asteburu ere 1.0 X 1.3 (INTENSIFIED)  = 1.3
argitu 2.0  = 2.0
ibili 2.0  = 2.0
atsegin 1.0  = 1.0
nagusitu izan berriro ere 1.0 X 1.3 (INTENSIFIED)  = 1.3
sortu 2.0  = 2.0
jarraitu 1.0 X 1/2 (REPEATED)  = 0.5
atsegin 1.0 X 1/2 (REPEATED)  = 0.5
-----
Average SO: 1.325
-----
Adjectives:
-----
polit 4.0  = 4.0
orokor 3.0  = 3.0
atsegin 2.0  = 2.0
maximo 1.0  = 1.0
baxu -3.0  X 1.5 (NEGATIVE) = -4.5
seguru 1.0  = 1.0
posible 1.0  = 1.0
azkar 2.0  = 2.0
eguzkitsu 3.0  = 3.0
maximo 1.0 X 1/2 (REPEATED)  = 0.5
natural 2.0  = 2.0
baxu -3.0 X 1/2 (REPEATED)  X 1.5 (NEGATIVE) = -2.25
bero 1.0  = 1.0
-----
Average SO: 1.05769230769
-----
```

4.7 irudia: EGU40 iritzi-testuko hitzei sentimenduen sailkatzailearen erregelak aplikatuta.

```
---------
Total SO: 1.03863636364
---------
```

4.8 irudia: EGU40 iritzi-testuaren sentimendu-balentzia.

jakin bakoitzean sailkatzaileak duen asmatze-tasa eta, azkenik, sailkatzaileak corpuseko domeinu bakoitzean duen asmatze-tasa.

4.18 Taulako emaitzen arabera, sailkatzailearen asmatze-tasa 0,71koa da. Izan ere, corpuseko test zatian dauden 48 testuetatik 34 iritzi-testuren orientazioa semantikoa ondo esleitu baitu. Iritzi-testuek duten orientazio semantikoaren ikuspegitik emaitzak aztertuta, emaitzek bestelako interpretazioak egiteko aukera ematen dute.

Emaitzak	**Testu-kopurua**	%
Ondo	34	0,71
Gaizki	14	0,29

4.18 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak.

4.19 taulako emaitzetan, tresnaren funtzionamenduari buruzko emaitza argigarriak ageri dira. Ikusten den moduan, sentimenduen sailkatzailearen asmatze-tasa orientazio positiboko iritzi-testuetan 1ekoa da. Emaitzak zeharo ezberdinak dira orientazio negatiboko iritzi-testuen kasuan. Mota horretako iritzi-testuekin, tresnaren asmatze-tasa 0,33koa da; izan ere, 24 orientazio negatiboko iritzi-testu egon arren, 8 iritzi-testu bakarrik identifikatu ditu negatibo moduan. Beraz, esan liteke, sentimenduen sailkatzaileak joera bat duela iritzi-testuei orientazio semantiko positiboa esleitzeko. Joera hori bat dator Alistair eta Dianak (2005) dioenarekin. Diotenez, lexikoian oinarritutako sentimenduen sailkatzaileek orientazio semantikoa esleitzeko joera erakusten du eta hori lotuta dago unibertsalki gizakiek hizkuntza positiboa erabiltzeko duten ohiturarekin Boucher eta Osgood (1969).

Testuen orientazio semantikoa	**Asmatutako testu-kopurua**	%
Positiboak	24/24	1
Negatiboak	24/8	0,33

4.19 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak iritzi-testuen orientazio semantikoa aintzat hartuta.

Azkenik, emaitzak domeinuen arabera aztertuta, domeinu guztietan asmatze-tasa antzekoa dela ikus daiteke, literaturan izan ezik.

4.20 taulak erakusten duenez, domeinuetan asmatze-tasa txikiena eguraldian, musikan eta politikan dago; asmatze-tasa 0,63koa da. Kirola eta zinema tartean kokatzen dira 0,75eko asmatze-tasarekin eta, azkenik, emaitza onenak literaturak lortzen ditu, testu baten orientazio semantikoa izan ezik, beste guztiena asmatu baitu tresnak bertan (0,88ko asmatze-tasa). Domeinutik domeinura ez dago ezberdintasun handirik, domeinu bakoitzean orientazio semantiko positibo eta negatiboko iritzi-testu kopuru bera dagoelako.

Domeinua	**Testu-kopurua**	**%**
Eguraldia	8/5	0,63
Musika	8/5	0,63
Politika	8/5	0,63
Kirola	8/6	0,75
Zinema	8/6	0,75
Literatura	8/7	0,88

4.20 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak domeinuen arabera.

4.4 Laburpena

Kapitulu honetan, tesi-lanean garatutako baliabideak zein diren azaldu dugu. Lehenik eta behin, Euskarazko Iritzi Corpusa izeneko sei domeinutako 240 iritzi-testu biltzen dituen corpusa (Alkorta *et al.*, 2016) eratu dugula aipatu behar da.

Bigarrenik, sortu dugun *Sentitegi* izeneko euskarazko sentimenduen lexikoiaren (Alkorta *et al.*, 2018) berri ere eman dugu. Ikusi den moduan, sentimenduen lexikoiaren bi bertsio sortu ditugu, bata domeinuari lotu gabea (8.140 sarrerakoa) eta bestea, corpuseko domeinuei lotutakoa (1.237) eta izaera murriztaileagoa duena, kolokazioak eta adierazpenak ez baititu kontuan hartzen. Sorkuntzak emandakoa esplikatu ondoren, ebaluatu egin dugu. Horretarako, Euskarazko Iritzi Corpuseko maiztasun handieneko lehen 100 hitzak hartu ditugu, bi anotatzaileek hitzei sentimendu-balentzia esleitu die eta ondoren, bertatik urre-patroia sortu da. Jarraian, urre-patroi hori euskarazko sentimenduen lexikoiak zerrendako hitzei eman dien sentimendu-balentziarekin alderatu dugu. Emaitzek erakusten dutenez, sentimenduen lexikoiko hitzek duten sentimendu-balentzia egokia da (Pearson korrelazioa 0,76koa delako), baina berez sentimendu-balentzia eduki beharko luketen hitz batzuei lexikoiak ez die balentzia esleitu eta, ondorioz, Pearson korrelazioa 0,54koa da.

Azkenik, tesi-lan honen hirugarren emaitza euskarazko sentimenduen sailkatzailea da. Euskarazkoa sortzeko ingelesezko SO-CAL sentimenduen sailkatzailean oinarritu gara eta zenbait aldaketa egin behar izan ditugu. Esaterako, *Eustagger* testu lematizatzailea (Aduriz *et al.*, 2003) inplementatu dugu, baita *Sentitegi* euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018) ere. Sentimenduen sailkatzaile hori ebaluatu egin dugu eta, emaitzek erakusten dutenez, tresnak kasuen % 71etan iritzi-testuaren orientazio semantikoa ondo sailkatzen du.

BALENTZIA ALDATZAILEAK HIZKUNTZA MAILA EZBERDINETAN

5. KAPITULUA

Balentzia-aldatzaileak

Kapitulu honetan, hizkuntzako maila ezberdinetan dauden euskarazko balentzia-aldatzaileak identifikatzeko lanaren emaitzak azalduko ditugu. Guztira lau hizkuntza maila landu ditugu: fonologiko eta morfologikoa (5.1 atala), sintaktikoa (5.2 atala) eta diskurtsokoa (5.3 atala).

5.1 Fonologia eta morfologia maila

5.1.1 atalean, balentzia-aldatzaile fonologiko eta morfologikoen sailkapena hiru modutan aurkeztuko dugu. 5.1.2 atalean, hainbat balentzia-aldatzaile morfologiko dituzten hitzei buruz arituko gara, 5.1.3 atalean, eragin ezberdinetako balentzia-aldatzaileen eta orientazio semantiko ezberdinetako hitzen arteko konbinazioak azalduko ditugu. Azkenik, 5.1.4 atalean, maila fonologikoak eta morfologikoak euskaran duten garrantzia nabarmenduko dugu.

5.1.1. Balentzia-aldatzaileen sailkapena

5.1 taulan, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako balentzia-aldatzaile fonologiko eta morfologikoen zerrenda ageri da maiztasunean oinarrituz sailkatuta. Guztira 32 balentzia-aldatzaile aurkitu ditugu eta horiek guztiak 1.623 aldiz agertzen dira corpusean. Gehienak atzizkiak dira eta hiru bustidura adierazkor eta aurrizki mota ageri dira.

Balentzia-aldatzaileek sentimendu-balentzian duten eraginean oinarritutako sailkapena 5.2 taulan dago. Bertan, balentzia-aldatzaileak lau multzotan

Hizkia	**Maiztasuna**	**Adibidea**
-tasun	307	zailtasun
-garri	250	barregarri
-tsu	170	gatazkatsu
-en	163	handien
-gabe	71	akatsgabe
des-	62	desorekatu
-egi	61	bikainegi
-dun	60	berezidun
-gile	60	inbertsiogile
-keria	57	handikeria
-s-/-z- >-x-	39	goxo
-ez/ez-	39/0	ezegonkor
-kide	38	garaikide
-ezin/ezin-	36/3	garaiezin
-xe/-txe	30	oraintxe
-txo	27	erasotxo
-ezia	27	dotorezia
-zale	27	bizizale
-gintza	16	osasungintza
-t- >-tt-	12	punttu
-gai	12	osagai
-gaitz	10	ulergaitz
-gura	8	logura
-gailu	8	neurgailu
-min	7	ikusmin
-ska/-xka	6	herrixka
-z- >-tx-	6	txoratuta
-tza	4	burujabetza
-tzar	3	krisitzar
-nahi	2	handinahi
-txa	1	neskatxa
a-	1	atenporalitatea
Guztira	**1.623**	

5.1 taula: Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako balentzia-aldatzaile morfologiko eta fonologikoak.

banatu ditugu: i) balentzia indartzen dutenak, ii) ahultzen dutenak eta iii) balentzian eraginik ez dutenak.

Indartu	**Ahuldu**	**Eraginik ez**
-garri	-gabe	-ezia
-tsu	des-	-tasun
-en	-egi	
-s-/-z- >-x-	-keria	
-zale	-ez/ez-	
-t- >-tt-	-ezin/ezin-	
-gura	-xe/-txe	
-min	-txo	
-z- >-tx-	-gaitz	
-tza	-ska/-xka	
-tzar	-txa	
-nahi	a-	

5.2 taula: Balentzia-aldatzaile morfologiko eta fonologikoak sentimendu-balentzian duten eraginaren arabera sailkatuta.

Bukatzeko, 5.3 taulan, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako hizkiak duten balio semantikoa irizpidetzat hartuta sailkatuta ageri dira. Bertan, hizkiak eta bustidura adierazkorra 12 multzo semantikotan banatu ditugu. Lehenik eta behin, ezeztapenarekin lotutakoak daude. Ezeztapenezko hizki horiek ezeztapen-marka moduan antzeko modu batean eragiten diote sentimendu-balentziari, baina hitzari eragin beharrean, esaldi edo sintagmari eragiten diote. Multzo horretan aurrizkiak eta atzizkiak aurki daitezke.

Datorren multzoan konparazioarekin lotutako hizkiak ageri dira: *-egi* hizkia, sentimendu-balentzia ahultzen duena eta *-en* hizkia, sentimendu-balentzia indartzen duena.

Hurrengo bi multzoak tamainarekin lotutakoa dira. Semantikoki handigarri funtzioa betetzen duen balentzia-aldatzaile bakarra aurkitu dugu: *-tzar*. Txikigarri funtzioa duten balentzia-aldatzaileak, aldiz, asko dira eta denak atzizkiak dira. Multzoa esanahiarekin lotura duten bi balentzia-aldatzaile aurkitu ditugu: *-tsu* eta *-tza* eta biek sentimendu-balentzia indartzen dute.

Ezeztapena	**Konparazioa**	**Handigarria**	**Txikigarria**
des-	-egi	-tzar	-xe/-txe
-ez/ez-	-en		-txo
-ezin/ezin-			-ska/-xka
-gabe*			-txa
(a-)			
Zailtasuna	**Nolakotasuna**	**Lanbidea**	**Jabetza**
-gaitz	-tasun	-gile	-dun
-garri	-keria	-gintza	-gabe*
	-ezia		
Nahia	**Hurbiltasuna/samurtasuna**	**Multzoa**	**Bestelakoak**
-gura	-s-/-z- >-x-	-tza	-zale
-nahi	-t- >-tt-	-tsu	
-min	-z- >-tx-		

5.3 taula: Euskarazko Iritzi Corpusetik lortutako hizkien sailkapen semantikoa.

Zailtasunarekin lotura duten beste bi balentzia-aldatzaile ere badaude: *-gaitz* eta *-garri*. Lehenak sentimendu-balentzia ahultzen du eta bigarrenak, berriz, sentimendu-balentzia indartzen du.

Bestalde, nolakotasuna adierazten duten hiru balentzia-aldatzaile daude, baina, *-keria* hizkiak sentimendu-balentzia ahultzen du eta *-tasun* eta *-ezia* atzizkiek sentimendu-balentzia indartzen dute.

Jabetza adierazten duten bi atzizki daude: *-dun* eta *-gabe*. Bigarren atzizkiak semantikoki jabetza edo ezeztapena adieraz dezake eta, ondorioz, bi multzoetan sailkatu dugu. Azken honek sentimendu-balentzia ahultzen du. Bustidura adierazkorrei dagokienez, hirurek semantikoki hurbilpena edota samurtasuna adierazten dute eta sentimendu-balentzia indartzen dute.

Azkenik, aipaturiko multzoekin zerikusirik ez duten balentzia-aldatzaile bakarra dago. Hizkia *-zale* da. Zaletasuna adierazten du eta sentimendu-balentzia indartzen du.

5.1.2. Hainbat balentzia-aldatzaile morfologiko dituzten hitzak

Hitz batean hainbat hitz agertzen direnean, hitz horrek egitura jakin bat du. Bi ezaugarri dituzte mota horretako hitzek:

- Hainbat balentzia-aldatzaile morfologiko eta balentzia-aldatzaile fonologikoa den bustidura adierazkorra batera agertzen diren hitzik ez dago. Hau da, hitz batean hainbat balentzia-aldatzaile morfologiko badaude, bustidura adierazkorrik ez da agertzen.
- Hitz batean gehienez hiru balentzia-aldatzaile morfologiko ager daitezke batera bata bestearen atzetik. Eta bitartean, hitzaren sentimendu-balentzian eta orientazio semantikoak aldaketak gertatzen dira.

5.4 taulan, balentzia-aldatzaile morfologiko bat baino gehiago dituzten hitzen egituraketa ageri da. Guztira, mota horretako 36 hitz eta 11 hizki aurkitu ditugu. Gehienez hiru hizki ager daitezke elkarren segidan eta hizkiak erantsi ahala hitzaren sentimendu-balentziak aldaketa ezberdinak jasaten ditu. Hirugarren mailara iristeko ezinbestekoa den bigarren mailan *-en* hizkia egotea. Eta bigarren mailatik hirugarrenera balentziaren ahultzea gertatzen da, nahiz eta hasieratik egoeratik balentzian indartze bat dagoen.

Hizkiak konbinatzeko aukerak eta beren ondorioak askotarikoak dira. Posible da hizkiak gehitu ahala hitzaren sentimendu-balentzia indartzea. Hori gertatzen da *prestigiotsuen* hitzean, *-tsu* eta *-en* hizkiek hitzaren balentziaren positibotasuna indartzen baitute. Hizkiak gehitu ahala hitzaren sentimendu-balentzia negatiboago bilakatzen ere joan daiteke. *Lotsagabekeria* hitzean, *-gabe* eta *-keria* hizkiek *lotsa* hitzaren sentimendu-balentzia ahuldu eta negatibo bilakatzen dute. Tarteko zerbait ere gerta daiteke, hots, hitzaren balentzia hasieran indartzea eta gero ahultzea. *Berritsukeria* hitzaren kasuan, *-tsu* hizkiak hitza positiboago bilakatzen du eta ondoren, *-keria* hizkiak positibotasun hori negatibo bilakatzen du.

5.4 taulan eta azalpenean adibide moduan jarritako hitzak (adibidez, *berritsu* eta *berritsukeria*) *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegietan egon daitezke. Hiztegietako sarrerak direnez, sentimenduen lexikoian

Maila	Balentzia	Hizkia
0	handi gehi harritu malenkonia aspertu berri prestigio lotsa garrantzi axola garbiz zehatz	
1	handinahi gehiegi harrigarri malenkoniatsu aspergarri berritsu prestigiotsu lotsagabe garrantzitsu axolagabe garbizale zehazgabe	-nahi (+) -egi (-) -garri (0) -tsu (+) -gabe (-) -zale (+)
2	handinahikeria gehiegikeria harrigarrien malenkoniatsuago aspergarriago berritsukeria prestigiotsuen lotsagabekeria garrantzitsuen axolagabeen garbizaletasun zehazgabetasun	-en (+) -tasun (0) -keria (-) -ago (+)
3	harrigarrienetako prestigiotsuenetariko garrantzitsuenetako	-eneta(ri)ko (+)

5.4 taula: Hitz batean hainbat hizki agertzeko moduak eta horien eragina hitzaren balentzian.

ager daitezke. Sentimenduen balentzian koherentzia mantentzeko, beharrezkoa da *berritsu* hitzak duen sentimenduen balentziari *-keria* balentzia-aldatzailearen eragina aplikatuta, eragiketa horrek eta sentimenduen lexikoian legokeen *berritsukeria* hitzak sentimendu-balentzia bera edukitzea.

5.1.3. Balentzia-aldatzaileak orientazio semantiko ezberdinetako hitzetan

5.1.1 atalean, balentzia-aldatzaile fonologiko eta morfologikoek balentzia indartu, ahuldu edo balentzian eraginik ez dutela adierazi dugu. Eragin ezberdinetako balentzia-aldatzaile horiek orientazio semantiko ezberdinetako hitzekin batu daitezke eta, ondorioz, hitz horien sentimendu-balentzian gertatzen diren aldaketak askotariakoak dira.

5.5 taulan konbinazio posible guztiak ikus daitezke. Aintzat hartuta hitz batek ++ edo −− orientazio semantikoa duela, balentzia horiek aldaketa ezberdinak jasan ditzakete balentzia-aldatzaileen ezaugarrien arabera. Are positiboago edo negatiboago bilaka daitezke edo, bestela, negatibo edo positibo izaterantz jotzen dute, kasu batzuetan, zeinu aldaketa jasateraino.

Hizkia / hitza	**Hizkia (↑)**	**Hizkia (↓)**	**Hizkia (0)**
Hitza (+)	+ + +	+/−	++
Hitza (−)	− − −	−/+	−−
Hitza (0)			

5.5 taula: Balentzia aldatzen duten hizkiek sentimendu-balentzia ezberdineko hitzetan duten eragina aztertzeko taula.

Ondoren, indartzaileak, ahultzaileak edota eraginik ez duten balentzia-aldatzaileak orientazio semaniko ezberdinetako hitzekin batzean gertatzen diren aldaketak aurkeztuko ditugu.

- Sentimendu-balentzia positibodun hitz bat eta balentzia indartzailea den hizki bat.
  Batura honetan, hitzaren balentzia positiboa are positiboago bilakatzen da. Esaterako, *eroso* hitzak +2 balentzia du eta *erosoen* hitzak balentzia are positiboagoa du.

- Sentimendu-balentzia positibodun hitz bat eta balentzia ahultzailea den hizki bat.
  Baturaren ondorioz, hitzak balentzia positiboa manten dezake, nahiz eta ahulagoa izan, edota hitzaren balentzia positiboa negatibo bilaka daiteke. *Ondu* aditzak +2 balentzia du eta *des-* ezeztapenezko balentzia ahutzailearekin batuz gero, balentzia negatibo bilakatzen da. Emaitza *desondu* −2 da[1]. *Estilistiko* (+3) hitzari *-egi* hizkia gehituz gero, aldiz, *estilistikoegi* hitzaren balentzia jaisten da.

- Balentzia positibodun hitz bat eta balentzia-aldatzailea ez den hizki bat.
  Konbinaketa honetan, hitzak bere balentzia mantentzen du. *Lasai* adjektiboak +2 balentzia du eta *lasaitasun* hitzak ere sentimendu-balentzia bera mantentzen du.

- Hitzak sentimendu-balentzia negatiboa eta balentzia indartzaile bat.
  Horren ondorioa hitzaren sentimendu-balentzia are negatiboago bilakatzea da. *Kalte* izenak −2 balentzia du *-tzar* hizki handigarria erabilita, *kaltetzar* hitzak are balentzia negatiboagoa du.

- Balentzia negatibodun hitz bat eta balentzia ahultzailea den hizki bat.
  Kasu horretan, hitzaren balentziak negatibo izaten jarraitzen du, baina intentsitatea ahulago batez edo, bestela, hitzaren balentzia negatiboa positibo ere bilakatu daiteke. *Sentitegi* lexikoian (Alkorta *et al.*, 2018), *labur* hitzak −1 balentzia du eta *laburregi* hitzak balentzia are negatiboa izango luke. Baina, *nahasi* hitzari, −2 balentziarekin, *-ezin* atzizkia erantsiz gero, *nahastezin* sortzen da eta *-ezin* ezeztapena denez eta ezeztapenak ±4 balentzia duenez, hitza ahuldu eta +2 sentimendu-balentzia izatera igarotzen da.

- Balentzia negatibodun hitz bat eta balentzian eragiten ez duen hizki bat.
  Baturaren ondorioz, hitzaren sentimendu-balentzian ez da aldaketarik

[1]Lan honetan, ezeztapena lantzeko hautatu dugun hurbilpena desplazamendu-ezeztapena da eta, horregatik, ezeztapen-markei nahiz ezeztapenezko hizkiei ±4 balentzia esleitu diegu.

gertatzen. *Baldar* nahiz *baldartasun* hitzek −2 sentimendu-balentzia dute, *-tasun* hizkiak ez duelako balentzian eraginik.

- Balentziarik gabeko hitz bat eta balentzia indartzailea den hizki bat. Hitzak balentziari ez duenez, balentzia-aldatzaileak ez du eraginik. Esaterako, *kolore* eta *mendi* hitzak neutralak dira *Sentitegi* sentimenduen lexikoiaren (Alkorta *et al.*, 2018) arabera, hau da, ez dute balentziarik, eta *koloretsu* eta *menditzar* hitzek ere ez dute.

- Balentziarik gabeko hitz bat eta balentzia ahultzailea den hizki bat. Kasu honetan ere, hitzak sentimendu-balentziarik ez duenez, ez dago balentziaren aldaketarik. Adibidez, *mendi* hitza eta *-txo* balentzia ahultzailea den hizkia batzean (*menditxo*), hitzaren sentimendu-balentzian ez da aldaketarik gertatzen.

- Balentziarik gabeko hitz bat eta balentzia-aldatzailea ez den hizki bat. Konbinazio honetan ere, hitzaren sentimendu-balentzian ez da aldaketarik gertatzen. Horren adibide den *koloretasun* hitzak (*kolore* + *-tasun*) ez du sentimendu-balentziarik eta balentzia-aldaketarik ere jasaten.

5.1.4. Maila fonologikoa eta morfologikoaren garrantzia euskaran

Polanyi eta Zaenenek (2006) ez dute balentzia-aldatzaile morfologiko eta fonologikoen inguruko aipamenik egiten testuinguruko balentzia-aldatzaileen azalpenean.

Baina, lan horrek balentzia-aldatzaileez ari denean, oro har, ingelesa hartzen du ikergaitzat. Horregatik, bertan zerrendatzen diren balentzia-aldatzaile mota guztiak beste hizkuntzetan ezin daitezke baliagarriak izan edota beste hizkuntzetan ingelesean ez dauden balentzia-aldatzaileak egon litezke.

Gure ustez, hori gertatzen da fonologian eta, batez ere, morfologian egon daitezkeen balentzia-aldatzaileekin. Euskara eta ingelesa morfologia tipologiaren aldetik ezberdinak dira, Ingelesa hizkuntza analitikoa da (Szmrecsanyi,

2012), hots, esaldietan hitzen arteko loturak hitz laguntzaileen bidez (preposizioak eta partikulak, esaterako) edo hitz hurrenkeraren bidez egiten dira flexioarekin egin beharrean. Euskara, aldiz, hizkuntza sintetiko eranskaria da (Euskaltzaindia, 1985). Morfologiak garrantzi handia du eta eranskaritasuna erabiltzen du morfemak erabiltzeko. Morfemok hitz erroaren esanahian aldaketak sor ditzakete. Har ditzagun bi adibide:

(47) Hori ikusi zuenean, irribartxoa atera zitzaion aurpegian.

(48) When she saw that, a giggle came on his face.

Bi adibide horietan, euskarazko eta ingelesezko adibide bana dugu eta esanahi bera dute, baina esanahi bera adierazteko moduak ezberdinak dira. (47) adibidean, irribarrea txikia dela adierazteko, *-txo* hizkia erabiltzen da. Ingelesezko (48) adibidean, ordea, beste estrategia bat erabiltzen da eta irribarre txikia adierazteko *giggle* hitza erabiltzen da. Arrazoi horrengatik eta euskaran morfologia oso aberatsa delako, morfologian egon daitezkeen balentzia-aldatzaileak aztertzea erabaki dugu.

Bestalde, fonologiako balentzia-aldatzaileak aztertzearen arrazoia beste bat da. Balentzia-aldatzaileen azterketari ikuspegi orokor bat eman nahi diogu eta, horregatik, fonologia ere lanean sartu nahi dugu. Harluxet hiztegiak (Fundazioa, 1995) dionez, bustidura fonema ez-sabaikari bat testuinguru jakin batean sabaikari bilakatzean datzan fenomenoa. Bustidura adierazkorra bereiztu egin behar da asimilazio bustiduratik. Oñederrak (1990) horrela egiten du bi bustiduren arteko bereizkuntza:

> Euskarazko fonologiaz ari garela, bustikuntza bi motatakoa izan daitekeela izan behar da kontuan: katea fonikoan ingurune jakin batean asimilazioz sortua edo, labur adierazteagatik, bustidura adierazgarria esaten zaiona. (Oñederra, 1990, 13 orr.)

Gaineratzen duenez, bi bustidurak hots klase berari jartzen zaizkio eta aldaketaren ondorioa ere bietan berdina da fonetikan gertatzen den aldaketa berdina delako. Berdintasunak aipatzeaz gain, desberdintasunak zertan diren ere aipatzen du:

> Bata ala bestea gertatzeko baldintzak dira aldatzen direnak, asimilazioa katea fonikoko inguruneak sortutako bilakabide sintagmatikoa den bitartean, bustidura adierazgarria hizkuntzaren eremu semantikoari (*lato sensu*) bait dagokio. (Oñederra, 1990, 14 orr.)

Hots, asimilazio bustiduran, asimilazioa gertatzearen baldintza fonetikoa da eta bustidura adierazkorrean, semantikoa. Beste modu batean esanda, hitz batean aldaketa semantikoa gauzatzeko egiten da bustidura adierazkorra eta nahita eragindako aldaketa fonetikoa da. Bustidura adierazkorraz txikitasuna, maitasuna edo goxotasuna adierazi nahi da.

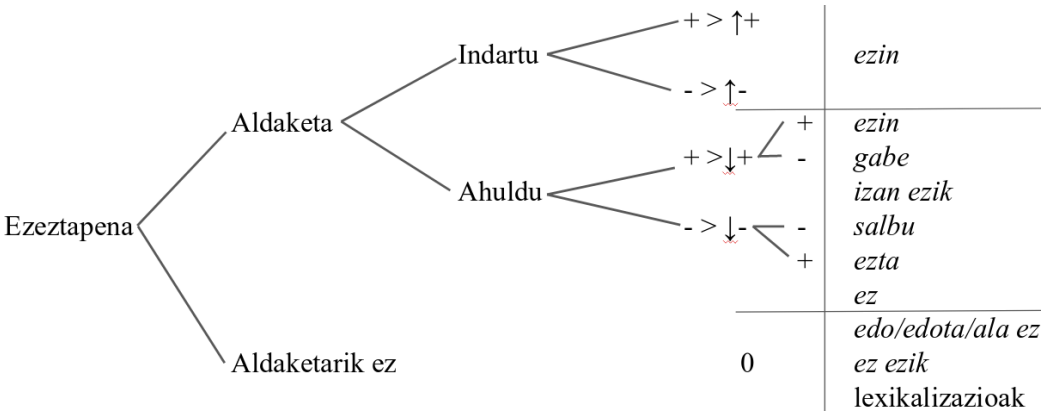
Bustidura adierazkorrak hitzen aldaketa semantikoa eragiten duenez, sentimenduen analisiaren ikuspegitik, balentzia-aldatzaile bat da eta horretan datza fenomeno fonologiko horren interesak.

5.2 Sintaxi maila

Sintaxi maileko 5.2.1 azpiatalean, ezeztapenezko balentzia-aldatzaileak aurkeztuko ditugu. 5.2.2 azpiatalean, trukaketa- eta desplazamendu-ezeztapenaren arteko ezberdintasunak azalduko ditugu, eta desplazamendu-ezeztapena metodologian zergatik erabili dugun ere adieraziko dugu. Azkenik, 5.2.3 azpiatalean, ezeztapen-markak eta beren irismena identifikatzeko erregelak *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) nola sortu eta ondoren, nola ebaluatu ditugun azalduko dugu.

5.2.1. Ezeztapenezko balentzia-aldatzaileak

5.1 irudian, ezeztapenezko balentzia-aldatzileak sentimendu-balentzian nola eragiten dute ikus daiteke.



5.1 irudia: Ezeztapen-marken eragin ezberdinak sentimendu-balentzian.

Ezin sentimendu-balentzia indartu dezakeen ezeztapen-marka bakarra da. Bestalde, sentimendu-balentzia ahultzen duten ezeztapen-markak gehiago dira: *ezin*, *gabe*, *izan ezik*, *salbu*, *ezta* eta *ez*. Azkenik, badaude zenbait egitura non ezeztapen-marka agertzen den eta ez duen sentimendu-balentzian eragiten. Egitura horiek ezeztapena juntagailuarekin, kontrastezko ezeztapena eta lexikalizatutako egiturak dira. Jarraian, 5.1 irudian azaldutako sailkapena xehetuago emango dugu.

Ezeztapen-markaren instantzia guztietatik (359), zazpi kasu antzeman ditugu ezeztapen-markak irismenaren sentimendu-balentzia indartu duena. Balentzia-aldaketa hori *ezin* ezeztapen-markak adjektiboari edo aditzondoari lotuta dagoenean bakarrik gertatzen da (corpusean agertu diren kasu guztien % 1,96).

(49) Dena nahasten da maisulan **ezin** ederragoa$_{+4}$ osatzeko. (MUS21).[2]

(49) adibidean, sentimendu-balentzia indartzen den adibide bat dago. Kasu horretan, ezeztapen-markak (*ezin*) konparaziozko atzizkia duen adjektiboari eragin eta bere sentimendu-balentzia indartu du. Hots, positiboa zen sentimendu-balentzia are positiboagoa bihurtzen du. Izan ere, esaldi horretan *ederragorik* ez dagoela esatea *ederrena* dela esatea da; beraz, sentimendu-balentziaren indartze bat dago.

Kasu gehienetan, aztertu dugun corpuseko instantziek sentimendu-balentzia ahultzen dute. Hainbat ezeztapen-marka motek ahultzen dute sentimendu-balentzia: i) *ez*, ii) *gabe*, iii) *ezin*, iv) *izan ezik*, v) *salbu* eta azkenik, vi) *ezta*. Kasuen % 89,98etan (323 instantzietan) gertatu da sentimendu-balentziaren ahultzea.

(50) Horrek **ez** die eragotzi$_{-2}$ ordea, 57 milioi euro ematea San Mames klub pribatuari! (POL30)

(51) Bruch-en kontzertua, munduko interesgarriena izan **gabe**, lan erakargarria da, oso, erraza entzuteko, eta bakarlari honen bertsioak dotoreziaren *plus*a izan zuen dudarik gabe. (MUS22)

(52) Irabazi$_{+2}$ **ezin**ik jarraitzen du Eibarrek, baina oso puntu ona eskuratu du Getaferen zelaian. (KIR17)

(53) "Ez baitute ezertarako balio izan, egun bateko soldata galtzeko$_{-2}$ **izan ezik**". (POL17)

[2]Esaldian dauden anotazio ezberdinen esanahiak hauek dira: **beltz koloreak** ezeztapen-marka adierazten du, azpimarkatutakoak ezeztapenaren irismen-eremua seinalatzen du. Berez, ezeztapen-markek ez lukete egon beharko azpimarkatuta, baina hori ez *ezin* ezeztapen-markaren kasuan. Izan ere, ezeztapen-marka horrek bi esanahi ditu: i) ezeztapena bera eta ii) posibilitatea. Horregatik, azpimarkatuta dago.

(54) E.T. eta Indiana Jones filmak **salbu**, noski. (ZIN18)

(55) **Ezta** Euskal Herria, Espainiako Estatua, Portugal edo Italia moduko beste herrialdeei ere. (POL08)

(50) adibidean, ezeztapen-markaren eragin-eremua esaldi osoa da eta *eragotzi* aditzaren −2 sentimendu-balentzia ahultzen du. Beraz, zeinu positiboa (+4) hartzen du ezeztapen-markaren eraginez. (51) adibidean, *gabe* ezeztapen-markak sintagma bati eragiten dio (*munduko interesgarriena izan*). (52) adibidean, *ezin* ezeztapen-markak bere ezkerretara dagoen *irabazi* aditzaren +2 sentimendu-balentzia ahultzen du. (53) adibidean, ezeztapen-marka *ezik* da eta *izan* aditzarekin batera *egun bateko soldata galtzeko* sintagma ezeztatzen du. (54) adibidean, *salbu* ezeztapen-markak izen-sintagma bati (*E.T. eta Indiana Jones filmak*) eragiten dio. Azkenik, (55) adibidean, *ezta* ezeztapen-markak sintagma osoari eragiten dio.

Azkenik, aurretik aipaturiko ezeztapen-markek sentimendu-balentzian eraginik ez duten kasuak ere badaude. Kasu hori sentimendu-balentzia indartzea baino ohikoagoa da, corpusean instantzien % 8,08an, 29 instantzietan, gertatu baita. Ezeztapen-markak berak ez du sentimendu-balentzian eraginik horrelako kasuetan: i) ezeztapen-marka juntagailuarekin agertzen denean, ii) ezeztapen-marka kontrastezko ezeztapenaren parte denean, eta iii) ezeztapen-marka egitura lexikalizatu batean agertzen denean (hots, egitura horiek beren berezko esanahia dutenean eta horietako batzuk hiztegietako sarrerak direnean).

(56) Cate Le Bon bai edo ez, hemen ez dago erdibidekorik. (MUS33)

(57) Ikuspuntu politikotik$_{-1}$ ez ezik, ekonomikotik$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(58) Sei puntu baino ez dituela, hamaseigarren postuan da Reala sailkapenean. (KIR27)

(56) adibidean, idazleak aurkako bi ideia adierazten ditu (bai eta ez). Horrenbestez, ez da balentzia-aldaketarik sortzen. (57) adibidean, kontrastezko

ezeztapen bat dago eta informazioa gehitzeko funtzioa betetzen du (Silvennoinen, 2017). Hots, ezeztapen-marka emendiozko juntadura batean agertzen. Horrelako egituretan, *ez ezik* ezeztapen-markak bere aurretik dagoen elementua (*Ikuspuntu politikotik* adibidearen kasuan) ezeztatzen du. Baina, sentimendu-balentziaren ikuspegitik, ez da balentziarik ezeztatu eta ondorioz, ahuldu. Esaldian, ezeztapen-marka daraman lehen zatiak bere atzetik informazioa emendatzen edo gehitzen du. Azkenik, (58) egitura lexikalizatuaren adibidea da.

5.2.2. Trukaketa- eta desplazamendu-ezeztapena

Sentimenduen analisian eta zehazki, lexikoian eta hizkuntza-ezagutzan oinarritzen den ikuspegian, bi hurbilpen daude ezeztapena lantzeko. Bi hurbilpen horiek trukaketa-ezeztapena (*switch negation*, ingelesez) (Sauri, 2008) eta desplazamendu-ezeztapena (*shift negation*, ingelesez) dira. Azter ditzagun bi hurbilpenak beheko bi adibideaetan[3].

(59) Edaritegia [**ez** da ona$_{+3}$]$^{-3}$, baina bertako musika ona$_{+3}$ da.

(60) Edaritegia [**ez**$_{-4}$da ona$_{+3}$]$^{-1}$, baina bertako musika ona$_{+3}$ da.

Lehen adibideak (59) trukaketa-ezeztapenari dagokio eta bigarrenak (60), berriz, desplazamendu-ezeztapenari. Lehen adibidean, *ez* ezeztapen-markak ez du balentziarik eta ezeztapen-markaren eragina da bere irismeneko parte den *ona* hitzaren sentimendu-balentzia (+3) alderantzikatzea. Beraz, trukaketa-ezeztapenean, irismeneko sentimendu-balentziari zeinua aldatzen zaio, positibotik negatibora edo alderantziz. Bigarren adibidean, desplazamendu-ezeztapenaren adibidea ageri da. Kasu horretan, *ez* ezeztapen-markak balentzia jakin bat du (−4) eta ezeztapen-markaren eragina da irismeneko hitzen sentimendu-balentziaren baturari (adibidean, −4) sentimendu-balentzia gehitzea. Irismeneko sentimendu-balentzia negatiboa bada, ezeztapen-markaren balentzia +4 izango dela.

Bi hurbilpen horien artean, guk desplazamendu-ezeztapena erabili dugu. Hurbilpen hori sintaxi mailako ezeztapenean eta ezeztapenarekin loturiko

[3](59) eta (60) adibideak (Taboada *et al.*, 2011) lanetik hartu eta itzuli dira.

hizkiekin erabili dugu. Balentzia-aldatzaile horri ±4 balentzia esleitu diogu eta, ondoren, balentzia duten hitzetan zer aldaketa eragiten duten aztertu dugu. Taboada *et al.*ek (2011) aipatzen duenez, azken hurbilpen hori hobeto dabil eta ez dago kontraesanik trukaketa-ezeztapenean bezala. Azter dezagun trukaketa-ezeztapenak duen akatsa.

(61) Bikaina$_{+5}$ → Ez bikaina$_{-5}$. Anker$_{-5}$.

(62) Ez bikaina$_{-5}$. Ez ona$_{-3}$.

(61) adibidean, trukaketa-ezeztapena erabiliz, *bikaina* adjektiboa +5 sentimendu-balentzia izatetik, ezeztapen-markaren eraginez, −5 sentimendu-balentzia izatea igarotzen da. Baina, hemen lehen kontraesana sortzen da; izan ere, intuitiboki *ez bikaina* eta *anker* elkarrengandik urrun daude, baina hurbilpen horren arabera, biek sentimendu-balentzia bera dute.

Bigarren kontraesana (62) adibidean beha daiteke. Bertan *bikaina* (+5) eta *ona* (+3) ezeztatuta daude. Ezeztapenaren eraginez, *ez bikaina* egiturak −5 sentimendu-balentzia du eta *ez ona*k aldiz −3. Baina biak alderatzen baditugu, *ez bikaina ez ona* baino positiboagoa dela ohartuko gara. Hala ere, trukaketa-ezeztapena jarraituz, *ez bikaina* egiturak sentimendu-balentzia negatiboagoa du *ez ona*k baino. Horrengatik guztiarengatik, mugimendu ezeztapena hurbilpena hautatu dugu eta ezeztapen-markek ±4 balentzia izango dute.

5.2.3. Ezeztapen-markak eta beren irismena identifikatzeko erregelak

Metodologian, 3.2.2 atalean, ezeztapen-markek aditzen eta sintagmen sentimendu-balentzia nola eragiten duten identifikatu ondoren; ezeztapen-markak eta beren inguruneko hitzen gramatika-kategoriekin 5.6 taulako erregelak osatu ditugu. Erregela horiek *Murriztapen Gramatika*n, MGn, (Karlsson *et al.*, 1995) erregelak sortzeko baliatu dira.

Murriztapen Gramatika (Karlsson *et al.*, 1995) testuingurua iruditu zaigu egokiena ebaluazioa egiteko, gramatikaren ezaugarriak ondo uztartzen baitira

Zenb.	Ezeztapen-marka	Erregelen egitura
1	ezin	PM *ezin* + [adjektiboa/adberbioa] (+ atzizki konp.) PM
2	ez	PM [(IS +) aditza +] *ez* PM PM [(IS +)] *ez* [+ ad. lag. (+ IS) + aditza (+ IS)] PM PM [IS +] *ez* PM
3	ez	PM *ez* [+ IS +] *ez* [+ IS] (...) PM
4	gabe	PM [IS/AS/sintagma +] *gabe* PM
5	ezin	PM [(IS) + aditza + *ezin*] PM PM [*ezin* (+ ad. lag.) (+ IS) + aditza (+ IS)] PM
6	izan ezik	PM [IS/sintagma] + *izan ezik*
7	salbu	PM [IS] + *salbu* PM
8	ezta	PM *ezta* + [IS/sintagma] PM
9	ez	PM [aditza/*bai*] *edo/ala/edota ez* PM
10	ez	PM [IS] *ez ezik* PM
11	ez, gabe, ezin, ezean	Egitura lexikalizatuak

5.6 taula: Sentimendu-balentzian eragin ezberdinak dituzten ezeztapenak identifikatzeko proposatu diren erregeletako batzuk.

ebaluatu nahi ditugun alderdiekin. Hiru ezaugarri nagusi eta abantaila ditu gramatika horrek:

- *Murriztapen Gramatika* (Karlsson *et al.*, 1995) hizkuntzarekiko independentea da. *Murriztapen Gramatika* analisi morfologikoan oinarritzen da, edozein testu analizatzeko helburuarekin. Gainera, egoera finituko mekanismoetan oinarritzen da, eta hori hizkuntzalaritzaren ikuspegitik gramatikaren formalismoa oso intuitiboa eta erabilerraza da.

- Desanbiguazio-erregelak. Gramatikako erregelek analisi morfologikoaren eta funtzio sintaktikoen desanbiguazioa gauzatzen dute. Erregelek testuinguru jakin batean zuzenak edo egokiak ez diren interpretazioak kentzen dituzte.

- Islapen-erregelak. Erregela horiek etiketatze morfologikoan eta sintaktikoan egon daitezkeen hutsuneak osatzen dituzte, testuingurua kontuan hartuz. Guk ezeztapen-markaren testuingurua, hau da, irismena identifikatu nahi dugu eta, horregatik, islapen-erregelak egokiak dira guretzat.

5.2.3.1. Murriztapen Gramatikako erregelak sortzeko eta ebaluatzeko prozedura

Erregelen egituraketa

Jarraian urratsez urrats, 5.6 taulako erregelak ingurunera nola egokitu ditugun esplikatuko dugu. *Murriztapen Gramatika* (Karlsson *et al.*, 1995) testuinguruan sartu behar ditugun elementuak hauek dira[4]:

- Puntuazio-markak, ezeztapen-markak eta egitura lexikalizatuetako hitzak elementuen zerrendetan (LIST) sartzea erabaki dugu. Ia erregela denetan agertuko diren hitzak dira eta hitz jakinak dira, hots, beti forma bera dute.

  (63) LIST EZ = ”ez”;

  (64) LIST BESTERIK = ”beste”;

  (63) eta (64) adibideetan, berriz, *ez* ezeztapen-marka eta egitura lexikalizatuetako *besterik* hitza, hurrenez hurren, elementuen zerrendetako elementu bakarrak dira, modu horretan, hitz horiek ezeztapen-marken erregela ezberdinekin konbinatzeko aukera edukitzeko.

- Irismeneko hitzen gramatika-kategoriak eta haien posizioa islapen-erregeletan sartu ditugu. Irismenean, aurreko kasuetan ez bezala, ez dakigu zein hitz ager daitezkeen bertan. Horregatik, ezin ditugu irismeneko hitzekin elementuen zerrendak osatu. Dakigun bakarra irismenean ager daitezkeen hitzen gramatika-kategoria eta ezeztapen-marketatik haien distantzia dira, 5.6 taularen bidez lortu ditugunak. Horrenbestez, informazio hori islapen erregeletan adierazi dugu:

  (65) MAP (!salbuBUK) TARGET (ADB) IF (0C SALBU);
  MAP (!salbuHAS1) TARGET (IZE) IF (1C SALBU);

  (65) adibidean, bestalde, bi erregela ageri dira *larunbata salbu* bezalako egiturak identifikatzeko. *!salbuBUK* etiketa duen erregelan, 0 posizioan SALBU elementuen zerrendak egon behar duela adierazten du

[4]Elementu horiek ikusgai daude A eranskinean.

eta, gainera, elementu zerrendakoak adberbioa izan behar du. Bestalde, *!salbuHAS1* etiketan duen erregelan, *salbu* ezeztapen-markaren aurretik izen batek egon behar duela adierazten da. Kasu horretan, hitza edozein izan daiteke, baina badakigu izenaren gramatika-kategoriakoak izan behar duela, 5.6 taularen bidez. Aipaturiko etiketa horiek izango dira sortu ditugun erregelak corpusetik pasatzean utziko duten aztarna.

Erregelak ebaluatzeko prozedura

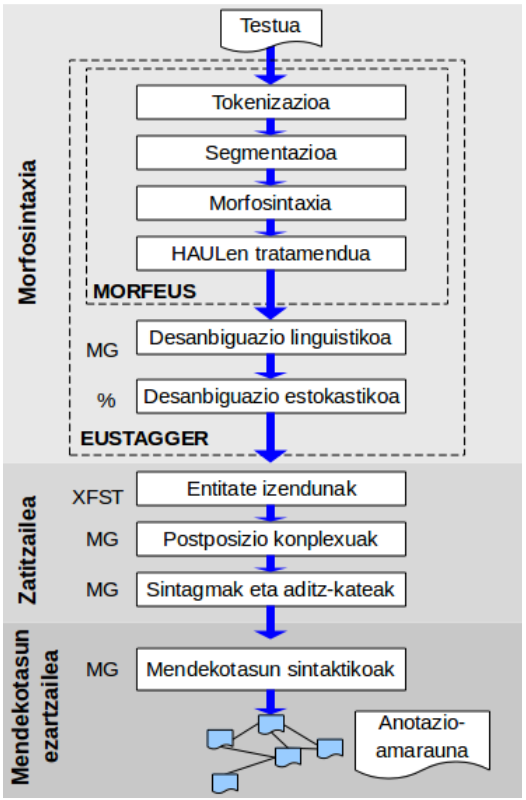
Erregelen ebaluazioa etiketatzaile batek gauzatu du, baina, aldez aurretik, etiketatzaile horren eta beste etiketatzaile baten adostasun maila neurtu dugu, ebaluazioaren kalitatea ebaluatu ahal izateko. Bi etiketatzailek Euskarazko Iritzi Corpusaren (Alkorta *et al.*, 2016) % 10a (2.706 hitz) etiketatu dute. Bi etiketatzaileen arteko adostasuna zein hitz etiketatzeari dagokionez, kappa 0,91koa da. Bi etiketatzaileak etiketatutako beharreko hitzak nola etiketaturi (ETIK_ONDO, ETIK_FALTA eta ETIK_GAIZKI) dagokionez, kappa 0,69koa da, zeina Landis eta Kochen (1977) arabera *sendoa* den.

Bi etiketatzaileen arteko adostasuna neurtu ostean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 144 iritzi-testu erabili ditugu *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelak ebaluatzeko. 144 iritzi-testuko corpusari Ixa taldearen analisi-katea (Aduriz *et al.*, 2004) pasa diogu. Analisi-katearen arkitektura 5.2 irudian ikus daiteke.

Ixarko analisi-katearen oinarrian EDBL-LBDBL baliabide lexikal konputazionala (Aldezabal *et al.*, 2006) dago eta analisi-kateak hiru atal ditu: morfosintaxia, zatitzailea eta mendekotasun ezartzailea. Corpusari analisi-katea pasa ondoren, emaitza hitz bakoitzeko analisi bakarra izan da 5.3 irudian agertzen den moduan.

Bertan, analisi-katearen arabera, *azaroaren* hitza izen (IZE) arrunta (ARRU) da eta bere kasua genitiboa (GEN>) da.

Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 144 iritzi-testuei analisi-katea pasa ostean, *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelak aplikatu ditugu corpusean eta erregela horiek corpuseko analisietan beren etiketa utzi dute. 3.2.2 atalean aipatu dugun moduan, etiketa erregelak ebaluatzeko erabili ditugu.



5.2 irudia: Ixa taldearen analisi-katea (Oronoz, 2009).

```
"<azaroaren>"
        <Correct!> "azaro" IZE ARR GEN NUMS MUGM ZERO w67,L-A-IZE-ARR-26,
            lsfi6 @IZLG>
```

5.3 irudia: Ixa taldeko analisi-katearen emaitza.

Murriztapen Gramatika (Karlsson *et al.*, 1995) esleitutako etiketak zuzenak edo okerrak diren, edo etiketaren bat falta den ebaluatzeko hiru kasuistika bereiztu ditugu[5]:

- ETIK_ONDO. Sorturiko erregelak dagokion hitzari etiketa jarri badio ontzat joko dugu. Kasu horretan, beraz, erregelak ondo identifikatu du

[5]Erregelak ebaluatzeko etiketak corpusaren zati batean aplikatuta 3.9 irudian daude ikusgai.

ezeztapenarekin loturiko elementuren bat (ezeztapen-marka, irismena edo egitura lexikalizatua) corpusean.

- ETIK_FALTA. Sorturiko erregelak dagokion hitzari etiketa ez badio jartzen okertzat joko dugu. Beste kasu batzuetan, gerta liteke erregelak ezeztapenarekin loturiko elementuren bat ez identifikatzea corpusean eta kasu horretan erabiliko dugu etiketa hori.
- ETIK_GAIZKI. Sorturiko erregelek ez dagokion hitzari etiketa jartzen badio okertzat joko dugu. Azken kasuan, gerta liteke erregelek ezeztapenarekin loturarik ez duen hitz bat ezeztapenarekin lotura duen elementutzat jotzea eta, kasu horretan ere, kasuistika hori okertzat joko dugu.

5.2.3.2. Erregelen ebaluazioaren emaitzak

Lehenik eta behin, emaitzak ezeztapen-markek sentimendua balentzian eragiten dutena (indartu, ahuldu edo eraginik ez) kontuan hartuta emango ditugu. Ondoren, emaitzak ezeztapenari lotutako elementuetan (ezeztapen-markak, irismena eta egitura lexikalizatuak) arabera emango ditugu.

5.7 taulako datuen arabera, erregelek ezeztapen-markekin lotutako osagaien hitzak ondo identifikatzen ditu, F1 puntuazioa 0,86 baita. Zehatzago esanda, erregelek ezeztapenarekin lotutako hitzak identifikatzerakoan 0,91ko doitasuna eta 0,80ko estaldura izan dute.

Eragina	**Doitasuna**	**Estaldura**	**F1**	**Hitzak (instantziak) ✓/✗/✢**
Indartu	1,00	1,00	1,00	4 (2): 4/0/0
Ahuldu	0,93	0,80	0,86	1.050 (192): 784/63/203
Eraginik ez	0,97	1,00	0,98	36 (19): 35/1/0
Guztira	0,93	0,80	0,86	1.090: 823/64/203

5.7 taula: Emaitzak ezeztapen-markek sentimendu-balentzian eragiten dutenaren ikuspegitik.

Ezberdintasun batzuk daude F1 puntuazioan, ezeztapen-markek balentzian eragiten dutenaren arabera. Esaterako, ezeztapen-markak indartzen duenean edota eraginik ez duenean, F1 puntuazioa altua da. Baina, eurei loturiko

hitz eta instantzia[6] kopurua baxua da. Ezeztapenak balentzia ahultzearen kasuan, aldiz, F1 puntuazioa 0,86 da, baxuagoa. Indartzearen kasuan, ezeztapenari lotutako hitz guztiak (4) ondo identifikatu dira. Ahultzean, 784 ondo identifikatu dira, 63 gaizki eta 203 ez dira identifikatu. Eraginik ez dagoen kasuetan, 35 hitz ondo identifikatu dira eta bakarra gaizki. Ezeztapena ahultze eraginaren kasuan, F1 puntuazioa 0,86 da. Emaitzak ezeztapenari loturiko elementu-motan oinarrituz ageri da 5.8 taulan.

	Doitasuna	**Estaldura**	**F1**	**Hitzak (instantziak): ✓/×/✢**
Ezeztapen-markak	1,00	0,96	0,98	195 (195): 188/0/7
Egitura lexikalizatuak	0,96	1,00	0,98	28 (16): 27/1/0
Irismena	0,91	0,75	0,82	867 (195): 601/63/196
Guztira	0,93	0,80	0,86	1.090: 823/64/203

5.8 taula: Emaitzak ezeztapenari loturiko elementu-motaren ikuspegitik.

Bertan beha daitekeenez, ezeztapen-markak eta egitura lexikalizatuak oso ondo identifikatu dira, F1 puntuazioa 0,98 baita bietan. Zazpi ezeztapen-marka ez dira identifikatu eta corpuseko hitz bat egitura lexikalizatutzat hartu da, nahiz eta ez den horrela. Irismenean, ordea, F1 puntuazioa baxuagoa da, 0,82koa, hain zuzen ere. Irismenaren parte diren 196 hitz ez dira identifikatu eta irismenaren parte ez diren 64 hitz irismeneko zatitzat hartu dira. Horretaz gain, 867 hitzetatik 196 ez dira irismeneko parte bezala jo, nahiz eta hala diren.

Errore-analisiak erregelen osaketari buruzko informazioa eman digu. Irismenari dagokionez, erregelek ez dituzte 196 hitz identifikatu (identifikatu gabeko zazpi ezeztapen-markak kontuan hartu gabe). Hitz horiek zer dela eta identifikatu gabe gelditu diren aztertu ondoren, kasu gehienetan, puntuaziomarkaren murriztapenarekin lotura dutela ohartu gara. Hau da, puntuaziomarka murriztapena irismena esaldi barrura mugatzeko sortua izan da, baina

[6] 5.7 taulan, eskuineko zutabean, hitzei eta instantziei loturiko datuak jarri ditugu. Lehenik eta behin, hitz-kopurua agertzen da, ondoren, parentesi artean, instantzia-kopurua eta ondoren, ezeztapenari loturiko hitzak ondo, gaizki edo ez diren asmatu. Ikusten den moduan, hitz eta instantzia-kopuruak ezberdinak dira. Ezeztapen-marken kasuan, hitz bat instantzia bat da. Irismenaren kasuan, instantziako hitz-kopurua ezberdina da. Guztira 192 irismen daude corpusean eta horiek 1.050 hitz dituzte. Azkenik, egitura lexikalizatuetan, hitz batetik hiru hitzerainokoa izan daiteke instantzia bat, betiere egitura lexikalizatu horren ezaugarrien arabera.

horrek zeharkako kalte batzuk sortzen ditu. Esaldiren batean zerrendaketa bat dagoenean, eta zerrendaketa hortako elementuak koma bidez bereizita daudenean, puntuazio-markaren murriztapena ez da ondo ibiltzen, zerrendaketako elementu bakarra identifikatzen baitu, esaldia hor bukatzen dela interpretatzen duelako.

5.3 Diskurtso maila

5.3.1 azpiatalean, diskurtso mailan aztertu ditugun balentzia-aldatzaileak aurkeztu ditugu. 5.3.2 atalean, Egitura Erretorikoaren Teoria (RST) (Mann eta Thompson, 1988) deskribatuko dugu.

5.3.1. Balentzia-aldatzaileak: nukleartasuna eta unitate zentrala

Balentzia-aldatzaileen azterketa diskurtso mailan hiru ikuspegitatik egin dugu. Alde batetik, nukleartasunak balentzia-aldaketan eraginik duen aztertu dugu (5.3.1.1 atala). Beste aldetik, unitate zentralak balentzia-aldaketan eraginik ere baduen aztertu dugu (5.3.1.2 atala). Azkenik, unitate zentralak Azkenik, erlaziozko diskurtso-egiturak unitate zentraletik duen distantzia bakoitzean zenbateko maiztasunez agertzen diren aztertu dugu (5.3.1.3 atala).

5.3.1.1. Orientazio semantikoa eta nukleartasuna

Balentzia-aldatzaileen eta nuklearitatearen arteko lotura aztertzeko diskurtso-erlazioak aztertu ditugu. Zehazki, zer diskurtso-unitatek (nukleoak edo sateliteak edota lehen testu-zatia edo azken-testu zatia) duen erlazio osoarekin orientazio semantikoaren adostasun gehien neurtu dugu. 5.9 taulan, orientazio semantikoaren adostasunari buruzko emaitzak azter daitezke, nuklearitatea eta osagaien lekua aintzat hartuz[7]. Emaitza horietan, adostasunik handiena +1 balioa da eta adostasunik baxuena 0 balioa.

5.9 taulako emaitzen arabera, kasu gehienetan, nukleo-unitateak erlaziozko diskurtso-egitura osoaren orientazio semantikoarekin adostasun handiagoa erakusten du satelite-unitateak baino. Orientazio semantikoaren adostasunean nukleo-unitateak eta satelite-unitateak duten ezberdintasuna ez da oso handia. 0,73 da nukleo-unitatearen adostasuna eta 0,64, berriz, satelitearena. Hala ere, joera hori ez da gertatzen erlaziozko diskurtso-egitura mota guztietan. Ebaluazioa taldearen[8] kasuan, satelite-unitateak erlaziozko

[7]Letra xehez dauden erlaziozko diskurtso-egiturak multzoak dira. Letra larriz daudenak erlaziozko diskurtso-egitura bakarra dira.

[8]Ebaluazioa taldea EBALUAZIOA eta INTERPRETAZIOA erlaziozko diskurtso-

Adostasuna Xrekin	Instantziak	Nukleoa	Satelitea
Kausa taldea	71	0,83	0,66
Aurkaritza taldea	70	0,70	0,33
BALDINTZA	18	0,61	0,56
Ahalbideratzea taldea	6	0,60	0,5
Ebaluazioa taldea	140	0,69	0,82
Ebidentzia taldea	53	0,83	0,64
KONTRASTEA*	26	0,73	0,31
Guztira	384	0,73	0,65

5.9 taula: Orientazio semantikoaren adostasuna erlaziozko diskurtso-egituren eta nukleoaren/satelitearen artean.

diskurtso-egitura osoaren orientazio semantikoarekin adostasun handiagoa erakusten du satelite-unitateak baino. Kasu horretan, satelite-unitatearen adostasuna 0,82koa da nukleo-unitatearena 0,69 den bitartean. Bestalde, Aurkaritza taldeak[9] du nukleo-unitatearen eta satelite-unitatearen arteko ezberdintasunik handiena. Nukleo-unitateak 0,70ko adostasuna du eta satelite-unitateak, aldiz, 0,33. Beraz, puntuazioan 0,44ko ezberdintasuna dago. Bukatzeko, nukleo-unitateak Kausa taldean[10] eta Ebidentzia taldean[11] erlaziozko diskurtso-egitura osoaren orientazio semantikoaren adostasunik handiena dute, puntuazioa 0,83koa baita bietan.

Erlaziozko diskurtso-egitura osoaren orientazio semantikoaren adostasunik handiena lehen eta azken unitateak duen aztertzen badugu, emaitzak argigarriagoak dira. 5.10 taulak erakusten duen moduan, erlaziozko diskurtso-egituraren azken unitateak erlazio osoaren orientazio semantikoarekin adostasun handiagoa du. Azken unitatearen kasuan, adostasunaren puntuazioa 0,78koa da eta lehen unitatearen kasuan, berriz, 0,58koa. Beren artean dagoen aldea esanguratsua da. Kasu horretan ere, erlaziozko diskurtso-egitura mota guztiek ez dute joera erakusten. Ahalbideratzea[12] taldean, erlazioa-

egiturek osatzen dute.

[9]Aurkaritza taldea KONTZETSIOA eta ANTITESIA erlaziozko diskurtso-egiturek osatzen dute.

[10]Kausa taldea KAUSA, ONDORIOA eta HELBURUA erlaziozko diskurtso-egiturek osatzen dute.

[11]Ebidentzia taldea EBIDENTZIA eta JUSTIFIKAZIOA erlaziozko diskurtso-egiturek osatzen dute.

[12]Ahalbideratzea taldea AHALBIDERATU eta MOTIBAZIOA erlaziozko diskurtso-

Adostasuna Xrekin	Instantziak	Lehen unitatea	Azken unitatea
Kausa taldea	71	0,66	0,83
Aurkaritza taldea	70	0,30	0,73
BALDINTZA	18	0,28	0,89
Ahalbideratzea taldea	6	0,67	0,50
Ebaluazioa taldea	140	0,68	0,83
Ebidentzia taldea	53	0,79	0,68
KONTRASTEA*	26	0,35	0,69
Guztira	384	0,58	0,78

5.10 taula: Orientazio semantikoaren adostasuna erlaziozko diskurtso-egituren eta lehen osagaiaren/azken osagaiaren artean.

ren lehen unitateak azkenak baino orientazio semantikoaren adostasun handiagoa dute. Ahalbideratzearen kasuan, lehen unitatearen adostasun puntuazioa 0,67koa da eta azken unitatearena 0,50ekoa. Bestalde, Ebidentzia taldean puntuazioak 0,79 eta 0,68 dira, hurrenez hurren. Lehen unitatearen eta azken unitatearen artean, puntuazioaren ezberdintasunik handiena BALDINTZA erlazioan gertatu da. Orobat, lehen unitarearen puntuazioa 0,28koa da eta azken unitatearena 0,89. Hots, 0,71ko puntuazioaren tartea dago. Azkenik, BALDINTZA erlazioko, Ebaluazioa taldeko eta Kausa taldeko azken unitateek erlaziozko diskurtso-egitura osoaren orientazio semantikoarekin adostasun puntuaziorik handiena dute, puntuazioa 0,89koa baita BALDINTZAren kasuan eta 0,83koa Ebaluazioarenean eta Kausarenean.

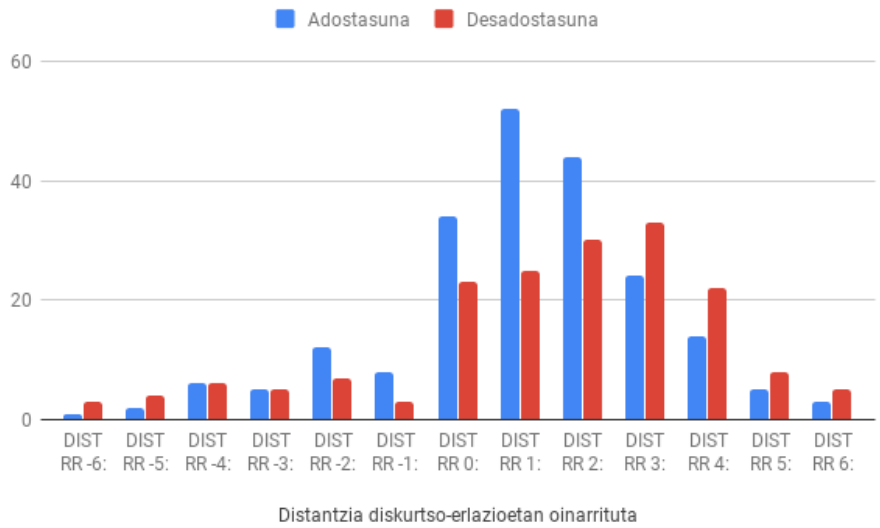
Laburbilduz, erlaziozko diskurtso-egituraren orientazio semantikoa baldintzatzen dute bai nukleartasunak, bai lehen edo azken unitatea izateak ere.

5.3.1.2. Orientazio semantikoa eta unitate zentrala

5.4 irudian, erlaziozko diskurtso-egitura guztiek beren testu osoarekin duten orientazio semantikoaren adostasun eta desadostasun instantziak agertzen dira. Unitate zentrala (DIST RR 0) da eta erlaziozko diskurtso-egiturak bere aurretik eta atzetik daude.

Erlaziozko diskurtso-egitura unitate zentraletik zenbat eta gertuago egon, erlaziozko diskurtso-egituren eta beren iritzi-testuen arteko orientazio semanti-

egiturek osatzen dute.



5.4 irudia: Testuek eta erlaziozko diskurtso-egiturek beren artean dituzten orientazio semantikoaren adostasun eta desadostasunak.

koaren adostasuna handiagoa da. −2, −1, 0, +1 eta +2 distantzietan orientazio semantikoaren adostasunaren instantzia gehiago daude desadostasunak[13] baino. Beste distantzietan, ordea, orientazio semantikoaren desadostasunaren instantziak gehiago dira edo bestela adostasun eta desadostasunen artean ez dago ezberdintasun nabarmenik.

Emaitzak erlaziozko diskurtso-egitura moten arabera aztertzen badira, zenbait ezberdintasun antzeman daitezke. Kausa taldearen kasuan, orientazioaren adostasunaren instantziak gehiago dira 0 eta +1 distantzietan bakarrik. Aurkaritza taldeak, aldiz, ez du erlaziozko diskurtso-egiturek distantzietan erakusten duten joera jarraitzen. Izan ere, Aurkaritza taldean, adostasun instantziak bi gune bereizitan dira nagusi. Hau da, adostasuna gune batean nagusi da (−2 distantzia); ondoren, desadostasun instantziak gehiago dira (0 distantzia) eta, azkenik, adostasun instantziak berriz nagusitzen dira (+1 eta +2 distantziak).

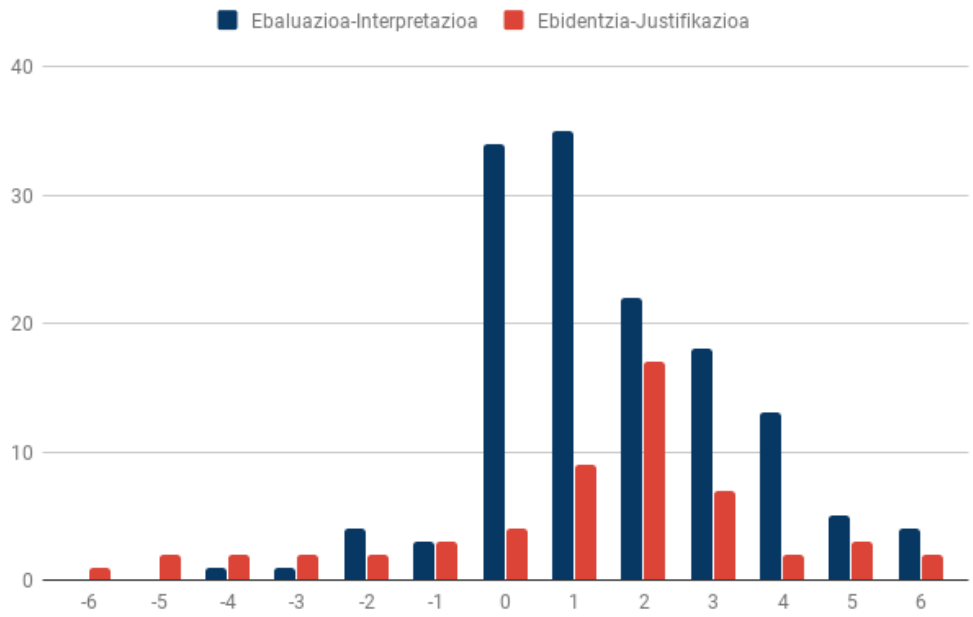
[13]Desadostasuna izendapenez, erlaziozko diskurtso-egiturak eta iritzi-testuak aurkako orientazio semantikoa duten instantziak adierazi nahi ditugu.

Laburbilduz, unitate zentralak erlaziozko diskurtso-egiturak eta iritzi-testuak duten orientazio semantikoaren adostasunean eragiten du, izan ere, erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, erlazioak duen orientazio semantikoak iritzi-testuak duenarekin gehiagotan bat egiten du. Gure ustez, unitate zentraletik gertuen dauden erlaziozko diskurtso-egituren gaiak iritzi-testuaren gai nagusitik hurbilago daudelako dago bien arteko orientazio semantikoaren adostasun handiagoa. Unitate zentraletik urrun dauden erlaziozko diskurtso-egituren, ordea, gaia iritzi-testuarengandik ezberdinak dira eta, ondorioz, orientazio semantikoan adostasuna ez da hainbestekoa.

5.3.1.3. Erlaziozko diskurtso-egituren agerpena distantzia bakoitzean

Iritzi-testuetan, erlaziozko diskurtso-egiturak distantzia bakoitzean zenbateko maiztasunez agertzen diren ere neurtu dugu. Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeen emaitzak aurkeztuko ditugu. Distantzia unitate zentraletik neurtu dugu. Distantzia negatiboa da unitate zentralaren ezkerretara eta positiboa unitate zentralaren eskuinetara. 5.5 irudian, Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeek instantzia gehien zer distantzietan duten beha daiteke.

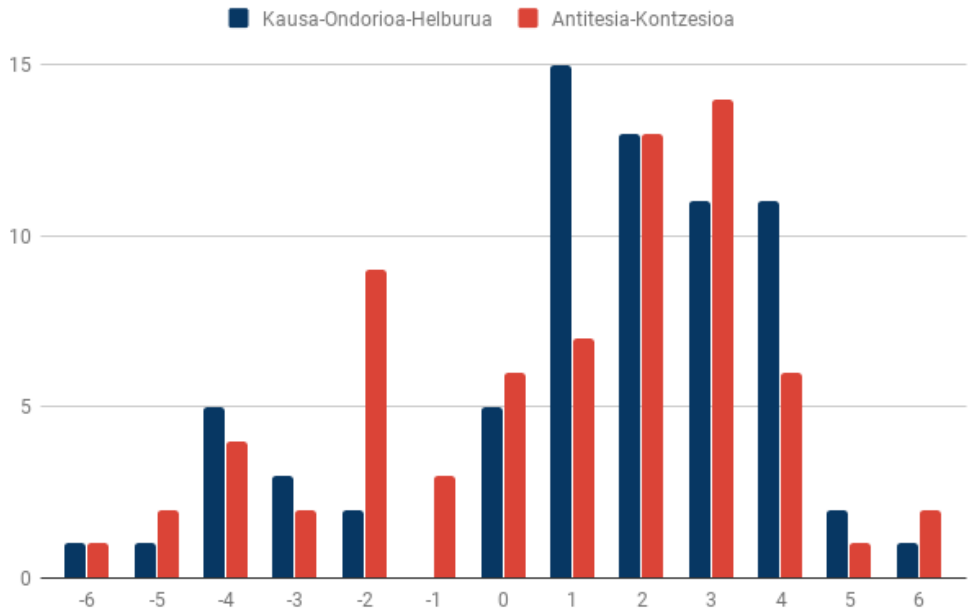
Ebaluazioa-Justifikazio erlaziozko diskurtso-egituraren taldearen instantzia gutxi daude unitate zentralaren aurretik, baina, 0 eta +1 distantzietan, bere instantzia kopurua asko igotzen da. Ondoren, berriz, instantzia kopurua jaitsi egiten da, unitate zentralaren aurreko egoerara itzuliz. KONTRASTEA erlaziozko diskurtso-egitura multinuklearrak ere modu berean jokatzen du. Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldean, ordea, distantzia guztietan dago erlaziozko diskurtso-egituraren talde horren instantziaren bat eta +1 eta +3 distantzien artean, bere instatzien kopurua nabarmen igotzen da. Beraz, Ebaluazioa-Interpretazioa erlaziozko diskurtso-egituraren taldeak distantzietan agerpen irregularra du eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeak, aldiz, agerpen erregularra. Bestalde, 5.6 irudiak Kausa-Ondorioa-Helburua eta



5.5 irudia: Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituren instantziak.

Antitesia-Kontzetsioa erlaziozko diskurtso-egituraren taldeen emaitzak erakusten ditu.

Bi erlaziozko diskurtso-egitura taldeek ezaugarri batzuk partekatzen dituzte, unitate zentraletik duten instantzia handieneko guneei dagokienez. Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldean bezala, erlazio horiek unitate zentralarekiko distantzia guztietan agertzen dira. Baina erlazio taldeek instantzien kantitatean bi goreneko gune dituzte. Bi multzo horietan, goreneko uneetako bat unitate zentralaren aurretik dago eta bestea unitate zentralaren ondoren. Antitesia-Kontzetsioa erlazioaren kasuan, bi goreneko uneak -2 eta $+3$ distantzietan daude. Instantzia maiztasunen bi goreneko une -4 eta $+1$ distantzietan agertzen dira Kausa-Ondorioa-Helburua erlaziozko diskurtso-egituraren taldean.



5.6 irudia: Kausa-Ondorioa-Helburua eta Antitesia-Kontzetsioa erlaziozko diskurtso-egitura taldeen instantziak.

2 Ikerketa hipotesia

1.2 ataleko ikerketa hipotesiari erantzunez, euskaran hizkuntza maila ezberdinetan testuinguruko balentzia-aldatzaileak daude.
Ezeztapena eta diskurtso-egiturako balentzia-aldatzaileak beste hizkuntzetan ere badaude, baina euskararen berezitasuna da morfologiazko balentzia-aldatzaileetan duen aberastasuna.

3 Ikerketa hipotesia

1.2 ataleko hipotesiari erantzunez, erlaziozko diskurtso-egituran nukleartasuna eta iritzi-testuetan unitate zentrala daude erlaziozko diskurtso-egituraren edota bertako osagaien orientazio semantikoa zeinu jakin batekoa izatea eragiten dutenak.
Nuklearitateak eta erlaziozko diskurtso-egituretakoko azken testuzatiak orientazio semantikoaren adostasun handiagoa dute. Beraz, elementu horiek diskurtso-unitateek erlaziozko diskurtso-egituraren orientazio semantikoa duten adostasuna indartu egiten dute.
Unitate zentralak testuko erlaziozko diskurtso-egiturek iritzi-testuaren orientazio semantikoarekin adostasun handiagoa izatea eragiten du. Erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, bere eta iritzi-testuaren orientazio semantikoaren adostasuna handiagoa da, hots, adostasuna indartu egiten du.

5.3.1.4. Unitate zentralaren azterketa

Unitate zentralaren azterketan, i) ezaugarri sintaktikoak aintzat hartuta domeinu bakoitzeko unitate zentralaren karakterizazioa eta ii) sentimendubalentziadun hitzen banaketa domeinu ezberdinetako unitate zentraletan aurkeztuko ditugu.

Unitate zentralen ezaugarri sintaktikoak

5.11 taulan, domeinu bakoitzeko unitate zentralean maiztasun handienez agertzen diren gramatika-kategoriak ikus daitezke[14].

Domeinuka iritzi-testuen unitate zentralek dituzten ezaugarriak identifikatu ostean, domeinuen artean zenbait berdintasun antzeman ditugu. Eguraldiaren domeinuko unitate zentralak beste domeinukoetatik ezberdinak dira hiru alderditan. Batetik, eguraldiaren domeinuak, denbora aipatzea ohikoa da (esaterako: *gaur*, *bihar* edo *asteburua*). Bestetik, beste domeinuetan ez bezala, ez da entitaterik agertzen bertan. Azkenik, aditza ez agertzea ere

[14]Gramatika-kategoriez gain, entitateak, denbora eta determinatzailea ere kontuan hartu ditugu unitate zentraleko hitzak maiztasunaren arabera aztertzean, mota horretako hitzak asko agertu direlako.

	Ize.	Adj.	Adi.	Adb.	Det.	Entitateak
Eguraldia	X	X		X		
Kirola			X		X	X
Musika	X		X			X
Politika		X				X
Filmak	X	X	X			X
Liburuak	X	X	X		X	X

5.11 taula: Unitate zentraletan bereizgarri diren ezaugarriak domeinuen eta gramatika-kategorien arabera.

ezaugarri bereizle bat da. Kirolaren domeinuari dagokionez, bertan determinatzaile ugari (partidetako emaitzak) agertzen dira. Aditzean zer entitatek irabazi edo galdu egin duen zehazten da.

Musika, film eta liburuetako domeinuek zenbait antzekotasun partekatzen dituzte. Izenak, adjektiboak, aditzak eta entitateak azaltzen dira unitate zentraletan. Azkenik, politikan, entitateak eta adjektiboak beste gramatika-kategoriak baino dezentez gehiago agertzen dira eta beste domeinuetatik, kasu horretan, ezberdindu egiten da.

Datu horiek aintzat hartuz, iritzi-testuen domeinuak unitate zentraletan gramatika-kategoria mota bat bestea baino gehiagotan agertzea eragiten du. Eguraldia eta kirolak ezaugarri bereizleak dituzte unitate zentralean eta beste domeinuek tarteko ezaugarriak dituzte. Unitate zentralen azterketa hori baliagarria izan daiteke etorkizunean iritzi-testuetako unitate zentralak antzemateko tresna edo baliabideak sartzeko.

Sentimendu-balentziadun hitzen banaketa unitate zentraletan

Corpuseko 192 unitate zentraletako hitzei sentimendu-balentzia euskarazko sentimenduen sailkatzailea erabiliz esleitu zaie. 5.12 taulak esleipen horren emaitzak erakusten ditu.

Corpuseko unitate zentraletan, 2.081 hitzetatik 258 hitzek (% 12,40ak) dute sentimendu-balentzia. Gramatika-kategorien ikuspegitik, 99 adjektiboek dute sentimendu-balentzia, 77na izen eta aditzek eta azkenik, bost adberbik. Datuak ehunekoetan emanda, adjektiboek sentimendu-balentziadun hitzen % 38 hartzen dute; izen eta aditzek % 30 eta azkenik, adberbioek % 2.

		%
Hitzak guztira	2.081	
Balentziadun hitzak guztira	258	12,40
Adjektiboak	*99*	*38*
Aditzak	*77*	*30*
Izenak	*77*	*30*
Adberbioak	*5*	*2*

5.12 taula: Euskarazko sentimenduen sailkatzaileak sentimendu-balentzia esleitutako unitate zentraleko hitzak gramatika-kategoriaren arabera sailkatuta.

Beraz, adjektiboetan sentimendu-balentziadun hitz gehiago agertzen diren arren, banaketa nahiko orekatua da. Domeinutik domeinura dauden ezberdintasunak ikus daitezke 5.13 taulan.

	Eguraldia		**Kirola**		**Literatura**		**Musika**		**Filmak**		**Politika**		**Guztira**	
Izena	7	0,16	11	0,33	17	0,36	10	0,26	12	0,33	20	0,33	77	0,30
Aditza	12	0,27	16	**0,48**	12	0,25	13	0,34	10	0,28	14	0,23	77	0,30
Adjektiboa	22	**0,50**	5	0,15	18	**0,38**	15	**0,39**	14	**0,39**	25	**0,42**	99	**0,38**
Adberbioa	3	0,07	1	0,03	-		-		-		1	0,02	5	0,02
Guztira	44		33		47		38		36		60		258	

5.13 taula: Sentimendu-balentziadun hitzen agerpena.

Sentimendu-balentziadun hitzen kopuruan, politikaren domeinuan 60 sentimendu-balentziadun hitz agertzen dira eta kirolean, sentimendu-balentziadun hitzak 33 dira. Eguraldiaren domeinuan, adjektiboaren gramatika-kategoriak nabarmenki balentziadun hitzen agerpen gehiago ditu (sentimendu-balentzia duten hitzen erdiak adjektiboak dira). Kirolaren domeinuan, aditza da sentimendu-balentziadun hitz gehien dituen gramatika-kategoria (aditzak % 48 dira). Literaturan, musikan, filmeetan eta politikan adjektiboak dira sentimendu-balentziadun hitz gehien dituztenak. Adjektiboak balentziadun hitzen % 38-42 hartzen dute, baina beste gramatika-kategoriak datu horietatik gertu daude. Laburbilduz, balentziadun hitzak dituen gramatika-kategoria ohikoena adjektiboena da, indar handiagoz eguraldian eta oso indar gutxiz kirolean non bertan aditza nagusitzen den modu nabarmen batean.

Atal honetako emaitzak eta 5.3.1.4 atalean lortutakoak, sentimenduen analisira begira unitate zentralak identifikatzeko, beren sentimendu-balentzia

eta orientazioa semantikoa kalkulatzeko eta domeinu bakoitzeko unitate zentralean zer gramatika-kategoriari eman behar zaion garrantzia antzemateko balio dezake. Gerora, bildutako informazio hori baliagarria izan daiteke euskarazko sentimenduen sailkatzailean unitate zentrala lantzeko modulu bat sortzeko.

4 Ikerketa hipotesia

1.2 ataleko hipotesiari erantzunez, domeinutik domeinura unitate zentralak ezaugarri bereziak dituela egiaztatu dugu.
Kirolaren domeinuan aditzak garrantzi handia du eta eguraldiaren domeinuan, aldiz, adjektiboak. Entitateak ere garrantzitsuak dira hainbat domeinutan.
Sentimendu-balentzia duten gramatika-kategorien kasuan, kirolaren domeinuan, aditzak ohikoenak dira eta beste gramatika-kategorietan adjektiboak.

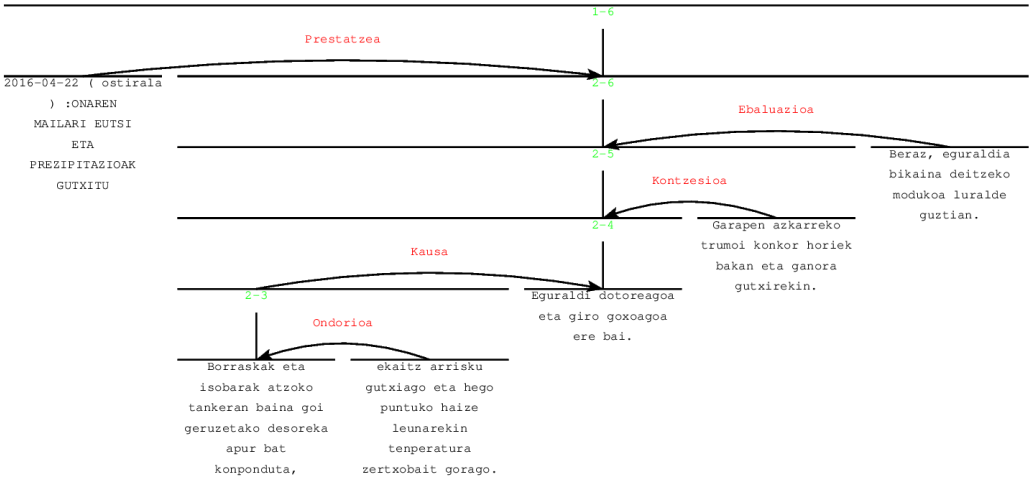
5.3.2. Egitura Erretorikoaren Teoria (RST)

Lan honetan, diskurtso-egitura aztertzeko Egitura Erretorikaren Teoria (*Rhetorical Structure Theory* edo RST) (Mann eta Thompson, 1988) hurbilpena hautatu dugu. Teoria honek testuaren egitura edo koherentzia deskribatzen du. Hizkuntzalaritza Konputazionalaren diziplinan eta erlaziozko diskurtso-egituran koherentzia aztertzeko gehien erabiltzen den teoria da eta euskaraz dagoen bakarrenetakoa da.

Hizkuntzalaritza Konputazionalean diskurtso-egitura lantzean bi fenomeno landu ohi dira: i) erreferentziazko fenomenoak: koerreferentziaren ebazpena lantzen da eta ii) erlaziozko fenomenoak: koherentzia erlazioen esleipena lantzen da. Egitura Erretorikoaren Teorian koherentzia erlazioak dira aztergai, baita gure lanean ere.

Mann eta Thompsonek (1988) RST konputagailu bidez testu-sorkuntza (*text generation*) egiteko asmoz sortu zuten. Testu-sorkuntza egin ahal izateko, testuaren ezaugarriak zehaztu behar dira eta testuen ezaugarririk garran-

tzitsuena koherentzia da. RSTk testuaren koherentzia deskribatzen duenez, konputagailua gauza izango litzateke testuen erlazio-egitura sortzeko edota prozesatzeko.



5.7 irudia: EGU06 iritzi-testua RSTz etiketatuta.

Testuaren egitura deskribatzeko, zenbait urrats bete behar dira. Hasteko, testuaren osagaiak identifikatu behar dira eta, ondoren, osagai horien arteko lotura edo erlazioa zehaztu. Testuan osagaia testu-zatia da. Testua aztertzeko testua zatitu egin behar da eta testua sortzeko testuko zatiak lotu egin behar dira. Testua aztertzean, testu-zatiak egiteari *segmentazioa* deritzo. Testu bat zatitzean, testu-zati bakoitzak proposizio bat adierazi behar du. 5.7 irudian, testua sei zatitan banatuta dago eta zati horietako bakoitzak proposizio bat adierazten du.

Ondoren, testu-zati horiek lotu egin behar dira eta hori egiteko koherentziazko erlazioak (*coherence relation*) ezartzen dira eta, testu-zatietan, koherentziazko erlazio horiek erlazio erretoriko (*rhetorical relation*) izena hartzen dute. RSTko zuhaitz-egituretan, erlazio erretorikoak errekurtsiboak dira. Hau da, erlazio bat beste erlazio baten unitatea da eta, modu horretan, testuko unitate guztiak zuhaitz-egituran aurkezten dira. 5.7 irudian, testu-zatiak erlazio erretorikoz (ONDORIOA, EBALUAZIOA, etab.) lotuta daude eta erlazioen artean errekurtsibitatea dago. Esaterako, ONDORIOA erlazio

erretorikoa KAUSA erlazio erretorikoaren unitatea da.

5.3.2.1. Nukleartasuna

RSTn nukleartasunaren bidez erlazio-egituretan unitate bakoitzak duen garrantzia zehazten da. RSTko zuhaitz-egituretan hierarkia bat osatzen da; izan ere, diskurtso-unitate guztiek ez dute garrantzi bera. Idazleak irakurlearri testuko mezua adierazi nahi dionean, unitate batzuek ideia garrantzitsuak dituzte eta beste batzuek, aldiz, bigarren mailako ideiak dira eta ideia nagusia ulertzeko baliagarriak dira. Beraz, diskurtso-unitate batzuk besteak baino garrantzitsuagoak dira. Diskurtso-unitate garrantzitsuenei nukleo-unitate (*nuclear unit*, N) deritze eta bigarren mailakoei satelite-unitate (*satellite unit*, S). Diskurtso-unitate bat nukleo- edo satelite-unitatea den ebazteari nukleartasuna (*nuclearity*) deritzo. 5.7 irudiko ONDORIOA erlazio erretorikoan, eskuineko diskurtso-unitatea ezkerrekoari lotzen zaio. Hori dela eta, ezkerreko unitatea nukleo-unitatea da eta eskuinekoa satelite-unitatea.

Nukleartasunak lotura du koherentziarekin, izan ere, idazleak testu batean adierazi nahi duena diskurtso-egiturako nukleo-unitateekin ulertzen da. Satelite-unitateekin, aldiz, testuaren ulermena zaila da, bertan dauden ideiak bigarren mailakoak direlako eta koherentziarik ez dagoelako. Erlazio nukleo-bakardunetan, satelitea nukleoari erlazio erretorikoz batzen zaio (N-S) eta erlazio horri *erlazio hipotaktiko* deritzo. Bi nukleoz (N-N) osatuta dauden erlazioei, aldiz, *erlazio parataktiko* deritze.

5.3.2.2. Erlazio erretorikoak

RSTn bi motatako erlazioak bereizten dira:

- Erlazio nukleoaniztunak (N-N). Mota honetako erlazioak nukleoz bakarrik osatuta daude eta, horrenbestez, erlazioa osatzen duten unitate guztiak maila berean daude eta unitateen artean ez da hierarkiarik sortzen. Erlazio nukleoaniztasunak hauek dira: KONJUNTZIOA, KONTRASTEA, DISJUNTZIOA, BATERATZEA, LISTA, BIRFORMULAZIO NUKLEOANIZTUNA eta SEKUENTZIA.

- Erlazio nukleobakardunak (N-S/S-N). Mota honetako erlazioetan, erlazioa osatzen duten unitateak ez daude maila berean eta hierarkia bat sortzen da. Nukleo-unitateak (N) adierazi nahi denaren zati garrantzitsuena biltzen du eta satelite-unitateak (S) nukleo-unitatean aipaturiko gaia hobeto ulertzen informazioa ematen du. Erlazio nukleobakardunak esaldien eta paragrafoen barruan ager daitezke. Mann eta Thompsonek (1988) bi motatako erlazio nukleobakardunak bereizten dituzte.

  – Edukizko erlazioak (*subject matter*). Unitateen arteko loturaren efektuak izaera semantikoa duenean gertatzen da. Hots, idazleak irakurleari bi unitateen artean erlazio bera dagoela adierazi nahi dio. ZIRKUNSTANTZIA, BALDINTZA, ELABORAZIOA, EBALUAZIOA, INTERPRETAZIOA, METODOA, KAUSA, ONDORIOA, AUKERA, HELBURUA, ARAZO-SOLUZIOA, EZ-BALDINTZATZAILEA eta ALDERANTZIZKO BALDINTZA mota honetako erlazioak dira.

  – Aurkezpenezko erlazioak (*presentational*). Unitateen arteko loturak izaera erretorikoa duenean gertatzen da. Kasu honetan, lotura horrek helburutzat irakurlearengan efektu jakin bat sortzea du. ANTITESIA, TESTUINGURUA, KONTZETSIOA, AHALBIDERATZEA, EBIDENTZIA, JUSTIFIKAZIOA, MOTIBAZIOA, PRESTATZEA, BIRFORMULAZIOA eta LABURPENA aurkezpenezko erlazioak dira.

5.3.2.3. Unitate zentrala

Iruskietak (2014) aipatzen duen moduan, pertsona batek testu baten laburpena egiten duenean, testuaren koherentzia globala edo makroegitura esplizitu egiteko gaitasuna erakusten du. Laburpen zientifiko eta akademikoetan ohikoak dira “The principal aim of this paper is to investigate ...” moduko egitura edo formulak. Horiek kasurik gehienetan adierazgailuak izan ohi dituzte eta adierazgailu horiek gramatika-kategoria ezberdinetakoak izan daitezke. Lan akademikoetan, ohikoak dira gramatika-kategoria ezberdinetako adierazgailu hauek: izenak (*paper*, *article*, *method*, *result* eta abar), aditzak (*discuss*, *introduce*, *analy-* eta *stud-*, besteak beste), erakusleak (*this*, *some*

eta abar) eta azkenik, izenordainak (*we*, *I*). Hala ere, adierazgailu guztiek ez dute beti makroegiturako gai nagusia adierazten; izan ere, mikroegitura adierazteko ere erabil baitaitezke.

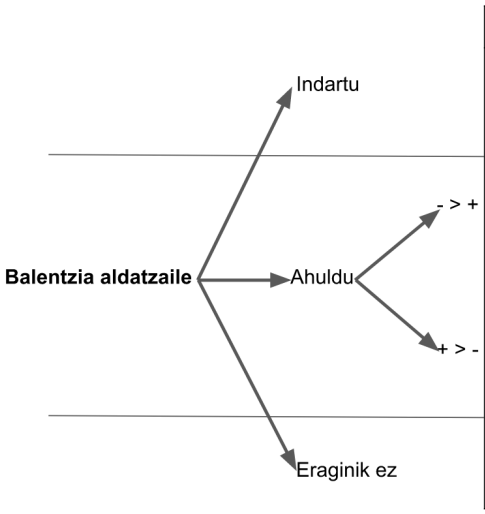
RST jarraituta eginiko zuhaitz-egitura batean, gai nagusia da erlazio-egiturako oinarrizko diskurtso-unitaterik (EDU) garrantzitsuena eta adierazgailuek hori identifikatzen lagun dezakete. Hau da, unitate zentrala da testuko EDU guztien artean garrantzitsuena eta zuhaitz-egitura eraiki nahi denean, bera erdigunetzat hartuta hasi behar da eraikitzen. Unitate zentralaren beste ezaugarri bat testuko diskurtso-unitate ez-errekurtsibo bakarra izatea da. Hots, unitate zentrala ez da inoiz beste erlazio baten satelite-unitatea izango.

RSTn oinarrituta egindako zuhaitz-egituran, beti egongo da unitate zentrala, nahiz eta tesi-adierazpen edo esaldi tematikorik ez egon. Izan ere, beti egon behar du zentralitatea duen nukleo-unitate bat eta ez-errekurtsiboa dena.

5.7 irudian, unitate zentrala *Eguraldia dotoreagoa eta giro goxoagoa ere bai* testu-zatia da. *Eguraldia* hitza makroegiturako gai nagusia izateak testu-zati hori unitate zentrala dela identifikatzen laguntzen digu.

5.4 Laburpena

Kapitulu honetan, euskaran hizkuntzako maila ezberdinetan dauden testuinguruko balentzia-aldatzaileak identifikatu eta horiek hitz eta sintagmen orientazio semantikoan eta sentimendu-balentzian duten eragina neurtu dugu. Emaitza 5.8 irudian ageri da.



Fonologia	Morfologia	Sintaxia	Diskurtsoa
<s>/<z> -> <x> <t> -> <tt> <z> -> <tx>	-garri -min -tsu -tza -en -tzar -zale -nahi -gura	ezin	· unitate zentrala · nukleoa · erlazioetako azken zatia
	-gabe des- -egi -keria ez-/-ez ezin-/-ezin -xe/-txe -txo -gaitz -ska/-xka -txa a-	ezin gabe izan ezik salbu ezta ez	
	-ezia -tasun	· edo/edota/ala + ez · ez ezik · egitura lexikalizatuak	

5.8 irudia: Euskarazko balentzia-aldatzaileak hizkuntza maila ezberdinetan.

Euskaran testuinguruko balentzia-aldatzaileak hizkuntza maila ezberdinetan daude eta beren eragina mota ezberdinetakoa da: hitzaren, sintagmaren edo esaldiaren sentimendu-balentzia indartu edo ahuldu egin dezakete. Halaber, badaude sentimendu-balentzian eraginik ez duten elementuak ere.

5.8 irudiko taulako lehen zutabean, euskarazko balentzia-aldatzaile fonologikoak zerrendatuta ageri dira. Bustidura adierazkorra balentzia-aldatzaile fonologikoa da eta beti hitzaren sentimendu-balentzia indartzen du. Izan ere, bustidura adierazkorra, samurtasuna edo gertutasuna adierazteko erabiltzen da eta mezuaren hartzaileari baliabide fonologiko hori erabiliz zuzentzean intentsitate indartu egiten da.

Bigarren zutabean, euskarazko hizkiak zerrendatuta eta sailkatuta ageri dira. Ahultzen duten balentzia-aldatzaile morfologikoak gehiago dira indar-

tzen dutenak baino. Kopuruaren ezberdintasunaren jatorrian ezeztapenezko hizkiak daude, horiek sentimendu-balentzia beti ahultzen baitute. Kontuan hartu behar da, ezeztapenezko hizkiek ±4 sentimendu-balentzia dutela, horrenbestez, hitzaren balentziaren zeinua alda dezakete baldin eta hitzaren sentimendu-balentzia ±3 edo baxuagoa bada.

Hirugarren zutabean, sintaxiko eta zehazki ezeztapeneko balentzia-aldatzaileak ageri dira. Ezeztapen-marka gehienek sintagmaren edo esaldiaren sentimendu-balentzia ahultzen dute. Kasu honetan ere, aintzat hartu behar da ezeztapen-marka horiek ±4 balentzia dutela eta sintagma edo esaldiaren sentimendu-balentzia ±3 edo txikiagoa bada, horien sentimendu-balentziaren zeinuan alderantzikatzea gertatzen da. Bestalde, *ezin* ezeztapen-markaren eskuinean adjektibo edo adberbioren bat badago, ezeztapen-marka horrek sentimendu-balentziaren indartzea eragina du. Bukatzeko, ezeztapen-markak ez du eraginik hautakariak diren lokailuekin agertzean, emendiozko lokailutzat edo ezeztapen kontrastibotzat jotzen den *ez ezik* egiturarekin agertzean eta ezeztapen-marka egitura lexikalizatu baten parte denean.

Azken zutabean, berriz, diskurtso-egiturari lotutako balentzia-aldatzaileak daude. Nukleartasunak, erlaziozko diskurtso-egituretako azken unitateak eta unitate zentralak eragina dute diskurtso-unitateek edota erlaziozko diskurtso-egiturek orientazio semantiko positiboa edo negatiboa izatean.

ONDORIOAK

6. KAPITULUA

Ekarpenak, mugak eta etorkizuneko lanak

Tesi-lan honen hasieran aipatu dugun moduan, gaur egun sarean iritzi-testu ugari egonda, testu horien informazio subjektiboa erauztea eta, zehazki, testu horiek balorazio positiboa edo negatiboa duten automatikoki jakitea abantaila bat da bai norbanakoentzat, bai gizartearentzat. Norabide horretan, euskarazko iritzi-testuetako informazio subjektiboa erauzteko tresna eta baliabideak garatu ditugu eta, halaber, euskarazko testuinguruko balentzia-aldatzaileak landu ditugu.

Jarraian, tesi-lan honetan euskarari dagokionez sentimenduen analisiaren inguruan egindako ekarpenak eta tesi-lanak izan dituen mugak aurkeztuko ditugu. Etorkizunera begira ditugun erronkak ere aipatuko ditugu.

6.1 Ekarpenak

Tesi hau euskarazko testu idatzien sentimenduen analisia lantzen duen lana da hizkuntzalaritza aplikatuaren ikuspegitik. Iritzi-testuak biltzen dituen corpus bat osatu eta sentimenduen lexikoi bat garatu ondoren, hitzen balentzietan eragiten duten fenomenoak aztertu ditugu. Jarduera horien ondorio dira garatu ditugun baliabideak eta tresnak.

6.1.1. Sentimenduen analisirako baliabideak eta tresnak

Euskarazko iritzi-testuen corpus etiketatua, euskarazko sentimenduen lexikoia eta dokumentu mailako sentimenduen sailkatzailea garatu ditugu.

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016). Corpus horrek sei domeinutako 240 iritzi-testu biltzen ditu. Diskurtso-egituraren nahiz subjektibitateaz etiketatuta dago. Diskurtso-egitura etiketatzeko RST hurbilpena (Mann eta Thompson, 1988) erabili dugu eta 70 iritzi-testu etiketatu ditugu. 192 iritzi-testuren unitate zentralak ere identifikatu ditugu. Halaber, testuek orientazio semantiko positiboa edo negatiboa duten ere zehaztu dugu. Gainera, 28 iritzi-testutan, erlaziozko diskurtso-egitura mota batzuei eta beren osagaiei (nukleoa eta satelitea) orientazio semantikoa (positiboa edo negatiboa den) nahiz sentimendu-balentzia (zenbakizko sentimenduen balioa) esleitu diegu.

- *Sentitegi* izeneko euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018). Sentimenduen lexikoi horrek 5 gramatika-kategoritako 1.237 hitz biltzen ditu. Bere oinarrian SO-CAL tresnaren gaztelaniazko lexikoia (Brooke *et al.*, 2009) dago eta ingelesezko lexikoiaz (Taboada *et al.*, 2011) ere aberastuta dago. Sentimenduen lexikoia sei domeinuetarako egokitu dugu (eguraldia, kirola, politika, musika, zinema eta literatura) eta, horretarako, Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) baliatu dugu.

- Tesi-lan honen beste ekarpen bat dokumentu mailako eta lexikoian oinarritutako sentimenduen sailkatzailea da. Euskarazko sentimenduen sailkatzailea garatzeko SO-CAL tresnan (Taboada *et al.*, 2011) oinarritu gara.

6.1.2. Euskarazko balentzia-aldatzaileen azterketa

Sentimenduen analisia lantzeko, hitzen sentimendu-balentzian eragiten duten hizkuntza maila ezberdinetako testuinguruko balentzia-aldatzaileak aztertu

ditugu. Balentzia-aldatzaileak aztertzerakoan, Polanyi eta Zaenenen (2006) lanean oinarritu gara.

- Fonologia eta morfologia mailako balentzia-aldatzaileak aztertu ditugu. Bertan ikusi dugunez, fonologian bustidura adierazkorra izeneko fenomenoa dago eta beren eragina da hitzaren sentimendu-balentzia indartzea. Morfologian, berriz, atzizkiak (adibidez, *-txo*/*-tzar*) aurkitu ditugu hitzaren balentzia indartu edo ahuldu dezaketenak. Gainera, ezeztapenarekin loturiko zenbait aurrizki (adibidez, *des-*) eta atzizki (adibidez, *-ezin*) ere aurkitu ditugu.

- Maila sintaktikoa ere aztertu dugu eta, zehazki, ezeztapen-markak eta irismena. Azterketak erakusten duenez, ezeztapen-markek hitzaren edo sintagmaren balentzia indartu (adibidez, *ezin*) edo ahuldu (esaterako, *gabe*) dezakete eta badaude aldaketarik eragiten ez duten kasuak (esaterako, *baino ez*) ere.

  Halaber, aipatzekoa da ezeztapen-marka batek (*ezin*) jokamolde bikoitza duela inguruan duen hitzaren gramatika-kategoriaren arabera. Ezeztapen-marka batzuk beste hitz jakin batzuekin maiztasun handiz agertzen direla ere erakutsi digute emaitzek. Egitura lexikalizatuak deitu diegu hauei. Irismenari dagokionez, *Murriztapen Gramatika* (Karlsson *et al.*, 1995) hurbilpena erabiliz, berau ezeztapen-markak eta egitura lexikalizatuak identifikatzea baino zailagoa dela behatu dugu, egitura eta luzera irregularrekoa baita.

- Azkenik, diskurtso maila ere landu dugu eta, horretarako, RST hurbilpena (Mann eta Thompson, 1988) erabili dugu. Lanean, erlaziozko diskurtso-egiturei eta unitate zentralari eman diogu garrantzia. Erlaziozko diskurtso-egituretan, ikusi dugu erlazioetako nukleoak edo azken EDUak osagai horien beraren eta erlaziozko diskurtso-egituraren arteko orientazio semantikoaren adostasuna indartzen dutela. Horretaz gain, emaitzek erakusten dute erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, erlazioaren beraren eta testu osoaren orientazio semantikoaren adostasuna indartu egiten dela. Horrenbes-

tez, emaitzek erakusten dute erlazioetan nukleoa eta azken EDUa eta testuan unitate zentrala balentzia indartzaile direla.

Unitate zentralean, balentziadun hitzak nola agertzen diren ere aztertu dugu. Emaitzen arabera, unitate zentraletako hitzen % 12,40k balentzia dute, batez ere, adjektibo, izen eta aditzek. Unitate zentralean, balentziadun hitzetan domeinuen arabera ezberdintasunak daudela behatu dugu. Esaterako, kirolaren domeinuan aditz-kategoriako hitzak dira balentziadunak, batik bat. Eguraldiaren domeinuan, ordea, adjektibo-kategoriako hitzak dira balentziadunak, batez ere.

6.2 Lanaren mugak

Atal honetan, tesi-lan honek izan dituen mugak aipatuko ditugu:

- Euskaran dauden testuinguruko balentzia-aldatzaile gehiago identifikatzen eta aztertzen jarraituko dugu. Tesi-lan honetan, hizkuntza maila bakoitzeko balentzia-aldatzaile bat aztertu dugu, baina oraindik badira aztertzeko geratu direnak. Hori da lan honek izan duen mugetariko bat. Aztertu gabe gelditu diren balentzia-aldatzaileen artean, baldintza, galde-perpausak edota konplexuagoa den ironia aipa ditzakegu. Balentzia-aldatzaileetako batzuek euskaran berezitasunak eduki ditzakete eta beste batzuk, aldiz, hizkuntza ezberdinetan antzekoak izan.

- Diskurtso mailako azterketa gehiago sakondu nahi dugu. Lan honetan, erlaziozko diskurtso-egiturak eta unitate zentrala aztertu ditugu. Baina, sentimenduen analisian eragin dezaketen diskurtso-egituran elementu gehiago egon daitezke. Horregatik, etorkizunean, haratago joan nahi dugu eta diskurtso-egiturako beste osagai batzuk ere aztertu nahi ditugu. Beste modu batean esanda, diskurtso-egituran dauden beste elementu batzuk aztertu nahi ditugu, azpiosagai zentrala esaterako. Halaber, erlaziozko diskurtso-egitura motetan ere sakondu nahi dugu eta beren artean egon daitezkeen ezberdintasunak aurkitu nahi ditugu.

6.3 Etorkizuneko lanak

Sentimenduen analisiaren azterketan eta beraren tratamendu konputazionalean hurrengo lan-ildoak aurreikusten dugu etorkizunerako:

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) are anitzagoa izatea. Gaur-gaurkoz corpusak egunkari eta webgune espezializatutako 6 domeinutako iritzi-testuak biltzen ditu. Etorkizuneko asmoa jendeak idazten dituen iritzi-testuak biltzea da. Metodologiaren garapenean aipatu bezala, jende arruntak idatzitako iritzi-testu gutxi aurkitu ditugu edota aurkitu ditugunen kalitatea ez da egokia izan diskurtsoa lantzeko. Asmoa da mota horretako testuak biltzea, oraindik aztertu gabe dauden bestelako fenomenoak ager baitaitezke. Adibidez, jende arruntak idatzita testuetan posible da letra bat errepikatuz hitzak luzatzea, baita emotikonoak bezalako elementuak agertzea ere, eta horiek orientazio semantikoan ere eragina izan dezakete.

- Corpusaren diskurtso-egituraren etiketatzea hobetu eta zabaldu. Mementoz, corpuseko 240 testuetatik 70 daude etiketatuta eta kopuru hori handitu egin nahi dugu.

- *Sentitegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) handitzea eta metodologian izandako gertaerei aurre egitea. Ikusi dugun moduan, sortu dugun lexikoiaren kalitatea ona da, baina oraindik ez da gauza subjektibitatea adierazten duten hitz edota esamolde guztiak identifikatzeko. Ondorioz, bere tamaina handitu nahi dugu eta domeinu gehiagotarako erabilgarria izatea lortu. Bukatzeko, hitz polisemikoen arazoa ere ebatzi nahi dugu, hau da, testuinguru ezberdinetan, aurkako bi orientazio dituzten hitzekin zer egin erabaki. Izan ere, *ikaragarria* bezalako hitzak aurkako orientazio semantikoa du, esaterako, pelikula bati buruz edo istripu bati buruz ari garenean.

- Dokumentu mailako sentimenduen sailkatzailea gehiago garatu nahi dugu. Tresna, mementoz euskarazko lexikoiaz, lematizatzaileaz eta erregela orokorrez osatuta dago, baina bertan aspektu gehiago integratu nahi ditugu. Esaterako, tesi-lan honetan landutako balentzia-

aldatzaileei lotutako modulua eratu nahiko genuke eta, modu horretan, tresnaren kalitatea igotzen den aztertu beharko litzateke. Modulu horretan, fonologiatik hasi eta diskurtso-egiturarainoko balentzia-aldatzaileei buruzko erregelak garatu nahiko genituzke.

Bibliografia

Aduriz I., Aldezabal I., Alegria I., Arriola J., Díaz de Ilarraza A., Ezeiza N., eta Gojenola K. Finite state applications for Basque. *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing*, 3–11, 2003.

Aduriz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M., eta Uria L. A Cascaded Syntactic Analyser for Basque. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 124–134, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24630-5.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., eta Maritxalar M. Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism. *Proceedings of Recent Advances in NLP (RANLP97), Tzigov Chark (Bulgary)*, 282–288, 1997.

Agerri R., Bermudez J., eta Rigau G. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Chair) N.C.C., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: A general lexical basis for the automatic processing of

Basque. *Proceedings of the IRCS Workshop on linguistic databases*. IRCS Workshop on linguistic databases., 2006.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., eta Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. *LREC-2002 Customizing knowledge in NLP applications workshop*, 1–6, 2002.

Alistair K. eta Diana I. Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*, 2005.

Alkorta J., Gojenola K., eta Iruskieta M. Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela, September*, 29 lib., 2016.

Alkorta J., Gojenola K., eta Iruskieta M. SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas*, 22(4), 2018.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria eta Patxi Salaberri (arg.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea, 127-134, Udako Euskal Unibertsitatea (UEU), Bilbo*, 2017.

Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B.n., eta Rotaetxe K. Euskal Gramatika Lehen urratsak (EGLU) II. *Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo*, 1985.

Banea C., Mihalcea R., Wiebe J., eta Hassan S. Multilingual Subjectivity Analysis Using Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 127–135, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613715.1613734`.

Barnes J., Badia T., eta Lambert P. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018a. European Languages Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L18-1104`.

Barnes J., Klinger R., eta Schulte im Walde S. Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages. *CoRR*, abs/1805.09016, 2018b. URL `http://arxiv.org/abs/1805.09016`.

Bautin M., Vijayarenu L., eta Skiena S. International sentiment analysis for news and blogs. *ICWSM*, 19–26, 2008.

Beigman Klebanov B., Burstein J., Madnani N., Faulkner A., eta Tetreault J. Building subjectivity lexicon(s) from scratch for essay data. *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'12, 591–602, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28603-2. URL `http://dx.doi.org/10.1007/978-3-642-28604-9_48`.

Benesty J., Chen J., Huang Y., eta Cohen I. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4. Springer, 2009.

Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G., eta Reynar J. Building a Sentiment Summarizer for Local Service Reviews. *WWW Workshop on NLP Challenges in the Information Explosion Era (NLPIX)*, 2008.

Blitzer J., Dredze M., eta Pereira F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P07-1056`.

Boldrini E., Balahur A., Martínez-Barco P., eta Montoyo A. EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. *Proceedings of the Fourth Linguistic Annotation Workshop*, 1–10. Association for Computational Linguistics, 2010.

Bond F., Ohkuma T., Da Costa L.M., Miura Y., Chen R., Kuribayashi T., eta Wang W. A Multilingual Sentiment Corpus for Chinese, English and Japanese. *Emotion and Sentiment Analysis*, page 59, 2016.

Boucher J. eta Osgood C.E. The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, 8(1):1–8, 1969.

Brooke J., Tofiloski M., eta Taboada M. Cross-linguistic sentiment analysis: From English to Spanish. *Proceedings of the International Conference RANLP-2009*, 50–54, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/R09-1010`.

Cambria E., Das D., Bandyopadhyay S., eta Feraco A. *A Practical Guide to Sentiment Analysis*, 5 lib. 01 2017. ISBN 978-3-319-55392-4.

Carenini G., Ng R., eta Pauls A. Multi-Document Summarization of Evaluative Text. *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E06-1039`.

Chen Y. eta Skiena S. Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 383–389, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P14-2063`.

Clematide S., Gindl S., Klenner M., Petrakis S., Remus R., Ruppenhofer J., Waltinger U., eta Wiegand M. MLSA - A Multi-layered Reference Corpus for German Sentiment Analysis. In Calzolari N., Choukri K., Declerck T., Dogan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors,

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, 3551–3556, 2012.

Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Cruz F.L., Troyano J.A., Pontes B., eta Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994, 2014.

Dadvar M., Hauff C., eta de Jong F. Scope of negation detection in sentiment analysis. *Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011*, 16–20. University of Amsterdam, 2 2011. ISBN not assigned.

Das D. eta Taboada M. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184, March 2018. ISSN 1574-020X. URL `https://doi.org/10.1007/s10579-017-9383-x`.

Das D. eta Martins A.F. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4(192-195): 57, 2007.

Ding X., Liu B., eta Yu P.S. A Holistic Lexicon-based Approach to Opinion Mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, 231–240, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. URL `http://doi.acm.org/10.1145/1341531.1341561`.

Ding X., Liu B., eta Zhang L. Entity discovery and assignment for opinion mining applications. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 1125–1134, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. URL `http://doi.acm.org/10.1145/1557019.1557141`.

Du W., Tan S., Cheng X., eta Yun X. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 111–120, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-889-6. URL http://doi.acm.org/10.1145/1718487.1718502.

Edmonds P. eta Hirst G. Near-synonymy and Lexical Choice. *Comput. Linguist.*, 28(2):105–144, June 2002. ISSN 0891-2017. URL http://dx.doi.org/10.1162/089120102760173625.

Egaña I. *Kritikarako hurbilketa literaturaren soziologiatik. Egunkari eta aldizkarietako euskal literatur kritikaren analisia (1975-2005).* Doktoretzatesia, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2013.

Egg M. eta Redeker G. How Complex is Discourse Structure? *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/796_Paper.pdf.

Eguchi K. eta Lavrenko V. Sentiment Retrieval Using Generative Models. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 345–354, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610124.

Elhuyar H.Z. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. Usurbil: Elhuyar, 2013.

Euskaltzaindia. Euskal Gramatika. Lehen urratsak I, Iruñea: Euskaltzaindia, 1985.

Euskara Institutua E. Sareko euskal gramatika, 2011. URL www.ehu.eus/seg.

Fernández J., Boldrini E., Gómez J.M., eta Martínez-Barco P. Análisis de sentimientos y minería de opiniones: el corpus emotiblog. *Procesamiento del lenguaje natural*, 47:179–187, 2011.

Fundazioa K.H. *Euskal hiztegi entziklopedikoa*. Klaudio Harluxet Fundazioa, 1995.

Gamon M. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `https://doi.org/10.3115/1220355.1220476`.

Gamon M., Aue A., Corston-Oliver S., eta Ringger E. Pulse: Mining Customer Opinions from Free Text. *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, 121–132, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-28795-7, 978-3-540-28795-7. URL `http://dx.doi.org/10.1007/11552253_12`.

Ghosh A., Li G., Veale T., Rosso P., Shutova E., Barnden J., eta Reyes A. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 470–478, Denver, Colorado, jun 2015. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S15-2080`.

Guo H., Zhu H., Guo Z., Zhang X., eta Su Z. Opinionit: A text mining system for cross-lingual opinion analysis. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 1199–1208, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. URL `http://doi.acm.org/10.1145/1871437.1871589`.

Halliday M.A.K., Matthiessen C., eta Halliday M. *An introduction to functional grammar*. Routledge, 2014.

Hassan A., Abu-Jbara A., Lu W., eta Radev D. A random walk: Based model for identifying semantic orientation. *Comput. Linguist.*, 40(3):539–

562, September 2014. ISSN 0891-2017. URL http://dx.doi.org/10.1162/COLI_a_00192.

Hatzivassiloglou V. eta McKeown K.R. Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. URL https://doi.org/10.3115/976909.979640.

Hu M. eta Liu B. Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. URL http://doi.acm.org/10.1145/1014052.1014073.

Hu Y.H., Chen Y.L., eta Chou H.L. Opinion Mining from Online Hotel Reviews A Text Summarization Approach. *Inf. Process. Manage.*, 53(2): 436–449, March 2017. ISSN 0306-4573. URL https://doi.org/10.1016/j.ipm.2016.12.002.

Iruskieta M. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, Informatika Fakultatea*, 2014.

Iruskieta M., Aranzabe M.J., Díaz de Ilarraza A., Gonzalez I., Lersundi M., eta Lopez de Lacalle O. The RST Basque TreeBank: an online search interface to check rhetorical relations. *4th workshop RST and discourse studies*, 40–49, 2013.

Iruskieta M., da Cunha I., eta Taboada M. Principles of a qualitative method for rhetorical analysis evaluation: A contrastive analysis English-Spanish-Basque. *Language Resources and Evaluation*, 49(2):263–309, 2015.

Jia L., Yu C., eta Meng W. The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. *Proceedings of the 18th ACM Confe-*

rence on Information and Knowledge Management, CIKM '09, 1827–1830, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. URL http://doi.acm.org/10.1145/1645953.1646241.

Jindal N. eta Liu B. Mining Comparative Sentences and Relations. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, 1331–1336. AAAI Press, 2006. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597348.1597400.

Kanayama H. eta Nasukawa T. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 355–363, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610125.

Karlsson F., Voutilainen A., Heikkila J., eta Anttila A., editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter & Co., Hawthorne, NJ, USA, 1995. ISBN 3110141795.

Karoui J., Farah B., Moriceau V., Patti V., Bosco C., eta Aussenac-Gilles N. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 262–272. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/E17-1025.

Kennedy A. eta Inkpen D. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22:110–125, 05 2006.

Kim J., Li J.J., eta Lee J.H. Evaluating Multilanguage-comparability of Subjectivity Analysis Systems. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 595–603, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858742.

Kim S.M. eta Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, 1–8, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-75-2. URL http://dl.acm.org/citation.cfm?id=1654641.1654642.

Kim S.M., Pantel P., Chklovski T., eta Pennacchiotti M. Automatically assessing review helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 423–430, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610135.

Kouloumpis E., Wilson T., eta Moore J. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.

Landis J.R. eta Koch G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 1977.

Leturia I., Arregi X., eta Sarasola K. Weba euskarazko corpus gisa. *Ekaia*, 27:281, 12 2014.

Li F., Huang M., Yang Y., eta Zhu X. Learning to Identify Review Spam. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, 2488–2493. AAAI Press, 2011. ISBN 978-1-57735-515-1. URL http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-414.

Li H., Cheng X., Adson K., Kirshboim T., eta Xu F. Annotating opinions in German political news. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 1183–1188, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/640_Paper.pdf.

Lim E.P., Nguyen V.A., Jindal N., Liu B., eta Lauw H.W. Detecting Product Review Spammers Using Rating Behaviors. *Proceedings of the 19th*

ACM International Conference on Information and Knowledge Management, CIKM '10, 939–948, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. URL http://doi.acm.org/10.1145/1871437.1871557.

Liu B. Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

Liu B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Liu B., Hu M., eta Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, 342–351, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9. URL http://doi.acm.org/10.1145/1060745.1060797.

Liu F., Li B., eta Liu Y. Finding opinionated blogs using statistical classifiers and lexical features. *Third International AAAI Conference on Weblogs and Social Media*, 2009.

Liu F., Wang D., Li B., eta Liu Y. Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 309–312, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N10-1042.

Long C., Zhang J., eta Zhut X. A Review Selection Approach for Accurate Feature Rating Estimation. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 766–774, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944654.

Lu Y., Tsaparas P., Ntoulas A., eta Polanyi L. Exploiting Social Context for Review Quality Prediction. *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 691–700, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. URL http://doi.acm.org/10.1145/1772690.1772761.

Mann W.C. eta Thompson S.A. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

McDonald R., Hannan K., Neylon T., Wells M., eta Reynar J. Structured Models for Fine-to-Coarse Sentiment Analysis. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 432–439. Association for Computational Linguistics, 2007. URL `http://aclweb.org/anthology/P07-1055`.

Meng X., Wei F., Xu G., Zhang L., Liu X., Zhou M., eta Wang H. Lost in Translations? Building Sentiment Lexicons using Context Based Machine Translation. *Proceedings of COLING 2012: Posters*, 829–838, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `https://www.aclweb.org/anthology/C12-2081`.

Miao Y., Su J., Liu S., eta Wu K. *SO-CAL Based Method for Chinese Sentiment Analysis*, 345–351. 12 2013.

Miller G.A. WordNet: A Lexical Database for English. *Commun. ACM*, 38 (11):39–41, November 1995. ISSN 0001-0782. URL `http://doi.acm.org/10.1145/219717.219748`.

Mohammad S.M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, 201–237. Elsevier, 2016.

Mohammad S.M., Salameh M., eta Kiritchenko S. How Translation Alters Sentiment. *J. Artif. Int. Res.*, 55(1):95–130, January 2016. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=3013558.3013562`.

Morsy S.A. eta Rafea A. Improving Document-level Sentiment Classification Using Contextual Valence Shifters. *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, NLDB'12, 253–258, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-31177-2. URL `http://dx.doi.org/10.1007/978-3-642-31178-9_30`.

Mujika L.M. Morfología de la composición lexical euskérica. *Fontes linguae vasconum: Studia et documenta*, 14(39):233–272, 1982.

Mullen T. eta Collier N. Sentiment analysis using support vector machines with diverse information sources. *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2004.

Na S.H., Lee Y., Nam S.H., eta Lee J.H. Improving Opinion Retrieval Based on Query-Specific Sentiment Lexicon. In Boughanem M., Berrut C., Mothe J., eta Soule-Dupuy C., editors, *Advances in Information Retrieval*, 734–738, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-00958-7.

Nakagawa T., Inui K., eta Kurohashi S. Dependency Tree-based Sentiment Classification Using CRFs with Hidden Variables. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 786–794, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858119`.

Ngoc Phu V. eta Thi Tuoi P. Sentiment classification using Enhanced Contextual Valence Shifters. 224–229, 10 2014.

O'Donnell M. RSTTool 2.4: A Markup Tool for Rhetorical Structure Theory. *Proceedings of the First International Conference on Natural Language Generation - Volume 14*, INLG '00, 253–256, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 965-90296-0-8. URL `https://doi.org/10.3115/1118253.1118290`.

Oñederra L. *Euskal fonologia: palatalizazioa: asimilazioa eta hots sinbolismoa*. Servicio Editorial de la Universidad del País Vasco= Euskal Herriko Unibertsitatea, 1990.

Oronoz M. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura*. Doktoretzatesia, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2009.

Otegi A., Imaz O., Díaz de Ilarraza A., Iruskieta M., eta Uria L. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84, 2017.

Pang B. eta Lee L. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 115–124, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL `https://doi.org/10.3115/1219840.1219855`.

Pang B. eta Lee L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008. ISSN 1554-0669. URL `http://dx.doi.org/10.1561/1500000011`.

Pang B., Lee L., eta Vaithyanathan S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL `https://doi.org/10.3115/1118693.1118704`.

Polanyi L. eta Zaenen A. Contextual valence shifters. 20 lib., 1–10. 01 2006.

Qu L., Ifrim G., eta Weikum G. The Bag-of-opinions Method for Review Rating Prediction from Sparse Text Patterns. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 913–921, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1873781.1873884`.

Quan C. eta Ren F. Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, 1446–1454, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3. URL `http://dl.acm.org/citation.cfm?id=1699648.1699691`.

Rao D. eta Ravichandran D. Semi-supervised Polarity Lexicon Induction. *Proceedings of the 12th Conference of the European Chapter of the As-*

sociation for Computational Linguistics, EACL '09, 675–682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609142.

Refaee E. eta Rieser V. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In Chair) N.C.C., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Riloff E., Patwardhan S., eta Wiebe J. Feature subsumption for opinion analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 440–448, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610137.

Rodriguez I., Martínez-Otzeta J.M., Lazkano E., eta Ruiz T. Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots. *International Conference on Social Robotics*, 666–675. Springer, 2017.

Rushdi-Saleh M., Martínez-Valdivia M.T., Ureña;a López L.A., eta Perea-Ortega J.M. OCA: Opinion Corpus for Arabic. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):2045–2054, October 2011. ISSN 1532-2882. URL http://dx.doi.org/10.1002/asi.21598.

San Vicente I., Saralegi X., eta Agerri R. EliXa: A modular and flexible ABSA platform. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 748–752, Denver, Colorado, June 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S15-2127.

Saralegi X., San Vicente I., eta Ugarteburu I. Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-resourced Languages. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, 96–108,

Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 978-3-642-37255-1. URL http://dx.doi.org/10.1007/978-3-642-37256-8_9.

Sarasola I. *Zehazki: gaztelania-euskara hiztegia*. Alberdania, 2005.

Sauri R. *A Factuality Profiler for Eventualities in Text*. Doktoretza-tesia, Waltham, MA, USA, 2008. AAI3304029.

Schulz J.M., Womser-Hacker C., eta Mandl T. Multilingual Corpus Development for Opinion Mining. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., eta Tapias D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Shin H., Kim M., Jang H., eta Cattle A. Annotation Scheme for Constructing Sentiment Corpus in Korean. *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 181–190. Faculty of Computer Science, Universitas Indonesia, 2012. URL http://aclweb.org/anthology/Y12-1019.

Silvennoinen O.O. Not only apples but also oranges: Contrastive negation and register. *Studies in Variation, Contacts and Change in English*, 19, 2017.

Snyder B. eta Barzilay R. Multiple Aspect Ranking Using the Good Grief Algorithm. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 300–307, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N07-1038.

Sorby S. Translating news from English to Chinese: Complimentary and derogatory language usage. *Babel*, 54:19–35, 01 2008.

Stone P.J., Dunphy D.C., Smith M.S., eta Ogilvie D.M. The General Inquirer: A Computer Approach to Content Analysis. 1966.

Stone P.J. eta Hunt E.B. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), 241–256, New York, NY, USA, 1963. ACM. URL http://doi.acm.org/10.1145/1461551.1461583.

Szmrecsanyi B. *Analyticity and syntheticity in the history of English*. Oxford University Press, 2012.

Taboada M. SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University, 2008.

Taboada M., Brooke J., Tofiloski M., Voll K., eta Stede M. Lexicon-based Methods for Sentiment Analysis. *Comput. Linguist.*, 37(2):267–307, June 2011. ISSN 0891-2017. URL http://dx.doi.org/10.1162/COLI_a_00049.

Tan P.N., Steinbach M., eta Kumar V. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN 0321321367.

Tsur O. eta Rappoport A. RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews, 2009. URL https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/180.

Turney P.D. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL https://doi.org/10.3115/1073083.1073153.

Turney P.D. eta Littman M.L. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. *CoRR*, cs.LG/0212012, 2002. URL http://arxiv.org/abs/cs.LG/0212012.

Turney P.D. eta Littman M.L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Trans. Inf. Syst.*, 21(4):

315–346, October 2003. ISSN 1046-8188. URL http://doi.acm.org/10.1145/944012.944013.

Vicente I.S. eta Saralegi X. Polarity lexicon building: to what extent is the manual effort worth? In Chair) N.C.C., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Vilares D., Garcia M., Alonso M.A., eta Gómez-Rodríguez C. Towards Syntactic Iberian Polarity Classification. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 67–73. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/W17-5209.

Villarroel Ordenes F., Ludwig S., De Ruyter K., Grewal D., eta Wetzels M. Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research*, 43(6):875–894, 2017.

Vinodhini G. eta Chandrasekaran R. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

Wan X. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 553–561, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613783.

Wan X. Co-training for Cross-lingual Sentiment Classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, 235–243, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9. URL http://dl.acm.org/citation.cfm?id=1687878.1687913.

Wang G., Xie S., Liu B., eta Yu P.S. Identify Online Store Review Spammers via Social Review Graph. *ACM Trans. Intell. Syst. Technol.*, 3(4):61:1–61:21, September 2012. ISSN 2157-6904. URL http://doi.acm.org/10.1145/2337542.2337546.

Wei B. eta Pal C. Cross Lingual Adaptation: An Experiment on Sentiment Classifications. *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, 258–262, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858842.1858890.

Westerski A. Sentiment Analysis: Introduction and the State of the Art overview. *Universidad Politecnica de Madrid, Spain*, 211–218, 2007.

Wiebe J. Learning Subjective Adjectives from Corpora. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 735–740. AAAI Press, 2000. ISBN 0-262-51112-6. URL http://dl.acm.org/citation.cfm?id=647288.721121.

Wiebe J. eta Mihalcea R. Word Sense and Subjectivity. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 1065–1072, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL https://doi.org/10.3115/1220175.1220309.

Wiebe J., Wilson T., Bruce R., Bell M., eta Martin M. Learning Subjective Language. *Comput. Linguist.*, 30(3):277–308, September 2004. ISSN 0891-2017. URL http://dx.doi.org/10.1162/0891201041850885.

Wiebe J.M., Bruce R.F., eta O'Hara T.P. Development and Use of a Gold-standard Data Set for Subjectivity Classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. URL https://doi.org/10.3115/1034678.1034721.

Wiegand M., Balahur A., Roth B., Klakow D., eta Montoyo A. A Survey on the Role of Negation in Sentiment Analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, 60–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858959.1858970`.

Wilson T.A. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. Doktoretza-tesia, University of Pittsburgh, 2008.

Yu H. eta Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL `https://doi.org/10.3115/1119355.1119372`.

Zhang W., Yu C., eta Meng W. Opinion Retrieval from Blogs. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 831–840, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. URL `http://doi.acm.org/10.1145/1321440.1321555`.

Terminologia eta laburdurak

Sentimenduen analisi (*Sentiment analysis*)

Orientazio semantiko (*Semantic orientation*)

Sentimendu-balentzi (*Sentiment valence*)

(Testuinguruko) balentzia-aldatzaile (*(Contextual) valence shifter*)

Sentimenduen sailkatzaile (*Sentiment classifier*)

Sentimenduen lexikoi (*Sentiment lexicon*)

Murriztapen Gramatika, MG (*Constraint Grammar, CG*)

Ezeztapen-marka (*Negation mark*)

(Ezeztapenaren) irismena (*Scope*)

Egitura lexikalizatu (*Lexicalized structure*)

Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory, RST*)

Unitate zentrala, UZ (*Central Unit, CU*)

Erlaziozko diskurtso-egitura (*Relational discourse structure*)

ERANSKINA

A. KAPITULUA

Murriztapen Gramatikako erregelak

Eranskin honetan ezeztapen-markak eta beren irismena identifikatzeko erregelak aurkezten ditugu. Erregelak Murriztapen Gramatika (MG) hurbilpenean idatzirik daude.

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT
   PUNT_GALD PUNT_ESKL PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = ``ez'';
LIST EZA = ``ez'';

LIST BAINO = ``baino'';
LIST BESTERIK = ``beste'';
LIST ZALANTZARIK = ``zalantza'';
LIST DUDARIK = ``duda'';
LIST BESTERIKGABE = ``besterik_gabe'';
LIST EZINEZKOA = ``ezinezko'';
LIST EZINEAN = ``ezinean'' "ezin'';
LIST NORA = ``nora'';
LIST EZEAN = ``ezean'';
```

```
LIST NORAEZEAN = ``nora_ezean'';

LIST EZINIZAN = ``ezin_izan'';
LIST EZIN = ``ezin" ``ezindu'';
LIST EZINHOBEA = ``ezin_hobe'';
LIST EZINIK = ``ezinik'';

LIST SALBU = ``salbu'';
LIST EZEZIK = ``ez_ezik'';
LIST IZANEZIK = ``izan_ezik'';

LIST GABE = ``gabe'';

LIST JUNTAGAILUA = ``edo'' ``ala'' ``edota'';
LIST BAI = "bai'';

LIST EZTA = ``ezta'';

#MAPPINGS

MAP (!gabe) TARGET (ADB) IF (0C GABE);
MAP (!ez) TARGET (PRT) IF (0C EZ);
MAP (!eza) TARGET (IZE) IF (0C EZA);

#LEXIKALIZAZIOAK

# (1)
#``baino ez'' egitura
MAP (!bainoezHAS) TARGET (LOT) IF (0C BAINO)    (1C EZ);

# (2)
# ``besterik ez'' egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)    (1
   C EZ);
```

```
#MAP (!besterikezBUK) TARGET (PRT) IF (-1C BESTERIK)
    (0C EZ);

# (3)
# ``zalantzarik gabe''
MAP (!zalantzarikgabeHAS) TARGET (IZE) IF (0C
    ZALANTZARIK) (1C GABE);
#MAP (!zalantzarikgabeBUK) TARGET (ADB) IF (-1C
    ZALANTZARIK) (0C GABE);

# (4)
# ``dudarik gabe''
MAP (!dudarikgabeHAS) TARGET (IZE) IF (0C DUDARIK) (1C
    GABE);
#MAP (!dudarikgabeBUK) TARGET (ADB) IF (-1C DUDARIK) (0
    C GABE);

# (5)
# ``besterik gabe'' antzemateko
MAP (!besterikgabe) TARGET (LOT) IF (0C BESTERIKGABE);

# (6)
# ``ezinezkoa'' antzemateko
MAP (!ezinezkoa) TARGET (ADJ) IF (0 EZINEZKOA);

#(7)
# ezinean antzemateko
MAP (!ezinean) TARGET (ADB) IF (0C EZINEAN);

#(8)
# ``nora ezean'' egitura.
MAP (!noraezeanHAS) TARGET (ADB) IF (0C NORA) (1C EZEAN
    );
MAP (!noraezeanBUK) TARGET (ADB) IF (-1C NORA) (0C
```

```
    EZEAN);
MAP (!noraezean) TARGET (ADB) IF (0C NORAEZEAN);

##ERREGELA ZEHATZAK

#(1)
# EZIN: ``ezin izan''
MAP (!ezinizanHAS) TARGET (ADI) IF (0C EZINIZAN);
MAP (!ezinizanTAR1) TARGET (ADL) OR (ADI) OR (ADT) OR (
    IZE) OR (ADJ) OR (DET) OR (IOR) OR (LOT) IF (-1
    EZINIZAN);
MAP (!ezinizanTAR2) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-2 EZINIZAN);
#MAP (!ezinizanBUK1) TARGET (ADI) OR (IZE) IF (-2
    EZINIZAN);
MAP (!ezinizanBUK2) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-3 EZINIZAN) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK3) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-4 EZINEAN) (NOT
     -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK4) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-5 EZINIZAN) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK5) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-6 EZINIZAN) (
    NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3
    PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK6) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-7 EZINIZAN) (
    NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA) (NOT -4
    PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA)
    (NOT -1 PUNTUAZIOA);
```

```
MAP (!ezinizanBUK7) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-8 EZINIZAN) (
    NOT -7 PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);


## EZIN orokorra
MAP (!EZIN) TARGET (IZE) IF (0C EZIN);
MAP (!ezinik) TARGET (ADB) IF (0C EZINIK);

# EZIN: ``ezinik''
MAP (!ezinik2) TARGET (ADI) IF (1C EZINIK);
MAP (!ezinik1) TARGET (IZE) IF (2C EZINIK);
##KOPONduta

# EZIN: ``ezin'' + beste aditzak
MAP (!ezinTAR) TARGET (ADT) OR (ADI) OR (ADL) OR (
    ADI_IZEELI) IF (-1C EZIN);
MAP (!ezinTAR2) TARGET (ADI) OR (ADL) OR (ADI_IZEELI)
    IF (-2C EZIN);
MAP (!ezinBUK) TARGET (ADI) OR (ADL) OR (ADI_IZEELI) IF
     (-3C EZIN);



# EZIN: ``ezin hobea''
MAP (!ezinhobea) TARGET (ADJ) IF (0C EZINHOBEA);
#MAP (!ezinadjHAS) TARGET (IZE) OR (ADB) IF (0C EZIN);
MAP (!ezinadjBUK) TARGET (ADJ) OR (ADB) IF (-1C EZIN);

# SALBU
MAP (!salbuBUK) TARGET (ADB) IF (0C SALBU);
MAP (!salbuHAS1) TARGET (IZE) OR (ADT) IF (1C SALBU);
MAP (!salbuHAS3) TARGET (IZE) OR (DET) IF (2C SALBU);
```

```
MAP (!salbuHAS3) TARGET (IZE) IF (3C SALBU);

# IZAN EZIK
MAP (!izanezik) TARGET (LOT) IF (0C IZANEZIK);
MAP (!izanezikHAS2) TARGET (IZE) IF (4C IZANEZIK);
MAP (!izanezikHAS1) TARGET (DET) IF (3C IZANEZIK);
MAP (!izanezikTAR1) TARGET (IZE) OR (DET) IF (2C
   IZANEZIK);
MAP (!izanezikTAR2) TARGET (ADI) OR (IZE) (1C IZANEZIK)
   ;

# EZ EZIK
MAP (!ezezik) TARGET (LOT) IF (0C EZEZIK);

#(2)
# EDO/EDOTA/ALA EZ
MAP (!juntagailuaHAS1) TARGET (ADI) OR (ADL) OR (ADT)
   OR (PRT) IF (1C JUNTAGAILUA) (2C EZ);
MAP (!juntagailuaHAS2) TARGET (ADI) OR (ADL) OR (IOR)
   IF (2C JUNTAGAILUA) (3C EZ);
MAP (!juntagailuaHAS3) TARGET (ADI) OR (IZE) IF (3C
   JUNTAGAILUA) (4C EZ);
MAP (!juntagailuaTAR) TARGET (LOT) IF (0C JUNTAGAILUA)
   (1C EZ);
#MAP (!juntagailuaBUK) TARGET (PRT) IF (-1C JUNTAGAILUA
   ) (0C EZ);

# EZ... EZ...
MAP (!ezez0) TARGET (PRT) IF (0C EZ) (2C EZ);
MAP (!ezez1) TARGET (DET) OR (IZE) OR (ADJ) IF (-1C EZ)
    (1C EZ);
MAP (!ezez2) TARGET (PRT) IF (-2C EZ) (0C EZ);
MAP (!ezez3) TARGET (DET) OR (IZE) OR (ADJ) IF (-1C EZ)
    (-3C EZ);
```

```
# EZ egiturak
MAP (!EZ-3) TARGET (IZE) IF (3C EZ) (NOT 2 BAINO) (NOT
    2 BESTERIK) (NOT 1 PUNTUAZIOA) (NOT 2 PUNTUAZIOA);
MAP (!EZ-2) TARGET (IZE) OR (ADI) OR (ADB) IF (2C EZ) (
    NOT 1C BAINO) (NOT 1C BESTERIK) (NOT 1C PUNTUAZIOA);
MAP (!EZ-1) TARGET (ADI) OR (ADJ) OR (ADL) OR (ADB) IF
    (1C EZ);
#MAP (!EZ0) TARGET (PRT) IF (0C EZ);
MAP (!EZ1) TARGET (ADT) OR (ADL) OR (ADI) OR (ADB) OR (
    IZE) OR (DET) OR (IOR) IF (-1C EZ) (NOT -2 BAINO) (
    NOT -2 BESTERIK);
MAP (!EZ2) TARGET (ADI) OR (ADL) OR (ADT) OR (DET) OR (
    IZE) OR (ADJ) OR (ADB) OR (IOR) IF (-2C EZ) (NOT -3C
     BAINO) (NOT -3C BESTERIK) (NOT -2 PUNTUAZIOA) (NOT
    -1 PUNTUAZIOA);
MAP (!EZ3) TARGET (ADI) OR (ADL) OR (ADT) OR (IZE) OR (
    ADJ) OR (IOR) IF (-3C EZ) (NOT -4C BAINO) (NOT -4
    BESTERIK) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (
    NOT -1 PUNTUAZIOA);
MAP (!EZ4) TARGET (ADI) OR (ADL) OR (IZE) OR (DET) OR (
    IOR) IF (-4C EZ) (NOT -5C BAINO) (NOT -5 BESTERIK) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ5) TARGET (ADI) OR (IZE) OR (IOR) OR (ADT) OR (
    ADL) IF (-5C EZ) (NOT -6C BAINO) (NOT -6 BESTERIK) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ6) TARGET (ADI) OR (IZE) OR (IOR) IF (-6C EZ) (
    NOT -7C BAINO) (NOT -7 BESTERIK) (NOT -5 PUNTUAZIOA)
     (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ7) TARGET (ADI) OR (IZE) OR (IOR) IF (-7C EZ) (
    NOT -8C BAINO) (NOT -8 BESTERIK) (NOT -6 PUNTUAZIOA)
```

```
    (NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3
   PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ8) TARGET (ADI) OR (IZE) OR (ADL) OR (IOR) IF
   (-8C EZ) (NOT -9C BAINO) (NOT -9 BESTERIK) (NOT -7
   PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
   (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
   PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ9) TARGET (ADI) IF (-9C EZ) (NOT -10C BAINO) (
   NOT -10 BESTERIK) (NOT -8 PUNTUAZIOA) (NOT -7
   PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
   (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
   PUNTUAZIOA) (NOT -1 PUNTUAZIOA);

MAP (!izeezaHAS) TARGET (IZE) IF (1C EZA);
MAP (!izenezaBUK) TARGET (IZE) IF (0C EZA);



# GABE egiturak

MAP (!gabeAUR) TARGET (IZE) OR (ADI) OR (ADJ) OR (IOR)
   IF (1C GABE);
MAP (!gabeAUR2) TARGET (IZE) OR (ADI) IF (2C GABE);
MAP (!gabeAUR3) TARGET (IZE) IF (3C GABE);

# EZTA
#MAP (!EZTA0) TARGET (LOT) IF (0C EZTA);
MAP (!EZTA1) TARGET (ADT) OR (ADL) OR (ADI) OR (ADB) OR
    (IZE) OR (DET) OR (IOR) IF (-1C EZTA) (NOT -2 BAINO
   ) (NOT -2 BESTERIK);
MAP (!EZTA2) TARGET (ADI) OR (ADL) OR (ADT) OR (DET) OR
    (IZE) OR (ADJ) OR (ADB) OR (IOR) IF (-2C EZTA) (NOT
    -3C BAINO) (NOT -3C BESTERIK) (NOT -2 PUNTUAZIOA) (
   NOT -1 PUNTUAZIOA);
MAP (!EZTA3) TARGET (ADI) OR (ADL) OR (ADT) OR (IZE) OR
```

```
    (ADJ) OR (IOR) IF (-3C EZTA) (NOT -4C BAINO) (NOT
    -4 BESTERIK) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA)
    (NOT -1 PUNTUAZIOA);
MAP (!EZTA4) TARGET (ADI) OR (ADL) OR (IZE) OR (DET) OR
    (IOR) IF (-4C EZTA) (NOT -5C BAINO) (NOT -5
    BESTERIK) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA5) TARGET (ADI) OR (IZE) OR (IOR) OR (ADT) OR
    (ADL) IF (-5C EZTA) (NOT -6C BAINO) (NOT -6
    BESTERIK) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA6) TARGET (ADI) OR (IZE) OR (IOR) IF (-6C
    EZTA) (NOT -7C BAINO) (NOT -7 BESTERIK) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA7) TARGET (ADI) OR (IZE) OR (IOR) IF (-7C
    EZTA) (NOT -8C BAINO) (NOT -8 BESTERIK) (NOT -6
    PUNTUAZIOA) (NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA)
    (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1
    PUNTUAZIOA);
MAP (!EZTA8) TARGET (ADI) OR (IZE) OR (ADL) OR (IOR) IF
    (-8C EZTA) (NOT -9C BAINO) (NOT -9 BESTERIK) (NOT
    -7 PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA9) TARGET (ADI) IF (-9C EZTA) (NOT -10C BAINO
    ) (NOT -10 BESTERIK) (NOT -8 PUNTUAZIOA) (NOT -7
    PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
    (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
```

Tesi honen idazketa
2019ko urriaren 15ean
bukatu zen.

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Didactics of Language and Literature

PhD dissertation

Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels

---

Jon Alkorta Agirrezabala

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Didactics of Language and Literature

Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels

This summary is a shortened and translated version of the dissertation entitled "Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa", written by Jon Alkorta under the supervision of Dr. Mikel Iruskieta and Dr. Koldo Gojenola. It also includes the papers which the candidate has published in English on the research presented here.

Donostia, 15th of October 2019

Aitari, Amari eta Josuri

Acknowledgements

Lehenik eta behin, nire zuzendariak eskertu nahiko nituzke: Koldo eta Mikel, eman didaten laguntzagatik. Eskerrik asko beti nire atzean egoteagatik eta baita tesian zehar sortu diren zalantzetan laguntza eskaintzeagatik ere. Azkenik, eskerrik asko egin dizkidazuen iruzkinengatik (asko lagundu baitidate) eta ikerkuntzaren mundua nolakoa den erakusteagatik ere. Halaber, eskerrak Maxuxi eta Rodriri tesi-txostena hobetzen laguntzeagatik.

Eskerrik asko tesian lagundu didaten beste pertsonei ere. Eskerrik asko Arantxari *Eustaggerr*ekin izan ditudan une zailetan laguntzeagatik eta Intxari *include*k Latexen duen garrantzia erakusteagatik. Eskerrak Estherri Kanadatik Ixarako konexioa ahalbidetzeagatik, Kikeri emandako laguntzagatik eta Amaiari administrazioaren munduan laguntzeagatik. Eskerrik asko, halaber, Itziar Gonzalez-Diosi *Murriztapen Gramatika*rekin eta tesi-txostena vs. Latex borrokan laguntzeagatik. Eskerrik asko zerrenda honetan aipatu gabe gelditu diren baina lagundu didaten beste guztiei ere.

Eskertzekoa da tesia hasi eta amaitu arte inguruan mugitu denari ere: 318 bulegoari giro ona sortzeagatik, tupper txokoari askotariko gaiak (batzuetan, frikiak) aipatzeagatik, pintxo-poteetan eta egindako beste planetan atera zareten guztioi eta, azkenik, estatistika klaseetan, Oierri izandako pazientziagatik eta *nukleo duro*ri estatistika ikasteko laguntza morala emateagatik.

También tengo que agradecer a Maite Taboada por su ayuda en la estancia de Vancouver y en los trabajos relacionados con el discurso y análisis de sentimientos.

Familia ere ezin da aipatu gabe utzi. Eskerrik asko aitari, amari eta Josuri beti hor egoteagatik, eta laguntza behar izan dudanean laguntzeagatik.

Eskerrik asko kuadrillari ere, asteburuetan, musa eta *Comunio*rekin une

biziki ziragarriak emateagatik. Eskerrak baita Urkizu eta Gazteluko bazkariengatik eta Kanpezura eta Jakatik mendira egindako planengatik ere.

Bukatzeko eskerrik asko Amets eta Lierni pisukideei eta aipatzea ahaztu zaizkidan beste guztiei.

Mila-mila esker denoi!

Official acknowledgments

This dissertation was funded by the PRE_2014_1_190, PRE_2015_1_0121, PRE_2016_2_0153, PRE_2017_2_0041, EP_2017_1_0003 and PRE_-2018_2_0033 grants awarded by the Basque Government's Education, Universities and Research Department as well as by IXA group, Research Group of type A (2010-2015) (Basque Government, ref.: IT344-10) and READERS (CHIST-ERA, MINECO, ref.: PCIN-2013-003-C02-02).

Abstract

In the research work, we have taken the first steps on sentiment analysis from point of view of applied linguistics. The work developed consists of two aspects. On the one hand, based on the contextual valence shifter approach to sentiment analysis, we have identified valence shifters of different language levels in Basque, from phonology to discourse through morphology and syntax. Moreover, we have measured their effect on the sentiment valence of different linguistic elements.

The second aspect of this work focuses on the creation and development of tools and resources for sentiment analysis in Basque. Firstly, a corpus with 240 opinion texts has been built and it has been annotated: from the point of view of semantic orientation and discourse information. Secondly, a sentiment lexicon with 1,237 entries has been created. Finally, a document level and lexicon based sentiment classifier has been created based on the SO-CAL tool.

Contents

List of Figures

List of Tables

INTRODUCTION

1 Introduction

1.1 Motivation

Opinions are the basis of human activity and they have a significant impact on our behavior (Liu, 2012). Even our choices and beliefs about reality are largely conditioned by what the world sees and evaluates. As a result of this situation, we often seek the opinions of others when we make a decision. It is a situation that occurs in individuals and organizations.

In opinion texts, there are some concepts related to subjectivity. These include sentiment, evaluation, attitudes, and emotions, among others. All of them, in addition to opinion, are objects of study in sentiment analysis or opinion mining.

As Liu (2012) reminds us, the beginning and growth of this area is entirely associated with the creation of social networks. The social networks of the web are varied: comments, forum discussions, blogs, microblogs, and Twitter. All of these applications have made available a large volume of opinions in digital resources for the first time in human history. Therefore, it is one of the largest and most active growth areas in natural language processing since 2000. Sentiment analysis is also used in data mining, web mining and text mining.

In general, sentiment analysis has become important in business and society, which has expanded the field from computational science to management science and social science. Around it, the industry has developed and companies have also created services to work the area. So, it can be said that sentiment analysis can be found in many business or social fields.

As we have seen, sentiment analysis internationally has undergone considerable development. English is the world's *lingua franca* today and the majority of the works in sentiment analysis have been done in that language. In contrast, little has been done in Basque.

Sentiment analysis is a very broad area. Some tasks are connected with computer science and others with linguistics. Besides, some task involves two approaches: based on language knowledge and based on statistical methods.

From a linguistic point of view, the first step is to find out the opinion of the word or, in other words, the subjective information. The next step is to look for language-related phenomena that affect the subjective information of these words. Let us look at the following examples.

(1) I $[\text{like}_{+}]_{+}$ that movie.

(2) I $[\text{like a lot}_{+}]_{++}$ that movie.

(3) I do not $[\text{like}_{+}]_{-}$ that movie.

(4) I $[\text{would like}_{+}]_{0}$ very much that movie if it were more vivacious.

In Example (1), it is mentioned that a person liked the movie, so the opinion of the sentence is positive. In the same way, in Example (2), the opinion of the sentence is positive but the intensifier[1] *a lot* makes this sentence more positive than the previous sentence. In contrast, in Example (3), although the opinion of the verb (*like*) is positive, the opinion of the sentence is negative. Finally, in Example (4) there is only one word (*like*) with positive opinion but the sentence does not express an opinion, because the opinion is in a conditional mood.

In these examples, although there is one word that expresses opinion in all sentences (*like*), the opinion expressed by the whole sentence can be different. Some linguistic phenomena modify the meaning of words and with opinion. The phenomena that change the value of a word or phrase are called contextual valence shifters.

This aspect mentioned motivates our work. On the one hand, we aim to create tools and resources related to sentiment analysis for Basque to extract subjective content from Basque texts. On the other hand, we seek linguistic phenomena that influence words or phrases with opinion, from phonology to discourse.

[1]Intensifiers are usually adverbs or adjectives. They have little inherent semantic content but they intensify the meaning of the words or phrases.

Our first aim is to create basic tools and resources for sentiment analysis in Basque. Specifically, we want to create *i)* a corpus of opinion texts, *ii)* a sentiment lexicon and *iii)* a document-level sentiment classifier.

There are few corpora of opinion texts and sentiment lexicons in Basque and the current tools are not useful for our goals, because they do not identify relevant language features. In the case of opinion texts, some corpora or databases collect reviews and tweets that appear on websites but they are not useful to us because texts are short and, as a result, it is difficult to analyze certain elements of discourse structure.

In terms of lexicons, the available lexicons indicate the semantic orientation of words (for instance, Stone and Hunt (1963) indicates if the word is positive or negative) but they do not indicate intensity (for example, *good* and *excellent* are positive but their difference in intensity is not indicated) and therefore, they are not useful to us.

As far as the sentiment classifier is concerned, there are few works in Basque language that perform sentiment classification. One of them is EliXa (San Vicente et al., 2017) and it performs an aspect-level sentiment classification[2]. Our objective is to provide Basque with another resource for language processing and, for this purpose, we want to create a document-level sentiment classifier based on lexicon. It must be said that to meet this objective, it is necessary to fulfill the aforementioned objectives: it is necessary to create a sentiment lexicon to integrate in sentiment classifier and to study valence shifters is also needed to measure the changes that affect words with opinion.

Our second objective is to study linguistic phenomena that affect the sentiment valence of the words of lexicons created in the previous step and to measure their influence on words.

As we have seen in Examples (1), (2), (3) and (4), for a document-level classifier based on the sentiment lexicon, the lexicon alone is not enough. More linguistic information is needed to make a good sentiment classification. Therefore, based on the work of Polanyi and Zaenen (2006), we will aim to identify contextual valence shifters[3] at different language levels in Basque.

In short, this thesis combines Basque and sentiment analysis. There has been little work done on Basque in sentiment analysis and most are limited to

[2]In aspect-level sentiment analysis, the aim is to identify aspects related to the entity under study (if the entity is London, the aspects are economics, tourism, etc.) and to determine whether it is positive or negative.

[3]Contextual valence shifters are phenomena that cause changes in the sentiment valence of words and/or sentences. This change can be a strengthening or weakening of the valence.

the creation of a sentiment lexicon. We want to provide the Basque language with resources and tools, as well as investigate the features of Basque that affect sentiment classification.

1.2 General hypotheses

In the previous section, we have mentioned sentiment analysis as a motivation for the thesis, as well as its utility in society. In this section, we will outline the scope of the thesis. This work has four general hypotheses:

- **Research Hypothesis 1:** *The quality of the sentiment lexicon in Basque obtained by translation is comparable to those that can be derived from the corpus or lexical databases.*
  Sentiment lexicons assign the semantic orientation and sentiment valence to words of a texts and this is the first step in calculating the semantic orientation of a text. There are two common approaches to creating lexicons: *i)* corpus and *ii)* lexical databases.

  Our approach is a mixture of both approaches. We will use two sentiment lexicons (Spanish and English lexicons) as a basis and we will translate them by enriching the information provided by the Basque Opinion Corpus.

  In our opinion, the sentiment lexicon that has already created is comparable in quality to corpus-based and lexical database-based lexicons. If the translation methodology is optimal, no information will be lost or transformed along and it would maintain its initial quality. Besides, the use of a corpus with opinion texts would prevent the incorrect assignments of semantic orientation and sentiment valence to words with opinion.

- **Research Hypothesis 2:** *At all levels of language (from phonology to discourse) certain linguistic phenomena influence influence the sentiment valence of words (and phrases).*
  We think that this effect can either strengthen or weaken their sentiment valence. In our opinion, these linguistic phenomena can appear at different levels of grammar: for example, through affixes in phonology, through different syntactic phenomena or even through discourse. We believe it is necessary to use this kind of information in the creation

of a document-based sentiment classifier, otherwise, the results of the classifier could be poorer.

- **Research Hypothesis 3**: *In the discourse structure, there are constituents that affect the semantic orientation of EDUs and discourse relations.*
  From our point of view, in the case of EDUs or discourse relations, having a positive or negative semantic orientation is not a random event. In other words, we believe that these elements are affected by their position in the rhetorical structure tree. Specifically, we hypothesize that some or all of the factors in the discourse structure affect the semantic orientation of EDUs and discourse relations.

- **Research Hypothesis 4:** *Due to the domain, the central units in texts are different regarding the grammatical category semantic orientation of words.*
  Although the central unit is present in all texts, we believe that its features differ in domains. We think that the domain could affect in the way the words with semantic orientation appear in central units.

1.3 Goals

In the previous sections, we discussed the context of this work and the general hypotheses to guide our research. In this section, we will outline our specific goals to verify if they support our assumptions.

- **Objective 1:** Basic tools for studying sentiment analysis and creation of tools and resources.
  - ***Objective 1.1:*** To create a corpus of opinion texts in Basque, enriched with the semantic orientation and discourse information using *Rhetorical Structure Theory* (RST).
  - ***Objective 1.2:*** To generate a sentiment lexicon in Basque, indicating the semantic orientation of words by a numerical value.
  - ***Objective 1.3:*** To create a document-based sentiment classifier for Basque, based on sentiment lexicon.

- ***Objective 2:*** Identify and measure the effect of contextual valence shifters in Basque that affect the valence and sentiment of words:

- ***Objective 2.1:*** To identify phonological valence shifters and measure their effect.
- ***Objective 2.2:*** To identify morphological valence shifters and measure their effect.
- ***Objective 2.3:*** Regarding syntax, to measure how negation marks affect the sentiment valence of words or phrases.
- ***Objective 2.4:*** Regarding discourse structure, to measure the effect of the central unit on opinion texts and the effect of nuclearity on discourse relations.
- ***Objective 2.5:*** To study the appearance of discourse relations in the RST-tree.
- ***Objective 2.6:*** Analysis of central units of opinion texts: to study the grammatical category of words and the distribution of words with semantic orientation in central units.

1.4 Organization of the thesis report

This thesis consists of six chapters. In the following lines, we will summarize the contents of each chapter:

- **Chapter 1: Introduction.**
  In chapter one, we first discuss work motivation, general assumptions, goals, and research questions.

- **Chapter 2: Methodology and resources.**
  In this chapter, we will discuss the methodology and resources used to accomplish the goals mentioned in the introduction and to answer the research questions. Following an overview of the methodology, we will outline the steps to create a corpus of opinion text in Basque and a sentiment lexicon. After that, we will define the methodology for identifying the valence shifters. We will explain individually what has been done in phonology, morphology, syntax, and discourse. We will then explain the steps to create the document-level sentiment classifier in Basque.

- **Chapter 3: Resources developed.**
  In the third chapter, we will discuss the resources developed and their

evaluation as a result of following the defined methodology. Some of the resources we have mentioned include the Basque Opinion Corpus, a sentiment lexicon called *Sentitegi*, and a document-level sentiment classifier.

- **Chapter 4: Valence shifters at different language levels.**
  In this fourth chapter, we will also report on the results obtained. Specifically, we will explain the contextual valence shifters we have found at the phonological and morphological, syntactic, and discourse levels. Moreover, we have also measured their influence on the semantic orientation of words and/or sentences, or the sentiment valence of them.

- **Chapter 5: Conclusions and future work.**
  This chapter will first discuss the implications of the thesis work. The initial objectives will be mentioned and research questions will be answered. The contributions of the work will also be mentioned. Finally, we will focus on future work.

- **Terminology and abbreviations.**
  In this appendix, we will list the terminology and abbreviations mentioned in this thesis report.

METHODOLOGY

2 Methodology

We followed a specific methodology to accomplish the goals set out and we also used some specific resources for that. In this chapter, we will deal with them. There are three main steps in the methodology:

1- Create basic resources for carrying out research. In other words, we explain the creation of opinion text corpus and sentiment lexicon in Basque.

2- Identify the valence shifters. We analyze and measure the linguistic phenomena that may affect words with opinions at different language levels and their effects.

3- Develop a document-level sentiment classifier in Basque. This sentiment classifier will be based on a lexicon we developed and a sentiment classifier of the SO-CAL tool (Taboada et al., 2011).

2.1 Creating resources for sentiment analysis

In the first step, we created resources for sentiment analysis. We created the Basque Opinion Corpus and the sentiment lexicon in Basque.

2.1.1 The Basque Opinion Corpus

To create the Basque Opinion Corpus, we first had to decide what type of corpus we wanted to create. We decided to use the features of the corpus'

texts as follows:

- A clear appraisal of the opinion texts. Since some of the texts did not show a clear opinion, we preferred texts with an opinion either for or against.
- To have resources to express a rich syntactic structure and a clear appraisal. We wanted the opinion texts to have different syntactic structures and different methods of appraisal.

After, we collected Basque opinion texts from websites. These websites are either specialized or belong to magazines and newspapers. In short, we used three different sources: *i)* newspapers, *ii)* specialist websites and, finally, *iii)* several blogs.

After finding websites to provide the Basque Opinion Corpus, we established the corpus' characteristics:

- The corpus will be large and will contain 240 opinion texts.
- The corpus will cover many fields. Opinion articles will belong to six fields: weather, politics, sport, movies, music and literature.
- The corpus' texts will be appraised in a balanced manner. Each field will be balanced concerning its appraisal.

In the next step, we created a database with the collected opinion texts. In that database, we specified the following criteria for each opinion text:

- Code: We gave each opinion text a code as follows: *THEME-number appraisal*.
- Title: A title was given to each opinion text.
- Address: We stated where the opinion text is located on the Internet, showing its source to allow anyone to access it.
- Appraisal: In the database, we specified whether each opinion text was either positive or negative.
- Size: We also specified the number of words in each opinion text, using the *Analhitza* tool (Otegi et al. 2017).

Furthermore, we also adapted the format of opinion texts so they could be processed using natural language tools. We connected articles to *txt* format and UTF-8 codification formats. The development phase set contains 144 opinion texts and the training set 48.

We also wanted to measure whether the corpus is suitable for study and therefore compared it with another subjective corpus in English (*SFU Review Corpus*), and two objective corpora in both English and Basque (same subject as in the subjective corpus). We studied the following aspects:

- Presence of the first person. In English with personal pronouns and in Basque with verbs, we were able to measure the extent to which the first person was present. We studied this because people usually talk about their experiences in the first person singular or plural.

- Presence of adjectives. In all the grammatical categories, we measured the weight of adjectives in both the objective and subjective corpora. We analyzed the presence of adjectives because they are the most widely used grammatical category when expressing feelings.

- Negation marks. We wanted to study negation marks in Basque because they can influence combinations of feelings in words or phrases. We measured how many negation marks appeared in the Basque Opinion Corpus. In this case too, we were able to do this using computer-based resources.

- Finally, we tagged the Basque Opinion Corpus for discourse and subjectivity information. First, out of 240 texts, 70 were tagged using *Rhetorical Structure Theory* (RST).

	A1	**A2**	**Annotated texts in total**	**Double annotation**
Movies	30	9	30	9
Weather	15	5	15	5
Literature	5	25	25	5
Total	**50**	**39**	**70**	**19**

Table 2.1 – The number of opinion texts annotated by two annotators (A1 and A2).

As we can see in Table 2.1, overall 70 texts (29.16% of the corpus) were tagged and of these 19 (27.14% of those tagged) were tagged by two annotators. To complete the annotation, two taggers had to follow the criteria of Das and Taboada (2018). In different fields, there were certain differences concerning the tagging process. For example, annotators had to spend approximately 20 minutes on weather-related texts, whilst literary texts required around one hour to tag. The results of correspondences between taggers can be seen in Table 2.2. More precisely, we measured whether they correctly tagged the type of discourse relations.

Domain	**Agreement (%)**
Weather	43.59 (17/39)
Literature	41.67 (70/168)
Movies	37.73 (83/220)
Total	**39.81 (170/427)**

Table 2.2 – Inter-annotator agreement in the annotation of texts using *RST* approach (Mann and Thompson, 1988).

As the results in Table 2.2 show, when tagging the type of discourse relation the correspondence between two annotators is 39.81%. There were variations from one field to another. For weather, there was 43.59% agreement, whereas, with regard to cinema, this agreement dropped to 37.73%. For literature, the agreement is 37.73%.

	Components		**Attachment**		**Nuclearity**		**Relation**	
Domain	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**
Weather	20/37	0.54	9/37	0.24	22/37	0.59	15/37	0.41
Literature	84/155	0.54	67/155	0,43	105/155	0.68	48/155	0.31
Movies	112/221	0.56	88/221	0.40	147/221	0.67	68/221	0.31
Total	**216/413**	**0.52**	**164/413**	**0.40**	**274/413**	**0.66**	**131/413**	**0.32**

Table 2.3 – Qualitative evaluation of automatic tools for inter-annotator agreement.

We also used a qualitative evaluation. As Iruskieta et al. (2015) refers, this type of evaluation is called qualitative because it allows comparisons of discourse structures developed in different languages and/or by different people. How they have been evaluated so far have been quantitative, but the difference is that EDUs take into account textual parts, nuclearity and discourse relations. tool to measure the level of agreement between annotators.

The results are provided in Table 2.3. Unlike manual measurements, in this case, in addition to the type of discourse relation, certain other aspects were also measured. Regarding the type of discourse relation, there was 0.31 of correspondence which is 0.08 lower in comparison with the manual evaluation. Behind this difference, unlike with manual evaluations, is the fact that the tool takes into account the position of central subconstituents. Amongst other aspects that were evaluated, there was greater correspondence. In the case of connectors, there was 0.40 correspondence and in component and nuclearity aspects, correspondence was above 0.50, 0.52 and 0.66, respectively.

The corpus itself was tagged based on subjectivity. More precisely, certain semantic orientation components in texts and discourse structures were tagged.

Two annotators had to tag three rhetorical relation components: *i)* the rhetorical relation itself, *ii)* its nucleus or nuclei (in the case of CONTRAST's relation) and *iii)* its satellite. There were three tags to be assigned: *i)* positive semantic orientation, *ii)* negative semantic orientation and *iii)* neutral semantic orientation. Overall, the two annotators annotated the rhetoric relation of one annotator and 40% of its components.

To achieve this, certain guiding principles agreed upon. These describe the tasks of each annotator, including the most difficult, such as metaphoric phrases and how they will be annotated. As can be seen from the difference in tagging results between two different annotators, the Cohen kappa coefficient was 0.58. Therefore, there is *moderate agreement*. As shown by Table 2.4, the largest non-correspondence between two annotators related to neutral semantic orientation.

		A2			
		NEG	**NEU**	**POS**	**Total**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Total	92	120	181	393

Table 2.4 – Contingency table of two annotators with respect to the semantic orientation of rhetorical relations.

Instead of considering agreement between two annotators in terms of correct semantic orientation, the annotators assigned a sentiment value for

each rhetorical relation and EDU.

2.1.2 Creation and evaluation of the sentiment lexicon in Basque

In sentiment lexicon creation in Basque, we first studied the existing conditions. We can distinguish three aspects:

- Time. Starting with sentiment lexicon, we aimed to work on certain features of sentiment analysis. This means that our time needed to be divided between certain parts of sentiment lexicon creation and sentiment analysis. Consequently, we had limited time to create the Basque lexicon.

- Tools and resources. Another limitation we faced for lexicon creation was related to the features available with Basque tools and resources. Certain sentiment lexicons were created using different approximations (Chen and Skiena 2014, Cruz et al. 2014, Barnes et al., 2018, Saralegi et al. 2013 and San Vicente and Saralegi, 2016) but the way sentiment was assigned to words is not. Sometimes, the value assigned to a word is not on a scale and, therefore, we cannot measure the intensity of change that can occur in a word due to valence shifters. Other times, however, in the lexicon, values for word sentiment are on two scales (positive and negative), and due to the polysemy of these words, finally, there are cases when the scale itself is not appropriate. This being the case, we decided to translate the Basque sentiment lexicon.

- Quality. We aim to create a high-quality sentiment lexicon. Moreover, we wanted to create an up-to-date lexicon which can be improved in the future. Consequently, we wanted to create a sentiment lexicon with features similar to those of the the SO-CAL tool.

In the next stage, we decided to create the sentiment lexicon. Following the SO-CAL tool's specific lexicon method for Spanish, we decided to translate it. We also decided to use SO-CAL's English lexicon in the translation process. We saw the following advantages when taking our decision:

- Features of SO-CAL lexicons. To take into account different linguistic phenomena, the values of lexicon words and values used to express

feeling sare between 5 and +5. In our opinion, both the values and value scales are appropriate for the aspects we want to study in our sentiment analysis, especially, since value changes can be measured on that scale.

- Translation resources. As previously mentioned, finding resources to create the Basque sentiment lexicon is difficult and those which do exist are not suitable. However, in Basque lexicography, there are many resources available, like the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* dictionaries (Sarasola, 2005), available both online.

- Choice to compare and evaluate. The sentiment lexicon we shall develop uses the same features as certain other sentiment lexicons and it can be used to compare the results. This also provides an opportunity to evaluate the lexicon.

Next, we collected resources and tools to develop the sentiment lexicon. Overall, we used five resources or tools:

- The Spanish version of the SO-CAL lexicon.

- Two bilingual dictionaries: *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005).

- The English version of the SO-CAL lexicon. The English version of the lexicon contains 6,610 words in five grammatical categories: nouns, adjectives, verbs, adverbs, and intensifiers. We looked at whether the entries of the first Basque version appeared in this dictionary and also what type of sentiment value they have. Then, we decided to assign the word a sentiment value or assign it the same value as the English version.

- The Basque Opinion Corpus.

- *Key Word In Context* (KWIC) technique. The Basque word translated from Spanish was used to search for the word in the corpus and identify the word's context. This way, we assigned a sentiment value attached to the field of the word translated from Spanish into Basque.

Instead of collecting resources and tools, we translated the sentiment lexicon and created them. There are three stages in the translation process: *i)* translate the lexicon from Spanish into Basque, *ii)* clean up the Basque lexicon and *iii)* evaluate the Basque lexicon. We will carry out the translation process based on the examples in Table 2.5.

Phenomenon	SPA	SPA in group	EUS	ENG	The last value
F1	desacreditar “discredit”	desacreditar −2 “discredit”	ospea_kendu −2 “discredit” izena_kendu −2 “discredit” sona_kendu −2 “discredit”	-	-
F2	atrofiar “atrophy”	atrofiar −1 “atrophy”	atrofiatu −1 “atrophy”	-	-
F3	amago ‘feint”	amago −1 ‘feint” cicatriz −2 “scar”	seinale −1 “signal”	-	-
F4	franquismo “Francoism”	franquismo −2 “Francoism”	frankismo −2 “Francoism”	-	−2
F5	correcto “correct”	acertado +3 “right” correcto +3 “correct” decente −2 “decent”	zuzen +3 “right”	right +1 correct +3	+3

Table 2.5 – Example of five phenomena related to the translation process

1- Translation. In this major stage, the SO-CAL tool’s Spanish lexicon was translated.

   i) Automatic translation from Spanish to Basque. First, we translated the Spanish lexicon automatically into Basque using the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries. If a Spanish word could be translated into Basque in several ways, we took into consideration all of the possibilities. This way, the value of each word in Spanish was inherited by each of the translation possibilities.

   In Table 2.5, the word in the first column was translated into Basque (third column).

   ii) Filtering and grouping. In this stage, we filtered and grouped the translations obtained in Basque, that is to say, when translating

the Spanish lexicon certain Basque translations appeared as repetitions and we grouped these. As a result of this grouping, we collected both the translation and the value of the original word.

In Table 2.5, in the third column, phenomenon F3, *amago* ("feint") and *cicatriz* ("scar") in Spanish can both be translated as *seinale* ("signal") in Basque. Likewise, for phenomenon F5, the original Spanish *acertado* ("correct"), *correcto* ("correct") and *decente* ("decent") are all translated as *zuzen* ("correct").

iii) Only choosing translated words which are *Elhuyar* (Zerbitzuak, 2013) or *Zehazki* (Sarasola, 2005) dictionary entries. We studied the translations one by one to see whether they were entries in the two dictionaries. When translating and creating the lexicon, we only took into consideration those translations which were dictionary entries.

In Table 2.5, phenomenon F1, *sona_kendu* ("discredit"), *izena_kendu* ("discredit") and *ospea_kendu* ("discredit") are not dictionary entries and therefore we did not take them into consideration. However, *atrofiatu* ("atrophy") is a dictionary entry and so was taken into account.

iv) Selecting sentiment value. After removing translations which are not dictionary entries, in the remaining translations, we chose the Basque translations which matched Spanish sentiment values and meanings, whenever the Spanish word could be translated into Basque in several different ways. When the Basque translation had a single meaning in Spanish, we chose that meaning and corresponding sentiment value.

When choosing the Spanish meaning and corresponding Basque translation, we followed the procedure below:

* If the translated Basque word had a single translation (and value) in the Spanish original, we use the corresponding translation and value. This is what happened in F2 and F4.
* If the Basque translation corresponds to several words and values in Spanish, the sentiment value assigned to the word (and Spanish meaning) is based on the Basque Opinion Corpus. This was the case with F3 and F5.
* There were certain cases where the translation obtained did

not appear in the corpus but there were several and original words with sentiment valence in Spanish. In this situation, priority was given to the most widely or commonly used translation of the word.

2- Cleaning up. In this second stage, we cleaned up the initial draft version of the sentiment lexicon. More precisely, we adapted the lexicon to certain specific fields and using the SO-CAL's English lexicon we enriched and corrected it.

   v) Adapting the field and corpus. Intending to create the second version, we used the frequency of lexicon words in the Basque Opinion Corpus.
   In Table 2.5, phenomenon F2, *atrofiatu* ("atrophy") is a concept in health and since that field is not present in our corpus, we removed the concept from the lexicon.

   vi) Re-examining and improving each of the lexicon's entries. We translated each of the lexicon's words from Basque into English using the *Elhuyar* dictionary (Zerbitzuak, 2013), then we looked whether these words appeared in SO-CAL's English version of the lexicon. If the Basque entry appeared in the English lexicon, we gave the Basque entry the same value as the English entry. This means we prioritize the value of SO-CAL's English version over the Spanish version. However, if the Basque entry does not appear in the English version, we either removed or maintained this entry according to how appropriate the word was.
   In Table 2.5, phenomenon F5, the Basque word *zuzen* ("correct") corresponds in SO-CAL's English lexicon to *correct* and *right* with different sentiment values.

3- Evaluating the lexicon. In the final phase, the Basque sentiment lexicon was evaluated. For this, firstly, based on two annotators' annotations a gold standard was created and the results obtained by the gold standard and lexicon were compared.

   vii) Annotation of the corpus' most frequently used words' gold standard. Using *Analhitza* (Otegi et al., 2017), the Basque Opinion Corpus' 400 most frequent words were removed (100 words for each grammatical category) and two annotators as-

signed a sentiment value between 5 and +5. Next, based on both annotations, the gold standard was created.

viii) In the second version of the lexicon created, the sentiment lexicon was assigned to the 400 words used in the gold standard.

ix) The values assigned by the gold standard and lexicon were compared using the Pearson correlation. Using this correlation, the correspondence was measured in two ways:

- Pearson 1. In this measurement, in the list of 400 words, only words tagged by two annotators were taken into account. If a word is only tagged by one person it was not taken into consideration.
- Pearson 2. In this measurement, 400 words from the list were used. If a word is not tagged by one or two annotators, it was assigned the value 0, so that the word can be taken into account.

2.2 Identification of valence shifters

After developing the basic guidelines and tools for carrying out the research, we began to identify contextual valence shifters in Basque. In a step-by-step study, we examined different types of valence shifters on different language levels and their impact on words with opinion.

2.2.1 Phonology and morphology level: expressive palatalization and affixes

Firstly, we started to identify the valence shifters of phonology and morphology-level. With this aim, in the first step, we searched for several bibliographic sources that work on the morphology and phonology of the Basque language and we collected a list of language affixes. In this list, we included an example of the grammatical category of the affixes and their semantic meaning. Among the bibliographic sources we used to complete Table 2.6, we can mention (Urdangarin, 1982), (Euskara Institutua, EHU, 2011) and (Oñederra 1990).

Mopheme	Noun	Adjective	Verb	Semantic classification	Example
-zale	X			Liking (of)	Ardozale "liking of wine"
ez-	X	X	X	Negation	Ezberdin ("different")
[z] → [x]				Closeness/ smallness	Gazte ("young") → gaxte ("young") Zahar ("old") → xahar ("old")

Table 2.6 – Morphemes with their characteristics.

Table 2.6 shows examples of expressive palatalization and affixes collected from bibliographic sources. The first letter is *-zale* ("fan of"), it appears with names and it means liking. The second is a prefix *ez-* ("no"); it may appear with nouns, adjectives or verbs and denotes negation.

Finally, in the last example there is an expressive palatalization: [z] becomes in [x]. Like all expressive palatalizations, there is no grammatical restriction and it is used to indicate closeness or smallness. We followed the same procedure with all other affixes.

Word	Number of instances
zati ("part")	11
zertxobait ("a little")	11
zerua ("sky")	9

Table 2.7 – Examples of words extracted based on number of instances.

In the next step, we took 192 opinion texts from the Basque Opinion Corpus and we analyzed them using the *Analhitza* tool (Otegi et al., 2017). After that, we took all the words from the text listed by frequency. Table 2.7 shows the word list provided by the *Analhitza* tool (Otegi et al., 2017), and is based on frequency. As can be seen, the word *zertxobait* ("slightly") that contains the suffix *-txo* ("little") appears in the corpus eleven times.

We then looked at what affixes (prefixes and suffixes) and what expressive palatalizations of the Basque language in the corpus. To do this, we used the list we created in the first step and, based on that, we collected a list of all how the affixes and expressive palatalizations appeared in the corpus. In the list, we individually described their characteristics as shown in Table 2.8.

In Table 2.8, in Example (1), there are no affixes or expressive palatalizations. In Example (2), the suffix *-txo* ("small") is in the word and it weakens the sentiment valence of the word "attack" (-2). Finally, in Example (3), because of the expressive palatalization, [t] becomes in [tt], and as this change

	Word	Morpheme	Valence	Effect on valence	Semantic	Noun	Adj.	Verb
(1)	istorio ("story")							
(2)	erasotxo ("a little attack")	-txo	−2	Weaken	Smallness	X	X	
(3)	pattal ("ill")	[t] → [tt]	−2	Strengthen	Proximity			

Table 2.8 – A sample of morphemes and expressive palatalization in the corpus.

has reinforced the expressivity, the sentiment valence (−2) becomes even stronger and negative. In total, we found for 59 suffixes, 7 prefixes and 13 expressive palatalization instances in the Basque Opinion Corpus, and with all of them, we followed the same procedure.

2.2.2 Syntactic level: negation marks

Another aspect of language which can influence the value of words is syntax. However, in this case, changing the value will not affect a word, but rather a sentence or phrase which consists of a group of words. Just as morphology or phonology valence shifters influence a word, syntactic valence shifters may influence, in addition to words, phrases or sentences.

With the aim of studying how negation marks change the sentiment value of phrases, firstly, we collected a corpus with sentences which have negative marks. However, before collecting the corpus, we created a list of Basque negation marks based on Altuna et al. (2017) and Altuna et al. (1985). These are the negation marks we used for our research: *ez* ("no"), *ezin* ("to be unable to"), *gabe* ("without"), *ezik* ("except"), *salbu* ("except"), *ezta* ("not only") and *ezean* ("if not").

In the next stage, in 96 corpus texts, we collected sentences with at least one negation mark. In this part of our research, we used 96 out of 192 texts which were not part of the test. We did not use all the texts because we obtained most negation marks which appeared in (Altuna et al., 2017) and because despite analyzing more texts there were no new negation marks. Overall, we obtained 359 instances of negation marks. According to data provided by *Analhitza* (Otegi et al., 2017), these 359 negation marks were distributed amongst 320 sentences. Consequently, there are sentences with more than one negation mark. The sub-corpus of negation marks we obtained 5,515 words.

After that, we assigned a sentiment value to both words and full sentences

in these 320 cases. To achieve this, we added words to the *Eustagger* tool (Alegria et al., 2002) used to assign grammatical categories and/or lexical marks from the *Sentitegi* sentiment lexicon (Alkorta et al., 2018); in the lexicon, the words are distributed according to grammatical category.

(5) Pogostkinak ezin hobeki$_{+2}$ atera zituen. (MUS20)
Pogostkina pulled them off perfectly$_{+2}$.

(6) Irabazi$_{+2}$ ezinik jarraitzen du Eibarrek. (KIR17)
Eibar continues without winning$_{+2}$.

(7) Ikuspuntu politikotik$_{-1}$ ez ezik, ekonomikotik$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)
Not only from a political$_{-1}$ point of view but also economically$_{+3}$, Greece has brought hope to other southern European countries, including the Basque Country.

In Examples (5), (6) and (7), the tagger and the sentiment lexicon added to it assigned a value to four words: *hobeki* "better" (adjective), *irabazi* "win" (verb), *politiko* "politic" (adjective) and *ekonomiko* "economic" (adjective). The first two words were assigned value +2 and the adjectives 1 and 3, respectively. This way, with the negation marks and words with a sentiment value, we prepared the context for a linguistic analysis at the next stage.

When carrying out the linguistic analysis of the negation marks, what we took into consideration was the shift in orientation caused by the negation marks in the semantics and, more precisely, in the sentiment value. In other words, we wanted to study the consequences of a negation mark in the case of words or groups of words with values, its range and, consequently, in phrase values too. The analysis was done manually and subsequently, we followed the described procedure.

In the analysis, we looked at the corpus' sentences manually one by one. After identifying the sentences' negation marks, we stated their range. Finally, taking into consideration the consequence on both the field of scope's sentiment value and the full sentence, we grouped the sentence itself and its negation marks.

(8) Pogostkinak [ezin hobeki$_{+2}$] atera zituen. (MUS20)
Pogostkina pulled them off [perfectly$_{+2}$].

(9) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)
Eibar continues [without winning$_{+2}$].

(10) [Ikuspuntu politikotik$_{-1}$ ez ezik], [ekonomikotik$_{+3}$] ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)
.[Not only from a political$_{-1}$ point of view], but also [economically$_{+3}$], Greece has brought hope to other southern European countries, including the Basque Country.

In Example (8), the negation mark is *ezin* ("can not") and its range *hobeki* ("better") goes as far as the adjective. Since in the range of action there is only one word and it has a value, the range of action value is +2, as well as for the sentence. In this case, analyzing the influence of the negation mark, we realized that the negation mark *hobeki* ("better") consolidates the word with value, in fact, in intensity *ezin hobeki* ("perfectly") is stronger than *hobeki* ("better"). Therefore, we ranked the sentence itself and the negation mark's value in the group which they consolidate.

However, in Example (9), although it uses the same negation mark (*ezin* "can not"), we see its influence is different, for example, the range of the negation mark *ezin* ("can not") is *irabazi* ("win"), with value +2 and this weakens it. In other words, *irabazi ezinik* ("without winning") is weaker than *irabazi* ("win") in terms of sentiment value. Therefore, the negation mark and the sentence itself is part of a larger group, because it has been weakened. As can be seen with these two examples, it is possible for the same negation mark itself to be in two different functions.

Finally, in Example (10), the negation mark *ez ezik* ("not only"), with a sentiment value of −1, from a "political point of view" negates the word group, but from the perspective of sentiment analysis, no change occurs in the sentiment value of the group of words. In fact, in this case, after the negation, it is used to add information and not to negate the negation's range. Therefore, the sentiment value of words that are part of the structure is not influenced by the structure *ez ezik* ("not only") in its range.

Number	**Negation mark**	**Structure of the rule**
1	ezin	PM *ezin* + [adjective/adverb] (+ comparative suffix) PM
5	ezin	PM [(NP) + verb + *ezin*] PM PM [*ezin* (+ auxiliar verb) (+ NP) + verb (+ NP)] PM
10	ez	PM [NP] *ez ezik* PM

Table 2.9 – Proposed rules for identifying negation marks that have different effects on sentiment.

Although we wanted to add information about negation in tools like SO-CAL which are based on rules, it is necessary to identify negation-marks and their range. We, therefore, created and evaluated identification rules. We created and evaluated them in the *Constraint Grammar* approach (Karlsson et al., 2011).

Nevertheless, to identify negation marks and their range before creating rules, in the stage before this, we organized in a structured manner the information collected in the linguistic analysis, as can be seen in Table 2.9. When completing this table, we used (Altuna et al., 1985) as a reference, since it describes the position of negation marks in sentences.

One of the items collected in the fact that the structure for each rule is syntactic. The structure of rules in Table 2.9 has the following features: square brackets [] show the negation mark's range, brackets () show that the phrase is optional. Italics show the negation mark's or lexicalized structure and forward slash / indicates that different grammatical categories can appear in the scope of negation.

Other items that appear in the rules include groups of words and grammatical categories - NP and VP show noun- and verb-phrase, respectively - and adjectives, adverbs, verbs, auxiliary verbs, and phrases.

Finally, there is another item that appears in the rules: restricting punctuation marks (PM). This restriction in rules aims to place the range of the negation mark within the sentence and not use items from other phrases. The range can appear both before and behind the negation mark, and since we saw the risk of treating the range as a word we decided to add restricting punctuation marks to the rules. In the case of lexicalized structures, there is no such restriction. Since lexicalized structures have a stable structure (since they are groups of words which are repeated) this is not necessary.

In Table 2.9, after describing what type of features who has, we shall explain how those rules are created.

(11) (...) [ezin da baztertu$_{-1}$ ekaitz zaparradaren bat izatea.]
.[It can not be ruled out$_{-1}$ a storm shower]. (EGU35)

(12) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)
Eibar continues [without winning$_{+2}$]

Through Examples (11) and (12), we shall look at Table 2.9's fifth rule's two variants. In Example (11), after *ezin* ("can not") being a negation mark, the auxiliary verb *da* ("is") and main verb *baztertu* ("rule out") appear, and

finally, a subject which is a noun phrase *ekaitz zaparradaren bat izatea* ("being a storm shower"). Therefore, the structure obtained for this phrase would be: *ezin* "can not" + auxiliary verb + main verb + NP. All the phrase appears in brackets and this shows the range of the negation mark. If we study Example (12), we will see the main verb *irabazi* ("win") appears before the negation mark *ezin* ("can not") and this will be the negation mark's range. After following the same procedure with 36 other instances of the negation mark *ezin* ("can not") in the corpus, we obtained the rules which appear in Examples (13) and (14).

(13) PM [(NP) + verb + *ezin*] PM

(14) PM [*ezin* (+ auxiliar verb) (+ NP) + verb (+ NP)] PM

As Examples (13) and (14) show, the main verb always appears before or after the negation mark *ezin*. In the case of Example (13), the appearance of the noun phrase is optional as well as in Example (14), where the same thing happens. The noun phrase can appear before or after the main verb (for example, *ezin partidua irabazi "can not win the match"* or *ezin irabazi partidua* "can not win the match"). The auxiliary verb can also appear after the negation mark.

Furthermore, to finish, we added the restriction of punctuation marks to the rule, so as not to use the part of the phrase before or after the negation mark. How the *Constraint Grammar* (Karlsson et al., 2011) adaptation rule was used to identify the negation mark, its range and lexicalized structure can be seen in Figure 2.1 below.

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT PUNT_GALD PUNT_ESKL
    PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = "ez";
LIST BESTERIK = "beste";

# (2)
# besterik ez egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)   (1C EZ);

(...)
```

Figure 2.1 – Examples of rules for identifying negation marks and their range in *Constraint Grammar* context (CG3).

Figure 2.1 shows how we adapted to the context of *Constraint Grammar* (Karlsson et al., 2011) :

- LIST. Here certain words or other items, for instance, punctuation mark, are listed. In Figure 2.1, punctuation signs (LIST PUNTUAZIOA) and negation marks (LIST EZ, LIST BESTERIK) are listed.
- MAP command (reflection rules). Using these rules, the grammar inspects the specified structures in the corpus and the structures in the corpus are tagged. The rules consist of the following:
  - Tag assigned. In Figure 2.1, the tag has the ! sign at the beginning and it is assigned to the target word.
  - The target word. In Figure 2.1, the target of the rule is a determinant (DET) (*beste*, "other"). In order to tag the target word, this target word must fulfill the rule's condition.
  - The rule's conditions. In order to assign the tag *!besterikezHAS* to a determinant, which is in position 0, the negation mark *ez* ("not") must be after this determinant (in position 1).

When carrying out the evaluation, we used 48 texts from the Basque Opinion Corpus. These 48 texts had to be processed to adapt them to the context of the *Constraint Grammar* (Karlsson et al., 2011) and to do this we used the Basque morpho-syntactic disambiguator *Constraint Grammar*.

```
"<,>"<PUNT_KOMA>"
    PUNT_KOMA
"<batere>"
     "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG \%SINT
"<harrotu>"   S:278/0
     "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG \%
         ADIKAT S:278 !gabeAUR
"<gabe>"   S:141/0
     "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> \%SINT S:141 !gabe
"<\$.>"<PUNT\_PUNT>"
    PUNT\_PUNT
```

Figure 2.2 – The corpus tagged with *Constraint Grammar*'s rules (Karlsson et al., 2011) .

Each rule in the *Constraint Grammar* (Karlsson et al., 2011) leaves its tag on in every word that fulfills the conditions. In Figure 2.2, for example, the

rules we created in the *Constraint Grammar* (Karlsson et al., 2011) tagged two words (*!gabeAUR* and *!gabe*).

After applying the rules to 144 corpus test sections, these rules were evaluated. Firstly, the correspondence between two annotators to carry out the evaluation was measured using a small part of the corpus and, then, one person did the entire evaluation.

When evaluating, the annotators had to follow the procedure below:

1- The annotator used three tags when evaluating: ETIK_ONDO, when the rule's tag is correct; ETIK_FALTA, when the word in the corpus needed a tag and when the rule does not assign any and, finally, ETIK_GAIZKI, when the rule assigns a tag to the corpus' word, but the word does not need a tag.

2- In the corpus, the annotator studied each word one by one and as the evaluation progressed gave the words a tag. If the word was tagged correctly, the annotator assigned ETIK_ONDO (correct tag). However, if the tag was missing, the annotator assigned the ETIK_FALTA (missing tag) status. Finally, if the word had an incorrect tag (false positive), then it was marked as ETIK_GAIZKI (incorrect tag).

```
"<,>"<PUNT_KOMA>"
    PUNT_KOMA
ETIK\_FALTA "<batere>"
    "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG \%SINT
ETIK\_ONDO "<harrotu>"   S:278/0
    "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG \%
        ADIKAT S:278 !gabeAUR
ETIK\_ONDO "<gabe>"   S:141/0
    "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> \%SINT S:141 !gabe
"<\$.>"<PUNT\_PUNT>"
    PUNT\_PUNT
```

Figure 2.3 – The tags assigned by annotators to words.

Figure 2.3 presents an example of an evaluation. There are three words and all three should be tagged. Two out of three (*harrotu* "become arrogant" and *gabe* "without") are tagged as ETIK_ONDO, at the beginning of lines. However, this is not the case for the word *batere* ("not"). Although it is part of the range of the negation mark, because it influences the verb *harrotu* ("become arrogant"), it is not tagged.

Consequently, it is tagged as ETIK_FALTA, in this case too, at the beginning of the line. The aforementioned procedure was followed for agreement between two annotators as well as to evaluate the corpus test section.

3- Finally, to measure agreement, F1 score was used. As the results show, the score was 0.60.

2.2.3 Discourse level: rhetorical relations, their components, and central unit

At the discourse level, following the *RST* approach (Mann and Thompson, 1988), we focus on two aspects. On the one hand, we analyzed changes in sentiment valence in rhetorical relations. On the other hand, in the central units, that is, in the most important units of discourse in the text, we studied the relation between grammatical categories and opinion.

Rhetorical relations

In rhetorical relations, we have first defined our aims. In total, we set three goals:

- We want to measure the agreement of semantic orientation between the rhetorical relation and its components.
- Given the distance between rhetorical relations from the central unit, we want to measure the greatest agreement of the semantic orientation between rhetorical relations and the whole text.
- In opinion texts, we want to examine whether different types of rhetorical relations they appear in the same parts of texts. We will establish the position of rhetorical relations based on the distance to the central unit.

After setting our goals in rhetorical relations, we have taken a few steps. The following is a step-by-step methodology:

1- Assign semantic orientation to different components of the discourse level:

   – 240 opinion texts.

– The following rhetorical relations: EVALUATION/INTERPRETATION (140 instances), CAUSE subgroup (71), CONCESSION/ANTITHESIS (70), EVIDENCE/JUSTIFICATION (53), CONTRAST (26), CONDITION subgroup (18), ENABLEMENT/MOTIVATION (6). In total, 384 instances were assigned the semantic orientation.

– The components of the above rhetorical relations: semantic orientation has been assigned to the nucleus and the satellites as well as to the first and last EDU of the relations.

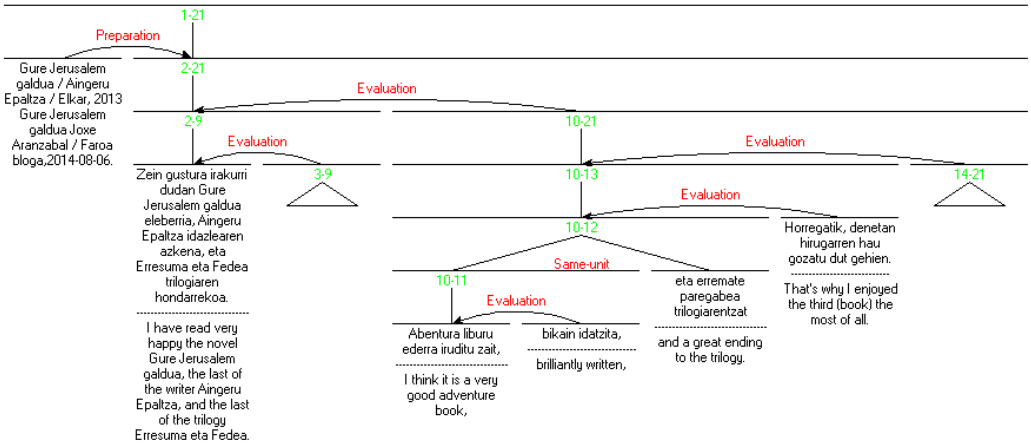


Figure 2.4 – Discourse-tree of the text (SENTFAR-01.)

2- Parameter analysis of rhetorical relations. After introducing the rhetorical relation of 28 literature texts in the database, we analyzed the following parameters. That is, in the second step of the methodology, in the manual labeling, we performed the parameter analysis on the 384 instances of the aforementioned rhetorical relations. We will use the EVALUATION relation (the EDUs 10-13) of Figure 2.4 to explain the parameters:

– Nuclearity. A rhetorical relation can be mononuclear or multinuclear. In Figure 2.4, the EVALUATION relation is N(ucleus)-S(atellite).

– Semantic orientation of rhetorical relations and EDUs. We have assigned three types of semantic orientation to rhetorical relations

and EDUs: positive, negative and neutral. First, we assign the semantic orientation to the EDUs and, then, to the whole rhetorical relation. In the EVALUATION of Figure 2.4 (the EDUs 10-12 and 13), the two EDUs and the whole relation have a positive semantic orientation.

- Distance to the central unit. The distance between the rhetorical relation and the central unit has been measured by the number of rhetorical relations between them. In our case, the distance of the EVALUATION relation (the EDUs 10-13) is +2.
- The type of rhetorical relation. The last parameter that we have taken into account is the type of rhetorical relation. The relation we take as an example is EVALUATION type. Consequently, the EDUs 10-12 presents a situation and the EDU 13 makes an evaluative comment about the situation.

3- In terms of semantic orientation, we measured the agreement among the semantic orientation labels. Agreement measurement has been made on rhetorical relations and their constituents. Manual evaluation was performed using F-measure.

Central unit

On the other hand, to find out the most common grammatical categories of the words and to analyze the distribution of words with semantic orientation in the central unit, we have performed the following steps:

1- Extract central units from the corpus. First, a linguist has selected central units from 192 texts of Basque Review Corpus. The test part of the corpus was not used. Before selecting the central units, the texts were segmented following the guidelines of the *RST* approach (Das and Taboada, 2018). On the other hand, we used the *RSTTool* tool (O'Donnell, 2000) for text segmentation.

In Figure 2.5, the text EGU01 can be seen in XML format. Each of these segments is an elementary discourse unit (EDU) and the task of the annotator, in connection with the next step, is to identify which of these EDUs is the central unit, in other words, the most important EDU in opinion texts.

```
<rst>
  <header>
    <relations>
    </relations>
  </header>
  <body>
    <segment id="1"> Gaurko eguraldia. HEGO EKIALDEKO HAIZEA ETA HODEI
        BATZUREKIN EPELDU EGINGO DA. </segment>
    <segment id="2"> Giro ona tokatuko zaigu gaur ere hego haize epelarekin
        .</segment>
    <segment id="3"> Borraska pixka bat hurbiltzeak haizea apur bat mugituko
         du</segment>
    <segment id="4"> baina </segment>
    <segment id="5">tenperatura igoaz, </segment>
    <segment id="6">giroa goxo mantenduko da leku gehienetan.</segment>
    <segment id="7"> Hodei zirrinta mehe batzuk agertuko dira zeru sabaian</
        segment>
    <segment id="8"> baina eguzkia apur bat lausotu arren,</segment>
    <segment id="9"> astro handia bistan edukiko dugu</segment>
    <segment id="10"> eta dotoreziak ez du bat ere galduko.</segment>
    <segment id="11"> Baditeke iluntze aldera ekaitz hodei batzuk garatzea
        eta euri zarrastaren batzuk botatzea agian. </segment>
    <segment id="12">Akats txiki horiek gora behera eguraldi bikaina oro har
        .</segment>
  </body>
</rst>
```

Figure 2.5 – Selection of central unit by two annotators in the EGU01 opinion text.

2- Inter-annotator agreement in central unit selection. To prove that the annotator has chosen the central unit correctly, another person has also selected the central units of these texts. To measure the agreement between two annotators, a total of 78 opinion texts (32.5% of the total corpus) were used. Percentage calculation was used to calculate the inter-annotator agreement.

Agreement	51	0.65
Disagreement	27	0.35
Total	78	1.0

Table 2.10 – Inter-annotator agreement when choosing a central unit in opinion texts.

As shown in Table 2.10, in 51 opinion texts of the 78 texts, two annotators considered the same EDU as a central unit. In the other

44 opinion texts, however, different EDUs have been chosen as central units. Therefore, the agreement in selecting the central unit of 78 opinion text was 0.65.

3- Analysis of the grammatical category of words in central units. In the next step, we analyzed the type of words in each central unit and whether one or more words appear at regular frequency. With this aim, we used the *Analhitza* tool (Otegi et al., 2017). We have analyzed these central units domain-by-domain.

4- Analysis of the results provided by *Analhitza* (Otegi et al., 2017). After that, we analyzed the results provided by the *Analhitza* tool (Otegi et al., 2017). In analyzing the results, we considered the following aspects:

   i) Frequency of words. We analyzed the 30 most frequent words of each grammatical category.

   ii) Grammatical category of words. Another feature of the words considered was their grammatical category. The tool directly classifies words into grammatical categories; which has made our work easier.

   iii) Domain of words. We also examined whether these words with high frequency are domain-related. In all domains, the classification was whether or not words belong to the domain (binary classification), except in weather. In weather, we also used the time concept in the domain classification, since we believe that the time is of particular importance. Therefore, in the weather domain, the word classification was trivial.

   For example, in words with the highest frequency in the weather domain, we classified the word “winter” as belonging to the domain and “appearance” as not belonging to the weather domain; because it is not directly related to this domain. Finally, since the word "weekend" indicates time, we classified it into a weather domain.

   iv) Assignation of sentiment valence to central units and their words. Another feature examined was the sentiment valence of the central units. We have used the Basque version of the SO-CAL tool with *Sentitegi* lexicon (Alkorta et al., 2018), which we created for the assignment of sentiment valence.

2.3 Lexicon-based document level sentiment classifier in Basque

The third major step of this work was to develop a document level sentiment classifier in Basque. With this aim, we wanted to develop the first Basque version of the SO-CAL tool (Taboada et al., 2011). Therefore, we used the *Eustagger* tool (Alegria et al., 2002) and the *SentiTegi* sentiment lexicon (Alkorta et al., 2018).

2.3.1 Integration of the *Eustagger* tool

The English version of SO-CAL contains lemmatized words with sentiment valence and some rules related to inflection. This is not valid for Basque, due to its rich inflection system and agglutinative nature.

To solve this difficulty, in the Basque version of SO-CAL, we have decided to integrate the *Eustagger* tool (Alegria et al., 2002) to lemmatize words of texts before assigning sentiment valence to them. In this way, the tool will first lemmatize the text and, then, check if the lemmatized word is present in the lexicon of the tool and if it is, it will assign a sentiment valence to the word in the text. Therefore, if a lemmatizer is integrated into the tool, the tool will first be able to lemmatize the text; then to check if the word lemmatized is in the lexicon, and finally, if the word is in the lexicon, to assign the sentiment valence to the word.

Figure 2.6 shows the changes made and a comparison of the structures of the SO-CAL tool in English and Basque. As can be seen, in English, there are some morphological rules (such as the deletion of the plural *-s* letter) and in this way, the words in the texts are transformed into lemmas. In Basque, meanwhile, because of morphological richness, the *Eustagger* lemmatizer (Alegria et al., 2002) is integrated to achieve the lemmatization. Then, the sentiment lexicon is applied and if the lemmatized words are in the lexicon, the tool assigns a sentiment valence to them. Finally, other rules are similar for English and Basque (assigning more weight to words with negative semantic orientation, not assigning sentiment valence in interrogative sentences, etc.) and they are useful for both languages.

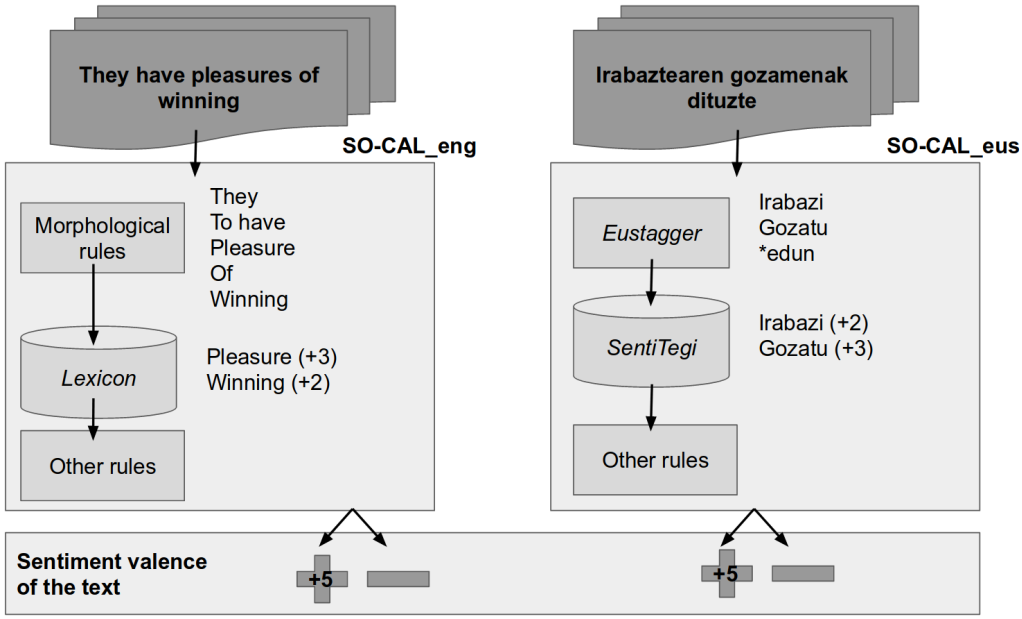


Figure 2.6 – Comparison of the structures of the English and Basque versions of the SO-CAL tool.

2.3.2 Integration of the *Sentitegi* sentiment lexicon

In the Basque version of the SO-CAL tool, the *Sentitegi* sentiment lexicon (Alkorta et al., 2018) has been integrated .

The tool itself contains a module for sentiment lexicons where we added Basque ones. For the tool to work, lexicons must be in *txt* format, with an entry and its valence in each line. Also, each grammatical category must have its file. The *Eustagger* lemmatizer (Alegria et al., 2002) will lemmatize the word in the text and identify its grammatical category, and, then, the tool will search that word in its grammatical category. We have added lexicons for four grammatical categories: noun, adjective, adverb, and verb.

Table 2.11 shows the change made. On the left, there is an English lexicon with a list of words and their sentiment valence. On the right, there is a part of the lexicon in Basque.

English		**Basque**	
Thriving	+3	Bikain	+5
Record-setting	+3	Maximo	+1
Leading	+3	Orokor	+3
Industrious	+3	Min	-2
Best-selling	+3	Polit	+4
Upset	+3	Gogor	-1
Clean	+2	Ahul	-2
Capacious	+2	Behar	-1
Cogent	+2	Txar	-3
Confident	+2	Zail	-2

Table 2.11 – Examples of sentiment lexicons in English and Basque.

2.3.3 Sentiment classifier evaluation

After integrating the *Eustagger* lemmatizer (Alegria et al., 2002) and the *Sentitegi* sentiment lexicon (Alkorta et al., 2018) in the sentiment classifier, we evaluated the classifier itself.

We have used 48 review texts from the Basque Opinion Corpus. As mentioned in this methodology section, the semantic orientation of the corpus opinion texts is annotated and we compared them with the results provided by the sentiment classifier. The tool gives numerical results and when evaluating, we considered the sign in the numerical result.

2.4 Summary

In this section, we describe the methodology of this thesis work. First, we created resources and tools for sentiment analysis. On the one hand, we built a Basque Opinion Corpus, collecting opinion texts from newspapers and specialized websites. Then we developed *SentiTegi* (Alkorta et al., 2018), a sentiment lexicon in Basque. To do this, we have translated the lexicon of the Spanish version of the SO-CAL tool using the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) bilingual dictionaries.

In the second step, we identified the valence shifters and measured their effect on words and phrases. We worked on phonological and morphological valence shifters. We collected information about them from bibliographic

sources and, then, we extracted their instances from a part of the corpus. In the next step, we assigned a sentiment valence to these instances with valence shifters using the *SentiTegi* lexicon (Alkorta et al., 2018). After that, we measured the effect of these instances in words. In other words, we examined if they reinforce, weaken, or do not affect the valence of words and phrases.

We also similarly dealt with negation marks. We collected a list of negation marks from bibliographic sources and searched them in a part of the corpus. Then, we assigned sentiment valence to the sentences with a negation mark and measured the effect of the negation marks. In the next step, using the above analysis, we developed rules for identifying negation marks and their scope and then adapted them to the *Constraint Grammar* (Karlsson et al., 2011) to evaluate them.

At the discourse-level, we obtained different kinds of rhetorical relations from a part of a corpus annotated with the *RST* approach (Mann and Thompson, 1988). Then, using both SO-CAL and *SentiTegi* sentiment lexicon (Alkorta et al., 2018), we assigned sentiment valence to the components of rhetorical relations and rhetorical relations. We also indicated some other features of rhetorical relations. With this information, we studied rhetorical relations from a sentiment analysis perspective. On the other hand, we also worked on the central unit by analyzing the grammatical categories of words that appear there and studying the distribution of words with sentiment valence.

In the next step, we adapted the English SO-CAL tool for sentiment classification into Basque. To this end, we first integrated the *Eustagger* tool (Alegria et al., 2002) for text lemmatization in the SO-CAL tool. Then, we replaced the English sentiment lexicon with the Basque *SentiTegi* lexicon (Alkorta et al., 2018).

RESULTS

3

Resources and tools

3.1 Article 1

Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis

***Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis*,**

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, IXA Group, University of the Basque Country

jon.alkorta@ehu.eus
koldo.gojenola@ehu.eus
mikel.iruskieta@ehu.eus.

Nowadays, it is very usual to read reviews about movies, products or tourist destinations before taking a decision. Reviews, as a particular genre, follow some genre constraints and also a specific discourse structure.

Following Taboada et al. (2016) discourse structure, along with syntax, is necessary to get a better account of sentiment analysis in review corpora. The aim of this paper is to present a corpus we have developed in order to study sentiment analysis in Basque. As far as we know, there is no polarity-balanced corpora for the study of sentiment analysis in Basque.

Corpus design

In order to fulfill this gap, we built a corpus for that purpose following this criteria:

1) Collect texts from specialized review websites (online magazines and newspapers).
   1.a) With a clear negative or positive evaluation.
   1.b) With rich syntactic structures and *opinionative* data.
   1.c) With balanced domains: 20 positive and 20 negative texts with similar word length.
2) Describe corpus information with: code, title, source, polarity and word length.
3) Analyze the corpus using different methods to measure the opinionative phenomena and evaluate its quality. Reliability has been measured comparing some characteristics of our corpus against other corpora.

Corpus

The corpus that we have built is composed of 240 texts in Basque corresponding to 6 domains (books, music, movies, weather, politics and sports). It contains 52,092 tokens and 3,711 sentences.

Regarding size, our corpus can be compared to other corpora built to analyze sentiment analysis: i) Emotiblog (Boldrini et al., 2010), that contains 100 texts for each language (Spanish, Italian and English; ii) the SFU Review Corpus (Taboada, 2008) is made up of 400 texts (8 domains), 289,270 tokens and iii) the Opinion Corpus for Arabic (OCA) (Rushdi-Saleh et al., 2011) has 500 texts and 215,948 tokens.

The quality of our corpus has been measured with respect to the following phenomena:

i) Presence of first person. Sentiments are usually expressed using the first person and, thus, we have measured if the frequency of the first person is different in objective and subjective corpora. A set of texts from Wikipedia has been taken as objective corpus (with the same topic and similar length as our corpus), because Wikipedia asks writers to include neutral or non-opinative texts. A language analyzer for Basque and English (ANALITZA), which is based on a set of tools for the automatic linguistic analysis based on IXA-pipes (Agerri et al., 2014), has been used to obtain the frequency of use of the first person.

With this tool we have measured the frequency of the first person of verbs (in Basque) and the frequency of pronouns (in English). Results demonstrate that the frequency of the first person is different in both corpora. While the presence of first person is about 0.12% (English Wikipedia) and 1.31% (Basque Wikipedia) in objective corpora, in subjective corpora the frequency increases up to 8.37% in our corpus and 11.80% in the SFU Review Corpus (Taboada, 2008). Results also show language differences: the first person is mostly plural in our corpus (5.10% plural and 3.27% singular) while the first person is mostly singular in the SFU Review Corpus (1.70% plural and 10.10% singular).

ii) Adjectives in the corpus. We have measured the frequency of adjectives, because adjectives are one of the most frequently used phenomena in sentiment analysis. However, we found that the frequency is similar (from 8% to 9%) in both languages and four corpora.

iii) Discourse markers. Relational discourse structure can change the polarity of a text, because there are coherence relations which describe the purpose or the conclusion of a text. Because of that, a text span with such relations is more important and, consequently, the polarity of the text span should be taken into account. In our corpus, we have seen some discourse marker signals, that signal *purpose* or *conclusion* discourse relations.

Some discourse markers that signal purpose are: *azken batean* 'in the end' (9), *laburbilduz* 'to sum up' (4), *azkenik* 'finally' (3), *amaitzeko* 'to finish' (3), etc. Moreover, the following discourse marker list signals the conclusion: *beraz* 'thus' (74), *ondorioz* 'consequently' (12), *hortaz* 'so' (4), etc.

The following example of our corpus shows the relevance of this phenomenon:

(1) *(...) aire masa* ***hotz*** *eta* ***ezegonkor*** *bat iritsiko zaigu (..) eguraldia benetan* ***gaiztoa*** *izango dugu. (...). Laburbilduz, etxean geratzeko moduko eguraldia*.

(...) **cold** and **unstable** air mass (...) very **bad** weather. (...). In short, the weather invites us to stay at home.

In Example (1), the first sentence states that the weather will be cold, bad and unstable and, after the underlined discourse marker, the prediction is summarized suggesting *to stay at home*.

Adversative discourse markers change the polarity of the previous text span. There are some adversative discourse markers in our corpus: *baina* 'but' (389), *ordea* 'however' (40), *hala ere* 'nevertheless' (38), *aldiz* 'while' (34), *berriz* 'whereas' (33), *dena den* 'even so' (16), *dena dela* 'anyway' (5), *haatik* 'though' (5), to cite some.

(2) *(...) Jada ezagutzen dugun istorioa, noski. Baina, horrek ez dio freskotasunik kendu (...).*

(...) We already know the story, of course. But this does not remove the freshness (...).

In Example (2), the first sentence may be said to have a negative polarity, but the discourse relation signaled with an adversative discourse marker in the second sentence inverts the polarity (from negative to positive).

iv) Irrealis Blocking. As Taboada et al. (2011) explains, there are some language forms that triggers an *irreal* context. We have found different examples in our corpus: a) conditionals and b) negative polarity items.

a) We found three types of conditionals in our corpus: i) non-hypothetical; ii) hypothetical and iii) unreal.

b) Besides, we have found various negative items for persons (*inor* 'nobody', 13 instances), things (*ezer* 'nothing', 29), mood (8), time (16) and space (5).

iv) Negation. Negation appears in different ways in our corpus: *ez* 'not' (718) modifying clauses; *gabe* 'without' (107) modifying noun phrases and *ezean* 'in the absence of / unless' (4) modifying subordinate clauses.

Conclusion and future work

In this work, we have created a Basque corpus for sentiment analysis and we have made a preliminary analysis of the data. The frequency of first person and discourse markers shows that the corpus is valid to study different opinionative phenomena. Moreover, we observe that the corpus has typical constructions (discourse marker, irrealis blocking and negation) analyzed in sentiment analysis which also suggest that our corpus contains opinionative data. In the future, we will tag this corpus using the frameworks of Rhetorical Structure Theory (RST, Mann & Thompson, 1988) and Appraisal Theory (Martin & White, 2005), to study how relational discourse structure modifies other language levels (semantic and syntactic) of sentiment analysis in Basque, following previous work (Alkorta et al., 2015).

References

Agerri, R., Bermudez, J., & Rigau, G. (2014, May). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC* (Vol. 2014, pp. 3823-3828).

Alkorta J., Gojenola K., Iruskieta M. & Perez A. (2015). Using relational discourse structure information in Basque sentiment analysis. 5th Workshop "RST and Discourse

Studies", in *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015), Alicante (España).*

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2010, July). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 1-10). Association for Computational Linguistics.

Mann, W. C., & 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.

Martin, J. & White, P. (2005) *The Language of Evaluation: Appraisal in English*. London: Palgrave McMillan.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, *62*(10), 2045-2054.

Taboada, M. (2008) The SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, *2*, 325-347.

3.2 Article 2

SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis

ISSN 2007-9737

SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta

IXA NLP Group, University of the Basque Country (UPV/EHU), Vizcaya, Spain

jon.alkorta@ehu.eus, koldo.gojenola@ehu.eus, mikel.iruskieta@ehu.eus

Abstract. The creation of a semantic oriented lexicon of positive and negative words is often the first step to analyze the sentiment of a corpus. Various methods can be employed to create a lexicon: supervised and unsupervised. Until now, methods employed to create Basque polarity lexicons were unsupervised. The aim of this paper is to present the construction and evaluation of the first semantic oriented supervised Basque lexicon ranging from $+5$ to -5. Due to the lack of resources, the Basque lexicon was created translating the SO-CAL Spanish dictionary by means of two bilingual dictionaries following specific criteria and then slightly corrected with the SO-CAL English dictionary and frequency data obtained from the Basque Opinion Corpus. Evaluation results show that the correlation between human annotators is slightly better than between a gold standard lexicon (obtained from human annotation) and the translated dictionary. This shows that the quality of the translated lexicon is satisfactory, although there is a space to improve it.

Keywords. Semantic oriented lexicon, manual translation method, Basque, sentiment analysis.

1 Introduction

Sentiment analysis is a task that classifies documents according to their polarity. This research area has had a big development in the last years due to social networks and Internet, which have increased the quantity of opinions and other types of text with emotion, and is in demand of methods for automatic processing.

There are many resources for sentiment analysis for the most used languages such as English [9], Chinese [15] and Spanish [5].

Additionally, competitions like SemEval [10] have greatly contributed to the development of resources and tools for sentiment analysis. However, the development is not symmetric on lesser used languages or languages in normalization process like Basque.

The semantic oriented lexicons are related to the lexical level and, so, they are useful and important in sentiment analysis. If the semantic orientation of the words is known, opportunities open up to calculate the semantic orientation of sentences and, therefore, the semantic orientation of texts taking into account syntax and discourse constraints.

The creation of the semantic oriented Basque lexicon has been semi-manual translating from the SO-CAL Spanish dictionary, and then enriching it with corpus analysis and the English SO-CAL dictionary. In the translation process, different bilingual dictionaries have been used. We have decided to use a semi-manual procedure to create our lexicon, in order to take into account some idiosyncratic characteristics of Basque language.

The aim of this paper is to present a semantic oriented lexicon for Basque. We will emphasize the process of creating this lexicon, and particularly the solutions adopted to solve the problems encountered.

The main contributions of this work are: i) the creation of a domain-specific semantic oriented Basque lexicon, ii) a description of a semi-manual technique to create the lexicon and iii) a thorough evaluation.

ISSN 2007-9737

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the methodology of the translation process. Then, Section 4 discusses the design decisions, while Section 5 describes the characteristics of the created lexicon in two stages. In Section 6 the quality of the lexicon is evaluated and, finally, Section 7 concludes the paper, also proposing directions for future work.

2 Related Work

There are various approaches for the creation of polarity lexicons, based on knowledge or on automatic methods. Each of the approaches has its advantages and drawbacks.

SO-CAL [14] is a dictionary-based tool to extract sentiment from texts. The dictionary was created manually, where words are annotated with polarity (positive or negative) and strength (semantic orientation: from ± 1 to ± 5). There are two versions of SO-CAL tool. The original version is the English SO-CAL and the Spanish version, the second one, is based on the previous version. The English and Spanish dictionaries (V1.11) contain 6,610 and 4,880 words, respectively.

A disadvantage of manually-created lexicons is the hard-work to make modifications. In contrast, they can be tailored to be domain-specific and, depending on the linguistic information used, they can treat a variety of different linguistic phenomena.

ML-SentiCon [6] is a multilingual polarity lexicon, where the lexicons have been automatically generated from an improved version of Senti-WordNet. It contains a Basque lexicon that contains 4,323 lemmas. The polarity values are situated between -1 and $+1$, in a continuous scale. Additionally, QWN-PPV tool [11] is able to generate multilingual polarity lexicons, including Basque. This unsupervised tool makes use of a corpus and WordNet.

The main disadvantage of these lexicons is that they are not domain-specific, so their results could vary from one domain to another. In contrast, their main advantage lies on the facility to create them.

Another characteristic of previous three works is that the sentiment value of words is in a scale, although the scale dimensions are different. However, there are works in which the sentiment value of words are not in scale. For example, in some works like [13], there are two non-numerical tags: *positive* and *negative*. Consequently, two words with different intensity are expressed with the same tag.

Methods to evaluate lexicons are different depending on each technique. Some works [3] use intrinsic methods where the result of the system is compared to a gold standard data set, predefined by evaluators. In contrast, there are other systems [4] which use extrinsic methods where the system is evaluated in an applied setting. Finally, some works [7] use both extrinsic and intrinsic methods.

The lexicon presented in this work differs from previous ones in several respects. SO-CAL dictionaries have also been manually created but, until now, they have dealt with languages which are not morphologically rich (Spanish and English) in contrast with Basque. Another relevant difference of this study has been the evaluation. We will apply an intrinsic evaluation and measure, using Pearson correlation, the agreement between two human annotators, and the reliability between the gold standard (based on human annotation) and the translated dictionary. Finally, the characteristic of the created lexicon is another interesting aspect. The words of the lexicon have the sentiment value in a scale from -5 to $+5$. This allows us to study how sentiment shifters of different linguistic levels (morphology, syntax and discourse) affect on sentiment analysis.

3 Methodology

In order to create a semantic oriented lexicon for Basque, we have adopted several decisions taking different factors into account:

i) **Time.** The creation of a semantic oriented lexicon for Basque is related to the project of linguistics-based Basque sentiment analysis and, for that reason, the time to create the lexicon is limited.

ii) **Resources.** The Basque language is still in a normalization process and this has some limitations to create corpora and to reuse computational resources. On the one hand, it is difficult to create a large opinion corpus of different topics. This situation could affect to the quality of the lexicon if the corpus is used for that. The collaboration of lexicographers would be ideal but it is a costly resource, not available. This situation adds a difficulty to create a semantic oriented Basque lexicon from zero.

iii) **Quality.** We want to develop the lexicon with the best possible quality (and in the less time possible) and with that aim we will first translate the lexicon, after that evaluate it and then improve our semantic oriented lexicon following an specific criteria.

3.1 Resources for Translation

We have used mainly four resources in the translation process.

i) **The SO-CAL Spanish Dictionary [14].** This dictionary is the source to create the Basque semantic oriented lexicon. It contains 4,880 words of five grammatical categories (noun, adjective, adverb, verb and intensifier).

ii) **Two Bilingual Dictionaries:** Spanish-Basque: Elhuyar dictionary [16] and Zehazki [12]. These dictionaries have been used to translate the Spanish SO-CAL dictionary. Moreover, they have also been used to check if the translated word is an entry of such dictionaries since we will work only with words which are entries of one of these dictionaries. Dealing with collocations and expressions is necessary but it is out of the scope of this work.

iii) **The Basque Opinion Corpus [1].** After getting the first version of the lexicon, each entry has been checked in the corpus to create a domain-based lexicon. The corpus contains 240 texts of six different domains.

iv) **The SO-CAL English dictionary [14].** This version which contains 6,610 words has been used to verify and enrich the already created domain-based lexicon.

Taking all the factors explained above into account and using the mentioned resources, we have decided to translate the SO-CAL Spanish dictionary to create the Basque SO-lexicon *Sentitegi*, following the methodology explained in Figure 1.

3.2 Translation Steps

Figure 1 shows the steps followed in the translation process. To begin with, a first version of a semantic oriented Basque lexicon has been created from the Spanish version of the SO-CAL dictionary. After that, the second version has been created enriching it with the English lexicon version (V1.11) and limiting it to the domains of Basque Opinion Corpus.

Some interesting phenomena have been detected in the translation process of SO-CAL dictionaries from Spanish and English versions (V1.11) to Basque. Table 1 shows these five phenomena.

- Phenomenon 1 (P1): the Spanish word is translated but the translation is not an entry of Elhuyar [16] and Zehazki [12] dictionaries, so we do not take it into account.
- Phenomenon 2 (P2): The Spanish word is translated, it is an entry of Elhuyar but the translation does not appear in the Basque Opinion Corpus. Consequently, it will appear in the first version (V1.0) but not in the second one (V2.O).
- Phenomenon 3 (P3): The Spanish word is translated, it is an entry, it appears in the corpus but it is not in the SO-CAL English dictionary. So, it will appear in the first version of the dictionary, but not in the second one.
- Phenomenon 4 (P4): The Spanish word is translated, it is an entry, it appears in the corpus and it is not present in the SO-CAL English dictionary. Then, it will be included in the (first and) second version.

ISSN 2007-9737

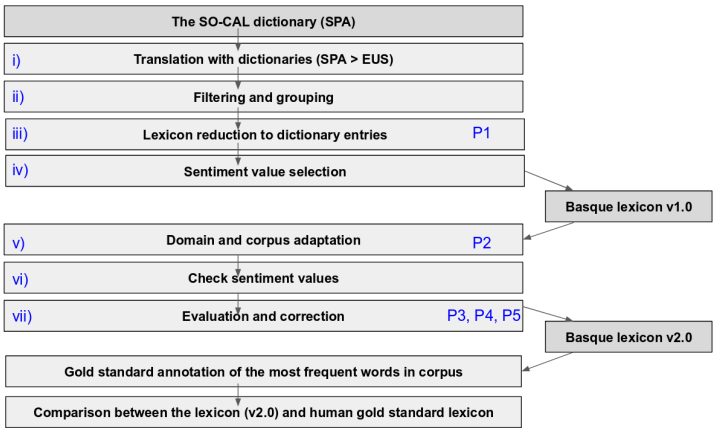


Fig. 1. Steps of the translation process. The enumeration in blue on the left indicates methodological steps. The blue code on the right (P1 to P5) indicates different phenomena in the translation process

– Phenomenon 5 (P5): The Spanish word is translated, it is an entry, it appears in the corpus and it also a word of the SO-CAL English dictionary. It will appear in the first and second versions. These last two phenomena are the same but the decision is different that depends on the characteristic of each word.

The translation process has been the following (see Figure 1):

i) **Automatic translation from Spanish into Basque.** The Spanish sentiment dictionary of SO-CAL has been translated using Elhuyar [16] and Zehazki [12] dictionaries. When one word of the dictionary has more than one entry, all the entries have been taken into account. The sentiment value of the Spanish word has been assigned to all the correlated elements in Basque.

   For example, the Spanish word *desacreditar* -2 "discredit" has been translated into Basque in different forms: *izena_kendu*, *ospea_kendu* and *sona_kendu* "discredit" with the same meaning. This example shows how one Spanish word could be translated in different forms to Basque. But these translations are not entries of the dictionary. Consequently, they have not been taken into account.

ii) **Filtering and grouping.** After translating all the words and transferring their sentiment values, the repeated words in Basque have been filtered and grouped.

   Table 1 shows how words in Basque (fourth column) can have one or more translations in Spanish (third column). The phenomena numbered 1, 2 and 4 have one translated word in Spanish whereas 3 and 5 have more than one.

   This phenomenon occurred because those words are polysemic. There are cases where two or more words in Spanish correspond to the same word in Basque and vice versa. Consequently, in some cases, each word in

Table 1. Words that belongs to five phenomena related to translation process

Phenomenon	SPA	SPA grouping	EUS	ENG	Value
P1	desacreditar “discredit”	desacreditar -2 “discredit”	ospea_kendu -2 “discredit”	-	-
P2	atrofiar “atrophy”	atrofiar -1 “atrophy”	atrofiatu -1 “atrophy”	-	-
P3	amago “feint”	amago “feint” -1 cicatriz “scar” -2	seinale “signal” -1	-	-
P4	franquismo “Francoism”	franquismo -2 “francoism”	frankismo -2 “francoism”	-	-2
P5	correcto “correct”	acertado “correct” +3 correcto “correct” +3 decente “decent” -2	zuzen +3 “correct”	right +1 correct +3	+3

Basque has several meanings and sentiment values in Spanish.

iii) **Dictionary entry: Check if the Basque translation is an entry in the Elhuyar [16] and Zehazki [12] dictionaries.** We have only accepted the translations which are entries of Elhuyar and Zehazki dictionaries. Consequently, Phenomenon 1 in Table 1 has occurred: *ospea_kendu* “discredit” is a collocation and not an entry, so we will not take it into account. In contrast, other words in the table are entries in the dictionary and they are maintained.

iv) **Sentiment value selection.** The value (and meaning in Spanish) of each word in Basque will be selected.

In order to choose the value, we have followed the following criteria:

- If the word in Basque has one translation (and value) in Spanish and if that translation is correct, the translation is selected. This is the case of phenomena 2 and 4 in Table 1. Sometimes the translation is not “correct” or “direct” as we will observe in Section 4.
- If the word in Basque has many translations (and values) in Spanish, the translation has been selected according to which translation is the best to use in the Basque Opinion Corpus [1]. We have analyzed the context of the words in the corpus using Key Word In Context (KWIC) format for concordance. This is the case of Phenomena 3 and 5 in Table 1.
- In the creation of the first version of the lexicon, there have also been cases where the word in Basque has not instances in the corpus. In these cases, the meanings that are used more frequently have been selected.

After these four steps, the first version of the Basque lexicon (V1.0) has been created. However, we detected some inconsistencies and we have felt the necessity to feed more information and, for that reason, we followed new steps to create the Basque lexicon (V2.0):

v) **Domain and corpus adaptation: New lexicon based on the Basque Opinion Corpus [1].** We have curated the first lexicon (Basque V1.0) and created the second version of this lexicon (Basque V2.0). This new lexicon has been curated with the information obtained form word frequencies we have extracted from the Basque Opinion Corpus.

The effects of this step are showed in Phenomenon 2 in Table 1. The word *atrofiatu* “to atrophy” does not appear in the corpus,

so it is not related to the domains of the corpus and, consequently, we do not take it into account. We do not take into account them because our work is limited to our corpus and we want to maintain as much as possible the coherence of SO values and avoid complexities which we do not see useful. In Table 1, Phenomena 3, 4 and 5 are not affected by this limitation while Phenomenon 2 is. With this procedure, the number of entries in the lexicon was reduced from 8,140 to 1,813 words, because it was manually checked and reviewed.

vi) **Curate and check SO values of each entry**: Find the English translations of each Basque entry in the SO-CAL English dictionary. Using the Elhuyar dictionary [16], we have translated the words in Basque to English and, after that, we have checked if the translated words are in the SO-CAL English dictionary. If the word is in this dictionary, we have maintain the dictionary entry and its value in the second version of the Basque dictionary. If the word is not in the English dictionary, almost in all cases. it was excluded from the second version in the Basque dictionary.

In Table 1, Phenomena 3 and 4 do not have any translation in the English dictionary and, consequently, their (English) column in Table 1 is empty. In contrast, Phenomenon 5 has two translations according to the English dictionary: *right* and *correct*.

vii) **Evaluation and correction**: Compare and choose the best translation and value. In this step, each word in Basque has the same value, most of the times, in Spanish and English (Basque V1.0).

There are 3 different cases in this situation:

- Phenomenon 3. There is not a word in the English version corresponding to the Basque word and the previous Spanish one is not accepted. In phenomenon 3, the word *seinale* "sign" has been assigned the value -1 (Table 1, fourth column) but there is not a corresponding value in the English version and, consequently, we have removed that value.
- Phenomenon 4. There is not a corresponding word in the English version for Basque and the previous Spanish translation and value are accepted. The word *frankismo* "francoism" is related to Spain and, for that reason, it appears in the Spanish version and not in English. In this case, we have maintained the assigned value.
- Phenomenon 5. The English translation and value are the same or better quality than the Spanish ones. Phenomenon 5 shows that the Spanish and English values agree, so we have assigned the value $+3$ to *zuzen* "correct". In other cases, the English and Spanish values differ. When this happens we decided that the English value will prevail to the Spanish one in the second version of the Basque dictionary, because the quality is slightly better in English as we previously report.

Phenomena 3 and 5 show how we have decided to give more relevance to the English version.[1]

4 Discussion

We explain in this section how we have solve the most fundamental problems we have found during the translation process:

i) **Source language is not always the preferred language.** English and Spanish could be the source language but we have chosen Spanish due to several reasons. The overall accuracy of the English SO-CAL is 76.62% while in the Spanish version is 71.81% [2]. In other words, the difference between them is not big enough. On the other hand, there are many more resources to translate

[1] Sometimes there is not a corresponding word in the English dictionary [16], an example and the explanation of what we have done in such cases is explained in Section 4.

Table 2. Examples of translations applying the coherence criteria

Criteria	EUS	Value	EUS	Value
A	errukigabe "ruthless"	-4	errukigabeko "(with) ruthless"	-4
B	tonto "stupid"	-3	tuntun "stupid"	-3
C	arduradun "responsible"	$+2$	arduragabe "irresponsible"	-2

the dictionary from Spanish to Basque than to translate from English to Basque. So, the translation from Spanish is more reliable and extended as shown in Table 1, where the phenomenon numbered 4 (*frankismo* "francoism") shows that although the English dictionary contains more items, there are some words in the Spanish dictionary that are not present in the English one.

In contrast, the English version has helped to check if the assigned value to the Basque word in the first version from Spanish is correct. In the cases where the value of the Spanish and English versions are different, we have preferred the English one as Phenomenon 3 (*seinale* "signal") shows. Due to this decision, the number of words of the lexicon has decreased from 1,813 to 1,237 entries.

ii) **Not one to one translation**. Another problem was presented when, in the translation, a Spanish word could be translated into Basque in different forms but with the same sense. We have decided to use all the translated words in Basque so as to get the higher recall possible. The first step, the automatic translation from Spanish into Basque, shows that one or more entries have been taken in Basque.

For example, the Spanish word *aparatoso* "showy, spectacular" has been translated into Basque in two different ways: *arranditsu* "spectacular" and *deigarri* "showy".

iii) **Domain adoptation of polysemic words**. There are some words that have opposite meanings according to their context. The best solution would be to create two entries but then it would be difficult to implement it in a system that does not distinguish between word senses. In this situation, we have decided to take only one meaning and we have used the Basque Opinion Corpus [1] to choose the meaning with the appropiate SO value.

For example, the Basque word *deigarri* "showy, spectacular" comes from Spanish *aparatoso* -3 "spectacular" or *llamativo* $+3$ "showy". Taking the context of the word in the corpus into account, we have disambiguated the word manually and chosen the value $+3$ for this word.

iv) **Coherence consistency.** In the process of choosing the value, we have to try (when the values match) to maintain the coherence of the values taking these criteria into account. Examples of the criteria are shown in Table 2.

A) Sometimes, the same word appears in different forms. For example, in the creation of the first version of the lexicon, it is usual that one word appears sometimes with genitive *-ko* "with" and other times with an elided genitive, and in both cases is a dictionary entry. In these cases, we decided to assign the same value. One of the cases is the adjective *berehala* "immediate". It appears with genitive suffix: *berehalako* "immediately" and without it *berehala* "immediate". We have assigned the same sentiment value $(+2)$ to both.

B) We assign (when the values match) the same value to words with similar meanings. For example, *tonto* "stupid" is used with man while *tuntun* with the same meaning is used with woman. We assign the value -3 to both.

Table 3. The semantic oriented Basque lexicons (V1.0 and V2.0)

	V1.0		V2.0	
Grammatical category	**Words**	**%**	**Words**	**%**
Noun	2,282	28.06	461	37.27
Adjectives	3,162	38.85	446	36.05
Adverbs	652	7.98	54	4.36
Verbs	1,657	20.36	276	22.32
Intensifiers	387	4.75		
Total	**8,140**	100	**1,237**	100

C) We also assign the same intensity (1 to 5), but opposite value (positive/negative) to antonymic words when the values coincide in Basque dictionary entry. In Basque, some prefixes (*des-* and *ez-* "dis-") and suffixes (*-ezin* "impossibility" "inability" and *-gabe* "without") are used to invert the meaning of the words and we have put special attention on these ones.

v) **"Incorrect" translations.** There have been some translations which are incorrect because of different factors. The Spanish word *provinciano* "backward" (-1) is employed to refer to people of Bizkaia and Gipuzkoa provinces. The Elhuyar dictionary [16] has defined the word as "inhabitant of Bizkaia or Gipuzkoa", a translation which is not useful for our purpose.

vi) **"Indirect" translations.** There have been some translations that we have considered as indirect. They are correct translations but since they have an extensive meaning and they are used in limited situations, they are not useful for us.

For example, the word *beltz* "black" could have two meanings: i) a color ii) "black, sad; gloomy, depressing" (figurative meaning). The figurative use of that word is less usual, there are other words with the same meaning and, taking into account that the word could complicate the correct sentiment value assignation of texts, we have decided not to assign any SO value.

The explained problems show the difficulty to translate a semantic oriented lexicon semi-automatically. This translation process is large and very detailed where the translation of the lexicon has different phenomena.

5 Results

As a result of the translation process, two versions of the semantic oriented Basque lexicon have been created. Table 3 shows the characteristics of these two versions.

The first version (V1.0) is the result of the first four steps in the translation process (Figure 1). It is translated directly from the Spanish SO-CAL dictionary with a strict criteria. But, unlike the second version (V2.0), the first version is not subject to the restrictions of being an entry of the Basque bilingual dictionaries and it was not improved taking into account the English SO-CAL dictionary, the Basque Opinion Corpus and other kind of features that work differently such intensifiers are considered as dictionary entries.

As a result of these considerations, the first versions have 8,140 entries and the second version 1,237, respectively. In both cases, nouns and adjectives are the grammatical categories with more entries. Verbs and adverbs are least frequent entries, whereas intensifiers have not been taken into account in the second version because they affect to other words, so we think that it is better to analyze differently assigning different values that does not go from -5 to -5 values.

Table 4. Examples of parallel lexicon

Word in lexicon	Value	SPA	Value	ENG	Value
bikain	$+5$	excepcional	$+5$	excellent	$+5$
on	$+2$	buen	$+2$	-	-
eskas	-1	escaso	-2	insufficient	-1
txar	-3	adverso	-3	bad	-3

Another interesting characteristic of the created lexicon is that it is parallel. That means that each word of the lexicon has it translations in English and Spanish and the sentiment values in each language also are included. This information appears in an orderly manner in the resource.

In Table 4, there are four examples showing the parallel lexicon. Sometimes, four sentiment values do not match because the Spanish and English SO-CAL lexicons have been created in different way. But the Basque word always matches with one of them. The examples of Table 4 are adjectives and they show how the sentiment values are in a scale.

Once we have implemented this lexicon in the Basque SO-CAL preliminary version, the created semantic oriented lexicon is useful to assign sentiment value to words as well as sentences, as is shown in the following examples:

(1) [*Halere, pentsa litekeenaren aurka, gaien urritasunak eta diskurtso errepikakorrak*$_{-6}$ *ez dakarte ñabardura aberastasunik, are gutxiago argumentu-mailako sakontasunik.*]$_{-6}$
(However, contrary to what is thought, the scarcity of problems and the repetitive$_{-6}$ discourses do not imply rich nuances, much less a plot depth.)$_{-6}$

(2) [*Arazo nagusia*$_{+2}$*, nire ustez, gaien*$_{+4}$ *emankortasun zalantzazkoan eta ekintzaren bilakaera eskasean*$_{-3}$ *datza.*]$_{+3}$
(The main$_{+2}$ problem is, I believe, the uncertain fertility of the topics$+4$ and the slow$_{-3}$ evolution of the action.)$_{+3}$

(3) *(...) [Emaitza ezustekorik*$_{-1.5}$ *gabeko istorio bat da, irakurlea epel*$_{-1.5}$ *uzteko arrisku dezente duen tonu arras moderatu batean emana.*]$_{-3}$
(The result is an unsurprising$_{-1.5}$ story, given in a moderate tone with a risk to leave the reader cold$_{-1.5}$.)$_{-3}$

As we show in the three examples the words of the dictionary have a SO value at the end of the word. To mention one, in Example 1, the Basque version of SO-CAL tool assigns the value -6 to the word *errepikakor* "repetitive". There is no another word with sentiment value according to lexicon, so the sentiment value of the sentence is also -6.[2] The methodology to calculate the semantic orientation of the sentence is similar in Examples 2 and 3.

6 Evaluation

In this section, we want to evaluate two aspects of the translation task. On the one hand, we want to evaluate the difficulty of the task. We think that the annotation of sentiment polarity is a difficult task because there is not a guide to follow and subjective perceptions must be, first, measured and, last, corrected if possible. On the one hand, the inter-annotator agreement of SO value annotation has been evaluated between two linguists annotators. On the other hand, we also want to measure the quality of the translated lexicon. With these in mind, a gold standard annotation has been created from the previous annotation and discussion by both annotators.

[2] In this sentence, the sentiment value of the word *errepikakor* "repetitive" in the lexicon is -4. But in SO-CAL tool, there are some mathematical operations related to linguistic phenomena that increase or decrease the sentiment value of the words. In this case, the sentiment value has increased to -6.

Table 5. Pearson correlation measurement and contingency table between two annotators

Grammatical category	Pearson 1	Pearson 2
Noun	0.87	0.59
Adjectives	0.71	0.60
Adverbs	0.93	0.82
Verbs	0.87	0.76
Total	**0.79**	**0.73**

Total categories			
	0	**NEG**	**POS**
0	187	12	27
NEG	14	42	5
POS	39	5	69

In order to evaluate these two aspects, we have extracted the most frequent 400 words (100 per each grammatical category) using Analhitza [8] from the Basque Opinion Corpus [1]. We have used Pearson correlation [17] to evaluate both tasks. Pearson correlation has been used in two different ways: i) Pearson 1: the correlation is measured taking into account only the annotated words by both annotators and ii) Pearson 2: the correlation is measured taking into account all words in the corpus.[3]

6.1 Correlation between annotators

We have decided to measure the correlation of two annotators to create the gold standard, taking into account the results achieved in the correlation coefficient. Table 5 shows the coefficient for each grammatical category, together with a contingency table.

Pearson 1 value shows that the correlation coefficient is high (0.79). This means that the value assigned is similar in a big percentage of the annotated words. The coefficients for different grammatical categories are situated between 0.71 and 0.93. In a similar way, Pearson 2 also shows high correlation, although it is slightly lower (0.73), with values between 0.59 and 0.82.

The contingency table of Table 5 shows that the biggest difference comes when one annotator has assigned a value to one word and the other one had not assigned any value and vice versa (90.19 % of all discrepancies 92 of 102).

After calculating this correlation, two annotators have discussed about their differences and after reaching consensus, a gold standard has been created.

[3] This means that there are cases where one word has been annotated by one annotator or by none of the them. When it happens, the un-annotated words value is 0 in order to calculate the Pearson correlation.

6.2 Correlation between the lexicon and gold standard

The correlation between the human gold standard lexicon and the translated lexicon shows some differences compared to the correlation between two annotators as presented in Table 6.

With Pearson 1, the cases in which the dictionary and gold standard contain an annotation for the word show similar correlation when compared to the results of two annotators (0.79). The correlation is high since the coefficients for the different grammatical categories are situated between 0.69 and 0.96. In contrast, Pearson 2 shows a lower correlation (0.54) and the coefficients of grammatical categories are situated between 0.47 and 0.59.

The interpretation of these results is that the values assigned to the dictionary and gold standard are similar (Pearson 1). But the difference from the previous result in Pearson 2 is created when the semantic oriented lexicon assigns value to the word and the annotator does no do it. This situation does not occur in the correlation between two annotators.

The contingency table shows us how the gold standard and the created dictionary differ. The discrepancy here also comes from the difficulty to assign a positive or negative value to a word. The difference is similar: 89.83 % of all discrepancies (106 of 118) are related to the decision to assign sentiment polarity to words. But here, in contrast with correlation between two annotators, the last version of the lexicon is more conservative, because the gold standard

Table 6. Pearson correlation measurement and contingency table between the gold standard and the Basque semantic oriented lexicon (V2.0)

Grammatical category	Pearson 1	Pearson 2
Noun	0.96	0.59
Adjectives	0.78	0.56
Adverbs	0.75	0.47
Verbs	0.69	0.54
Total	**0.76**	**0.54**

Total categories			
	0	**NEG**	**POS**
0	195	2	15
NEG	30	34	8
POS	59	4	53

annotates much more words than the lexicon does, decreasing the correlation in Pearson 2.

To sum up, the evaluation shows a high correlation in Pearson 1 in the case of two annotators and the lexicon and gold standard. The correlation coefficient is 0.79 and 0.76, respectively. In the case of Pearson 2, the correlation between two annotators remains high (0.73) but the correlation measure falls between the lexicon and gold standard (0.54).

7 Conclusion and Avenues for Future Work

In this paper we presented the first semi-manually created semantic orientation lexicon for Basque[4]. Time factor, few resources and quality pushed us to translate the SO-CAL Spanish dictionary to Basque.

The translation process has followed several steps. To summarize the steps, the English and Spanish SO-CAL dictionaries have been translated into Basque using two bilingual dictionaries. After that, the groups of words with the same meaning have been grouped and the best sentiment values according to the context of the Basque Opinion Corpus have been chosen. Finally, the created lexicon has been adapted to the domains of the Basque Opinion Corpus. The Basque sentiment lexicon has its limitations, since polysemy and figurative meaning phenomena were not considered and therefore are not totally solved.

Pearson correlation shows that the agreement coefficient is high between both annotators with respect to the following two factors: $i)$ assigning a value and $ii)$ deciding if a word has any value. In contrast, in the case of the comparison between human gold standard and translated lexicon, the correlation coefficient is high when the value is assigned but not in the case of deciding if the word has a value or not, which results has been lower. This lower coefficient appears mainly because there are less words annotated in our translated lexicon V2.0.

At present, the second version of semantic oriented lexicon is implemented in the Basque SO-CAL. In a foreseeable future, our aim is to improve this lexicon but considering morphosyntactic and discourse phenomena. This lexicon will be the basis of this system and we will consider how to enrich the system with sentence level and text level information.

[4] The semantic oriented Basque lexicon is available at: `http://ixa.si.ehu.es/node/11438`

Acknowledgements

Jon Alkorta's work is financed by a Basque Government scholarship (code: PRE_2017_2_0041). Koldo Gojenola's work is partially financed by the PROSA-MED (TIN2016-77820-C3-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). Mikel Iruskieta's work is partially financed by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). We would like to offer our special thanks to Maite Taboada, Arantxa Otegi and Oier Lopez de Lacalle.

References

1. **Alkorta, J., Gojenola, K., & Iruskieta, M. (2016).** Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. *Proceedings of Fourth International Workshop on Discourse Analysis (IWoDA16)*, pp. 58–62.
2. **Brooke, J., Tofiloski, M., & Taboada, M. (2009).** Cross-linguistic sentiment analysis: From english to spanish. *Proceedings of the international conference RANLP-2009*, pp. 50–54.
3. **Chetviorkin, I. & Loukachevitch, N. (2012).** Extraction of russian sentiment lexicon for product meta-domain. *Proceedings of COLING 2012*, pp. 593–610.
4. **Chetviorkin, I. & Loukachevitch, N. (2014).** Two-step model for sentiment lexicon extraction from twitter streams. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 67–72.
5. **Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008).** Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. *Procesamiento del lenguaje natural*, Vol. 41, No. 0.
6. **Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014).** Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, Vol. 53, pp. 113–120.
7. **Goyal, A. & Daumé III, H. (2011).** Generating semantic orientation lexicon using large data and thesaurus. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, pp. 37–43.
8. **Otegi, A., Imaz, O., de Ilarraza, A. D., Iruskieta, M., & Uria, L. (2017).** Analhitza: a tool to extract linguistic information from large corpora in humanities research. *Procesamiento del Lenguaje Natural*, , No. 58, pp. 77–84.
9. **Pak, A. & Paroubek, P. (2010).** Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, volume 10, pp. 1320–1326.
10. **Rosenthal, S., Farra, N., & Nakov, P. (2017).** Semeval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518.
11. **San Vicente, I., Agerri, R., & Rigau, G. (2014).** Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 88–97.
12. **Sarasola, I. (2005).** *Zehazki: gaztelania-euskara hiztegia*. Alberdania.
13. **Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966).** The general inquirer: A computer approach to content analysis.
14. **Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011).** Lexicon-based methods for sentiment analysis. *Computational linguistics*, Vol. 37, No. 2, pp. 267–307.
15. **Tan, S. & Zhang, J. (2008).** An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, Vol. 34, No. 4, pp. 2622–2629.
16. **Zerbitzuak, E. H. (2013).** Elhuyar hiztegia: euskara-gaztelania, castellanovasco. usurbil: Elhuyar.
17. **Zou, K. H., Tuncali, K., & Silverman, S. G. (2003).** Correlation and simple linear regression. *Radiology*, Vol. 227, No. 3, pp. 617–628.

Article received on 18/12/2017; accepted on 15/02/2018.
Corresponding author is Jon Alkorta.

4

Contextual valence shifters

4.1 Article 3

Saying no but meaning yes: negation and sentiment analysis in Basque

Saying no but meaning yes: negation and sentiment analysis in Basque

Jon Alkorta
Computer Languages and Systems
IXA group (UPV/EHU)

Koldo Gojenola
Computer Languages and Systems
IXA group (UPV/EHU)

Mikel Iruskieta
Didactics of Language and Literature Department
IXA group (UPV/EHU)

{jon.alkorta,koldo.gojenola,mikel.iruskieta}@ehu.eus

Abstract

In this work, we have analyzed the effects of negation on the semantic orientation in Basque. The analysis shows that negation markers can strengthen, weaken or have no effect on sentiment orientation of a word or a group of words. Using the Constraint Grammar formalism, we have designed and evaluated a set of linguistic rules to formalize these three phenomena. The results show that two phenomena, strengthening and no change, have been identified accurately and the third one, weakening, with acceptable results.

1 Introduction

Negation is a morphosyntactic operation in which a lexical item denies or inverts the meaning of another lexical item or language construction (Loos et al., 2004). The effect of the negation can be the change of semantic orientation (SO) and, according to Liu (2012), negation is called sentiment shifters because they change the semantic orientation of a word or a sentence.

With the aim of calculating the semantic orientation, the first step is to build a lexicon, but this is not enough, to grasp the correct SO-value of Example 1.

(1) [*Irabazi*$_{+2}$ *ezinik jarraitzen du Eibarrek*]$^{+2}$. (KIR17)
[(The soccer team) Eibar continues without winning$_{+2}$]$^{+2}$.

Following the semantic lexicon Sentitegi (Alkorta et al., 2018)[1], the semantic orientation of the word *irabazi* ("to win") is $+2$, and consequently, of the sentence also is $+2$. But we can notice that the semantic orientation of the sentence is clearly negative. The negator *ezin* ("can not") turns the positive oriented word *irabazi*$_{+2}$ ("to win") into a negative oriented one. Therefore, we think that addressing this phenomenon is crucial to obtain better results in the calculation of the SO of texts.

[1]The semantic lexicon is available on the web at: http://ixa.si.ehu.es/node/11438

The main aim of this work is to study how negation expressions and syntactic structures can change the semantic orientation of words, and to design a set of linguistic rules by means of Constraint Grammar (Karlsson et al., 2011) in order to identify these phenomena. According to our corpus study, different negation language forms can strengthen, weaken or have no effect on semantic orientation. These results go in the same direction as (Jiménez-Zafra et al., 2018b) where effects of negation within its scope are studied. We have centered our study on negation markers that unlike negation in verbs and nouns and negative polarity items, they only share information about negativity while others can share more information like aspect of action (e.g. *they denied going to the city*).

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes methodological steps. Then, Section 4 presents theoretical framework, while Section 5 gives a linguistic analysis. Section 6 shows results and error analysis, concluding with Section 7 and proposing directions for future work.

2 Related Work

There is a variety of works about negation and sentiment analysis in different languages and from different approaches.

For English, Liu and Seneff (2009) have presented a work where a parse-and-paraphrase paradigm is used to assign sentiment polarity for product reviews. If negation is detected, its polarity will be reversed (switch negation). If it has a value of $+5$, it will be reversed to -5, and vice versa. Following this, they have improved results (recall was improved in 45 %). The treatment of negation has been different in Taboada et al. (2011). In their work, when a negator is identified, the polarity value is not reversed; instead it is shifted toward the opposite polarity by a fixed amount. This approach is called shift negation. In

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 85–90
Brussels, Belgium, October 31, 2018.
https://doi.org/10.18653/v1/P17

Text	Text span	Dictionary words	SO value
MUS20	*Pogostkinak [ezin hobeki]*$^{+}$ *atera zituen* Pogostkina took them out [in an unbeatable way]$^{+}$	*hobeki* best, better	$+2$
KIR17	[*Irabazi ezinik*]$^{-}$ *jarraitzen du Eibarrek*, Eibar continues [without winning]$^{-}$	*Irabazi* to win	$+2$

Table 1: Polarity extraction of words (step 2) and linguistic analysis (step 3).

the creation of the semantic orientation calculator (SO-CAL tool), Taboada et al. (2011) have also treated negation in combination with other linguistic phenomena (like irrealis or intensifiers).

In Spanish, there are several works related to negation and sentiment analysis. In the case of Jiménez Zafra et al. (2015), firstly, they have analyzed what the effects of different negators in different sentences are. After that, they have created linguistic rules defined by the previous analysis. Finally, they developed a module that has been included in their polarity classifier system, improving results between 2.25 % and 3.02 % depending on the resource. Vilares et al. (2015) have used a syntactic approach for opinion mining on Spanish reviews. This system treats negation taking into account the scope and polarity flip caused by negation. According to their results, there is an improvement, due to the implementation of negation, among other reasons.

Our work is related to (Taboada et al., 2011) and (Jiménez Zafra et al., 2015) since it is based on a linguistic analysis and also because a set of rules that detect the negation language forms are created. As far as we know, there is not any work which analyzes negation in connection with sentiment analysis in Basque.[2]

3 Methodological steps

1- Negation corpus. We have extracted 359 negation instances of seven[3] negation markers. They were extracted from a total of 96 reviews of six different topics: movies, music, literature, politics, sports and forecast. We have selected those negation markers because they are the most frequent in the corpus.

2- Polarity extraction of every instance. We have created a polarity tagger, based on a POS tagger (Ezeiza et al., 1998) to enrich the corpus with POS information on a semantic oriented lexicon for Basque (Alkorta et al., 2018), to assign the semantic orientation value (SO value, between -5 and $+5$) to words, as shown in Table 1. There, the adverb *hobeki* ("best", "better") and the verb *irabazi* ("to win"), have a SO value of $+2$ in the lexicon.

3- Linguistic analysis. We have analyzed whether the negation markers can change the semantic orientation and the SO value of sentences. We have also tried to identify whether there are other phenomena related to negation with or without effects on semantic orientation. In Table 1, in MUS20, the negation marker appears near *hobeki* ("best"), an adverb. The result of this combination is strengthening. In contrast, in KIR17, the verb *irabazi* ("to win") is before the negator and the result is weakening. These two examples show the different performances of *ezin(ik)* ("can not"). Consequently, in Table 1, for example, this negation marker appears in two different groups. The same methodology has been used with other negation markers.

4- Constraint Grammar (CG3) rules for negation. Several rules have been proposed to detect each group, in order to identify the effects of negation based on the linguistic analysis presented in Section 5.

5- Evaluation. We use F_1 to evaluate the results using a different set of 46 reviews from the same corpus (Alkorta et al., 2016)[4].

4 Theoretical framework

In this section, we explain the three most important concepts, regarding our analysis: i) scope (negation analysis) and ii) switch and iii) shift negation (sentiment analysis approach to negation).

(2) *Berez pianorako konposatutako poliptiko txiki honek* **ez** *du bere naf kutsua galtzen*$_{-2}$ *bertsio orkestratuan.* (MUS01)
This small polyptych composed for the piano does **not** lose$_{-2}$ its naive sense in the orchestral version.

[2] Altuna et al. (2017) also analyze negation but their point of view is different, since they analyze events in Basque texts.

[3] The extracted negation markers from the Basque Opinion Corpus (Alkorta et al., 2016) are the following: *ez* ("not"), *gabe* ("without"), *ezin* ("can not"), *salbu* ("except (for)"), *izan ezik* ("except (for)"), *ezta* ("not even, not either") and *ezean* ("in the absence of" or "unless").

[4] A part of the corpus is available on the web at: http://ixa2.si.ehu.es/diskurtsoa/fitxategiak.php

(3) *−maitasun istorio konbentzional bat, grazia*$_{+3}$ *handirik*$_{+1}$ ***gabe****a−*. (LIB07)
*−*a conventional love story, **without** great$_{+1}$ grace$_{+3}$.[5]

According to Huddleston and Pullum (2002), the scope of negation is the part of the meaning that is affected by the negation marker, changing or not their SO value. In the examples above, the scope is underlined. As our study shows, there can be two kinds of semantic orientation in scope and these can be changed by negation markers. In Example 2, the SO value of the verb *galdu* ("to lose") and of its scope is -2. The negation weakens the SO value of the verb, reversing its SO. But, in Example 3, the SO values of the noun *grazia* ("grace") $+3$ and the adjective *handi* ("great") $+1$ assign a SO value of $+4$ to the scope which is positive. The negator *gabe* ("without") weakens the SO value.

According to Taboada et al. (2011), there are two approaches in sentiment analysis to weaken the negative SO value: i) switch negation and ii) shift negation.

(4) This pub is [**not** good$_{+3}$]$^{-3(switch)}_{-1(shift)}$ but the music from there is good$_{+3}$.

In the switch negation approach, the SO value of Example 4 is reversed. The SO value of the adjective *good* is $+3$ while the reversed SO value is -3. However, this criteria has a problem: if *excellent* is $+5$; *not excellent* would be more positive ($+1$) than *not good* (-2), but the SO value points to the contrary (*not excellent* is more negative than *not good*).

Otherwise, in the shift negation, the different negators have their own SO value and the results depend on the interaction of both SO values (the value of negation marker and negated word). Taking into account Example 4, the SO value of the negation *no* is -4 in the dictionary; so, when it modifies the word *good*, which has a SO value of $+3$, the sum value of scope is -1. This is the way how the shift approach solves the problem we describe in Example 4. We have decided to use the shift negation approach assigning a ± 4 SO value to the negators.

5 Linguistic analysis

In the theoretical framework of the shift negation, it has been considered that negation markers only weakens the SO value. Nevertheless, we have identified two other functions of these negation markers with low frequency, but relevant anyway from our point of view as the works of (Jiménez-Zafra et al., 2018a) and (Jiménez-Zafra et al., 2018b) show. As we observed in this study, the negation markers can strengthen, weaken or have no effect in the SO value of its scope as Figure 1 shows.

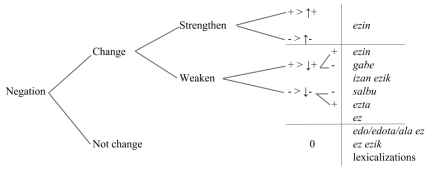


Figure 1: The effects of negation on semantic orientation according to negation markers.

The majority of negation markers usually weaken the semantic orientation of scope. But as we can see in Figure 1, the negation marker *ezin* ("can not"), for example, can strengthen or weaken the semantic orientation of scope. The weakening can be understood in two ways: i) if the word or scope of the semantic orientation is $+5$, $+4$, -5 or -4, their semantic orientation will not become negative because according to our methodology (shift negation), due to our SO value of the negators is ± 4. In contrast, ii) if the semantic orientation of scope or sentence is between -3 and $+3$, their semantic orientation will be reversed. iii) Finally, negation with conjunction, contrastive negation and lexicalized structures do not change the SO value of the scope.

5.1 Negation strengthening the SO

Among all the negation instances, we have observed some cases where the semantic orientation has been strengthened (1.96 %: 7 of 359). This happens when the negation marker *ezin* ("can not") modifies adjectives or adverbs.

(5) *Dena nahasten da maisulan* ***ezin*** *ederragoa*$_{(+4)}$ *osatzeko.* (MUS21)
Everything is mixed to create a masterpiece that can **not** be more beautiful$_{(+4)}$.

In Example 5, the negator modifies the adjective and, in this case, the negation with an adjective in a comparative structure is used to reinforce the positive SO value. The result

[5] **Bold** is used to mark the negator, underline means the scope of negation.

Example	Negation marker	Categorization	Instances
6	ez	[(NP +)] *ez* [+ aux. (+ NP) + verb (+ NP)]	214
		[(NP +) verb +] *ez*	18
		[NP +] *ez*	13
		ez [+ NP +] *ez* [+ NP] (...) (repetitive)	2
7	gabe	[NP/VP/clause +] *gabe*	41
	ezin	[(NP) + verb +] *ezin*	19
		ezin [(+ NP) +] verb [(+ NP)]	5
	salbu	[NP] + *salbu*	2
	izan ezik	[NP/clause] + *izan ezik*	1
	ezta	*ezta* + [NP/clause]	1
	ez, ezin	with any clear pattern	7
	Total		**323**

Table 2: Negation weakening the semantic orientation.

of negating a positive chunk *can not be more beautiful* is to be even more positive. In this case, the *masterpiece* is *very beautiful*.

5.2 Negation weakening the SO

In the majority of cases, the SO value is weakened due to negation. Several negation markers can weaken the semantic orientation. In our corpus, 89.98 % of cases (323 of 359) show a weakening of scope.

(6) *Horrek* ***ez*** *die eragotzi*$_{(-2)}$ *ordea, 57 milioi euro ematea San Mames klub pribatuari!.* (POL30)
It does **not** prevent$_{(-2)}$ them, however, to give 57 milion euros to San Mames private club!

(7) *Irabazi*$_{(+2)}$ ***ezin****ik jarraitzen du Eibarrek, baina oso puntu ona eskuratu du Getaferen zelaian.* (KIR17)
Eibar continues **without** winning$_{(+2)}$, but it has achieved a very good point in Getafe's (football) field.

In Example 6, the default word order of Basque (main verb + auxiliary verb) was reversed in a typical negation structure (*ez* "not" + auxiliary verb + main verb). In this example, the negation marker *ez* ("not") has an effect on all the words of the sentence, including the verb *eragotzi* ("prevent") which has a negative SO value (-2), weakening its SO value. In Example 7, the negation marker *ezin* "can not" negates the verb *irabazi* ("win"). Therefore, the negation marker *ezin* ("can not") works like an intensifier does with adjectives and adverbs (Example 5) while it has the opposite function with verbs and nouns (Example 7). Therefore, weakening negators can have a positive or negative (± 4) SO value, if the modified chunk (scope) has a positive or negative SO value. The same happens if the SO value is positive $+5$, because the result of the weakening (-4) will not change the polarity and the SO value will still be positive $+1$. In contrast, if the SO value of the modified chunk $+3$ or -3 or lower, the SO value will be reversed to a ± 1. This happens in Example 6 and Example 7. In the first example, the SO value of the scope is $+2$ (*eragotzi* ("prevent") -2 + *ez* ("not") $+4$ = $+2$). In the second one, the SO value of the scope is -2 (*irabazi* ("win") $+2$ + *ez* ("not") -4 = -2).

5.3 Negation with no effect

Negation with no effect on semantic orientation has happened in 8.08 % of our sample (27 of 359). In these cases, the negation does not modify any word with a SO value assigned. This can happen due to three reasons: i) the negator appears with a conjunction, ii) the negator is a part of contrastive negation and iii) the negator is part of a lexicalized structure (structures with their own meaning and sometimes also corresponding to dictionary entries). The scope concept is applicable only in the case of contrastive negation and the particle *ez* ("no") with a conjunction.

(8) *Ikuspuntu politikotik*$_{(-1)}$ *ez ezik, ekonomikotik*$_{(+3)}$ *ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira.* (POL08)
Not only from the political point of view, but also from the economic point of view, Greece has also hoped for other parts of southern Europe, including the Basque Country.

(9) *Sei puntu baino ez dituela, hamaseigarren postuan da Reala sailkapenean.* (KIR27)
With only six points, Real is in the sixteenth position in the classification.

Example 8 shows a contrastive negation with additive function (Silvennoinen, 2017). In other words, the negation mark does not negate the noun phrase, as in *ikuspuntu politikotik*$_{(-1)}$ ("from the political$_{(-1)}$ point of view"), actually it functions as conjunction

Example	Negation marker / lexicalized structure	Instances
	[verb/*bai* "yes"] + *edo/edota/ala ez* (*ez* with conjuction)	3
8	[NP] + *ez ezik* (contrastive negation)	2
9	*baino/besterik ez*	11
	Others lexicalized structures	13
	Total	**29**

Table 3: Negation without effects on semantic orientation.

and adds new information: *ekonomikotik ere* ("also from the economic point of view"). Structures of Table 3 have their own SO value, they can be considered as dictionary entries and they can appear in different positions in the sentence. In Example 9, the structure *baino/besterik ez* ("only") is an adverb.

6 Evaluation

6.1 Evaluation methodology

To tag the negation changes of the SO value, we have created negation rules based on previous studies.Rules have been implemented using Constraint Grammar (CG3) (Karlsson et al., 2011) to assign the correct value to the negated structures. The corpus of 96 texts has been tagged using the Basque morphosyntactic disambiguator based on the CG formalism (Aduriz et al., 1997). Then, a different set of 48 texts of the Basque Opinion Corpus has been used as test dataset to evaluate the rules. After that, the results have been analyzed manually, observing if the words have been annotated or not and, when annotated, whether they have the correct annotation.

Negation effects	**Prec.**	**Rec.**	**F_1**
Strengthen	1.00	1.00	1.00
Weaken	0.93	0.80	0.86
No effect	0.97	1.00	0.98
Total	0.93	0.80	0.86
Negated elements	**Prec.**	**Rec.**	**F_1**
Negation markers	1.00	0.96	0.98
Lexicalized structures	0.96	1.00	0.98
Scope	0.91	0.75	0.82
Total	0.93	0.80	0.86

Table 4: General results of negation effects and negated elements.

Most of the corpus was evaluated by one linguist, but with the aim to know the reliability of this evaluation a piece of the corpus (10 %) has been annotated by two linguists. Both annotators have followed a guideline to evaluate the output of CG3 rules. According to the results, the Cohen's kappa score is 0.93 for the annotation of the words that belong to negation and the kappa score is 0.69 for the annotation of words that have been annotated correctly, badly or is missed (which can be considered as *substantial* in (Landis and Koch, 1977)).

6.2 Results and error analysis

According to general results, the F_1 of the negation rules identifying elements related to negation is 0.86 (Precision is 0.93 while recall is 0.80).

In accordance with weakening and scope error analysis, these elements show lower F_1 score because they behave more irregularly. The components as well as the length in scope are more unpredictable. Moreover, some negators apply to lists of words with comma and, as some constraints in CG3 rules correspond to punctuation marks, they have not been detected. This suggests that the rules need more precision. So, the punctuation mark constraint is not enough. Therefore, some syntactic information is needed to detect these kind of structures.

7 Conclusions and Future Work

This work presents a negation analysis for Basque sentiment analysis based on Constraint Grammar rules. According to this study, the negation can affect the semantic orientation (SO value) in different ways: i) strengthening, ii) weakening or iii) having no effect. According to our evaluation to measure the identified words, the overall precision is 0.93, the recall 0.80 and the F_1 score 0.86. In line with error analysis, the punctuation mark constraint is not enough and more precise rules are needed in the negation weakening. In the near future, i) we want to implement these negation rules in a tool for automatic Basque sentiment analysis and ii) we want to continue with the analysis of negation: analyzing the scope in a bigger corpus and especially based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), studying if the position of negator in rhetorical structure has any effect on sentiment analysis.

References

Itziar Aduriz, José María Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Koldo Gojenola, and Montse Maritxalar. 1997. Morphosyntactic disambiguation for basque based on the constraint grammar formalism. *Proceedings of Recent Advances in NLP (RANLP97)*, pages 282–288, Tzigov Chark (Bulgary).

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2016. Creating and evaluating a polarity - balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis*, pages 58–62. Santiago de Compostela (Spain).

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2018. SentiTegi: building a semantic oriented Basque lexicon. In *Proceedings of the CICLing 2018*. Hanoi (Vietnam).

María Jesus Aranzabe Begoña Altuna and Arantza Díaz de Ilarraza. 2017. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria and Patxi Salaberri (pub.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea*, pages 127–134. Udako Euskal Unibertsitatea (UEU), Bilbo (Spain).

Nerea Ezeiza, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics.

Rodney Huddleston and Geoffrey Keith Pullum. 2002. The Cambridge grammar of English. *Language. Cambridge: Cambridge University Press*.

Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, M. Dolores Molina-González, and L. Alfonso Ureña-López. 2018a. Relevance of the SFU Review SP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish. *Information Processing & Management*, 54(2):240–251.

Salud María Jiménez Zafra, Eugenio Martínez Cámara, María Teresa Martín Valdivia, and María Dolores Molina González. 2015. Tratamiento de la Negación en el Análisis de Opiniones en Espanol. *Procesamiento del Lenguaje Natural*, (54).

Salud María Jiménez-Zafra, Mariona Taulé, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and M. Antónia Martí. 2018b. SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Jingjing Liu and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 161–169. Association for Computational Linguistics.

Eugene Emil Loos, Susan Anderson, Dwight H. Day, Paul C. Jordan, and J. Douglas Wingate. 2004. *Glossary of linguistic terms*, volume 29. SIL International Camp Wisdom Road Dallas.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Olli O. Silvennoinen. 2017. Not only apples but also oranges: Contrastive negation and register. *In Turo Hiltunen, Joe McVeigh and Tanja Sily (edit.), Big and Rich Data in English Corpus Linguistics: Methods and Explorations, VARIENG, Helsinki (Finland)*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, 21(1):139–163.

4.2 Article 4

Using relational discourse structure information in Basque sentiment analysis

Using relational discourse structure information in Basque sentiment analysis

El uso de la información de la estructura retórica en el analisis de sentimiento

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta y **Alicia Perez**
IXA Group. University of the Basque Country
jon.alkorta@ehu.eus, koldo.gojenola@ehu.eus, mikel.iruskieta@ehu.eus, alicia.perez@ehu.eus

Resumen: En este artículo presentamos un estudio sobre el análisis del sentimiento que explota información extraida de la estructura relacional discursiva en un corpus en euskera sobre crítica literaria. Para el análisis discursivo hemos utilizado la *Rhetorical Structure Theory* (RST) y para la polaridad el método QWN-PPV. Los resultados preliminares demuestran que el análisis del discurso es efectivo para el análisis de opiniones.
Palabras clave: Análisis del sentimiento, polaridad, relaciones de coherencia, unidad central, RST, crítica literaria

Abstract: This paper presents a study in sentiment analysis which exploits information of the relational discourse structure in a Basque corpus consisting of literature reviews. The QWN-PPV method was employed to label all the texts at element level and the *Rhetorical Structure Theory* (RST) was used to extract discourse structure information. The preliminary results show that discourse structure is effective for opinion mining.
Keywords: Sentiment analysis, polarity, coherence relations, central unit, RST, literary criticism

1 Introduction

Sentiment analysis is nowadays a well known topic where the opinion, sentiment or subjectivity (Pang and Liu, 2008) are studied. The opinion about films (Pang and Lee, 2004), the success of politicians (Tumasjan et al., 2010) and the opinion of consumers about products are some of the topics studied.

Different levels of language has been studied in Sentiment Analysis. Hatzivassiloglou and McKeown (1997) studied the word level, Yu and Hatzivassiloglou (2003) the sentence level and Pang, Lee, and Vaithyanathan (2002) the discourse level. These are some examples of these three levels extracted form our corpus[1]:

i) Lexical level: where words[2] and entities have their own polarity[3], as in Example (1).

(1) [...] *literatura ona*$_{(+)}$ *sortu ahal izateko.* BER01
[...] to create good$_{(+)}$ literature.

In the example the word *ona* (good) has a positive polarity, because its entry in a dictionary has a positive value.

ii) Syntactic level: where the function in the word ordering or the clause's syntactic function can change the polarity assigned in the lexical level.

(2) *xede*$_{(+)}$ *onak*$_{(+)}$ *ez dira nahikoak*$_{(+)}$ *izaten literatura ona*$_{(+)}$ *sortu ahal izateko.* BER01
good$_{(+)}$ goals$_{(+)}$ are not enough$_{(+)}$ to create a good$_{(+)}$ literature.

In Example (2), the negation *ez* (not) changes the polarity assigned by the dictionary entries to a negative polarity determined by the negation of an otherwise positive statement.

[1] References and links to see the annotated text are at the end of the examples.

[2] For example, SentiWordNet (Esuli and Sebastiani, 2010) is a lexical resource for opinion mining which assigns three sentiment scores to each synset of WordNet: positivity, negativity, objectivity.

[3] In the following examples the polarity will be marked with: (+) when positive, (−) when negative and (∗) when neutral. All the examples were analyzed with the QWN-PPV method, explained below.

iii) Discourse level: where the coherence relations can highlight or even change a clause polarity (micro-structure), or the overall polarity of a text (macro-structure).

In Example (3) the change of polarity is out of the sentence scope, at discourse level.

(3) *Dokumentazio lana, esan bezala, nabarmena*$_{(+)}$ *da, eta baliabideen erabileran idazleak duen ahalmena eta egindako lana bereziki azpimarratzekoak*$_{(+)}$ *dira. Baina, horiek horrela izanik ere, emaitza zalantzagarria*$_{(*)}$ *da.* BER04
The documentation work, as mentioned before, is spectacular$_{(+)}$, and the capacity of the writer to use the resources and the work done especially are to underline$_{(+)}$. But, although that is so, the result is doubtful$_{(*)}$.

In this example, there are several words with a positive polarity, but the polarity of the example is not positive because a contrast discourse relation signaled by the adversative connector *baina* (but) has changed it.

This example demonstrates the importance of rhetorical relations, which can change the polarity of the sentence. For that reason, it is necessary, from our point of view, to also consider the discourse structure information in sentiment analysis.[4]

Currently there exists an Opinion Mining system for Basque. We have used this tool[5], that assigns automatically a positive or negative polarity to words[6]. The system makes use of the QWN-PPV method (Vicente, Agerri, and Rigau, 2014), that automatically generates polarity lexicons annotated at synset and lemma level. For that purpose, QWN-PPV uses a Lexical Knowledge Base (WordNet) and a list of positive and negative elements. QWN-PPV is a method that detects elements from lexical level and, consequently, the method is unable to correctly detect the polarity of the examples mentioned before —see examples (2) and (3).

[4]Example (4) below also shows the importance of discourse structure considering the main topic of the text and the rhetorical relation, when assigning a polarity score.

[5]Ber2Tek Opinion Mining can be tested at `http://iritzierauzketa.ber2tek.eus/`

[6]The method does not signal a neutral polarity.

With the aim of fulfilling this gap, we want to develop a method based in two language levels: the lexical and the discourse level. To that end, we will estimate the importance that the discourse structure has in sentiment analysis. Our study, which is based on a theoretical approach based on discourse, exploits information from the relational discourse structure in a Basque corpus consisting of literature reviews. The theory we employ to that end is the *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987)[7]. This theory describes the structure and coherence of text and it has been useful in Sentiment Analysis and in other many NLP advanced tasks (Taboada and Mann, 2006). In this respect, this work is a first approximation to Opinion Mining using discourse structure in Basque. This study, as far as we know, is the first work on Basque and sentiment analysis from a discourse point of view. Therefore, it fullfils this gap in Basque, but it is also relevant for RST, because it includes a different language to some recent works like (Trnavac and Taboada, 2014) in English and (Zhou et al., 2011) in Chinese.

The rest of the paper is structured as follows. Section 2 lays out the related work. Section 3 sets out the methodology we used and Section 4 presents the results. Finally, Section 5 presents a discussion and establishes directions for future work.

2 Related work

As we have explained before, Sentiment Analysis has three linguistic different levels, while we will focus on the lexical level and the discourse level interaction. Inside the discourse level, there are various categorizations according to different viewpoints.

In the discourse level there are two possible methods: language model and knowledge-based model. The first one determines if a span of a text is subjective, while the second one finds words with its polarity in a dictionary and calculates a sentiment score for all the text with an algorithm.

i) In the language model approach, Alistair and Diana (2005) use Support Vector Machines to classify sentiment expressed by movie reviews. Firstly, they use unigram fea-

[7]Other theories worth to mention are the *Segmented Discourse Representation Theory* (SDRT) (Asher, 1993) and the *Penn Discourse TreeBank* (PDTB) (Miltsakaki et al., 2005).

tures and then, they couple bigrams. The bigrams are composed by a valence shifter and another word. The results are acceptable and adding a term-counting method helps get better results.
ii) The knowledge-based approximation can be divided in four approaches (Cambria et al., 2013): keyword spotting, lexical affinity, statistical methods, and concept-level techniques. In the experiments, we will use the QWN-PPV method, which is an (almost) unsupervised method, i.e., a statistical method.

According to (Zhou, 2013), from a theoretical point of view, there are two approaches to Sentiment Analysis: discourse-based and aspect-based.
a) In a discourse-based approach not all the sentences have the same importance. Several researchers have tried to measure the contribution of sentences or phrases to the polarity of the text. Discourse Structure based Sentiment Analysis is divided in two approaches: rule based and weight based ones. In both approaches the results have improved with the addition of discourse relations. Moreover, these works have shown that the combination of two paradigms can bring an overall improvement.

In the rule based approach, Somasundaran et al. (2009) use a supervised collective classification and a supervised optimization framework in order to improve polarity classification. Text spans are extracted according to their importance in discourse structure.

Vanzo, Croce, and Basili (2014) assign a sentiment polarity to entire tweet sequences using a Markovian formulation of the Support Vector Machine discriminative model, SVM_{hmm}. In contextual information, they take into account two aspects: the conversation and the user attitude or the overall attitude of the last tweets. The individual perspective is independent in context, so they consider the tweet as a multifaceted entity. Consequently, each vector contributes in one aspect of the overall representation. The evaluation shows that sequential tagging effectively improves the detection precision approximately 20% in F1 measure.

In the weight based approach, Polanyi and Zaenen (2006) demonstrate that the structure of the text gives important information to extract the opinion. They have found that connectors increase or decrease the intensity of polarity. In this way, discourse relations can also increase or decrease the intensity.

Inspired in this previous work, Taboada, Voll, and Brooke (2008) extract the most important spans of a text and then, they calculate the semantic orientation in two ways, where the most important spans weight more. First, they use RST and they extract all the nuclei of the text. After that, they use a topic classifier based in support vector machines, improving results.

The rhetorical information of a text can be extracted automatically with a discourse parser and then this information can be used in sentiment analysis to determine the polarity of the text in a more reliable way. For example, Taboada, Voll, and Brooke (2008) and Heerschop, Goossen, and Hogenboom (2011) use the Sentence-level PArsing of DiscoursE (SPADE) tool in order to extract the discourse relations automatically from the text (Soricut and Marcu, 2003).
b) In Aspect-based Sentiment Analysis, the subject of a review is important because it helps to predict the "relevant" polarity expressed in the text. In this way, words related with the aspect help to give more accurate results.

Lim and Buntine (2014) combine language model and aspect creating the LDA-based[8] Twitter Opinion Topic Model. This model uses strong sentiment words, hashtags, mentions and emoticons to predict the opinion modeling the target.

The OpeNER project (Opener, 2013) makes use of three components: *i*) opinion express, *ii*) opinion holder and *iii*) opinion target. There are four tag levels in the Annotation Tool of the project. Tagging is based in three parameters: positive / negative attitude, sentence "on-topic" and "to-the-point"[9]. In the project, topic sentence can be a touristic attraction, a restaurant or a hostel. The opinions indirectly linked with the reviewed entity are also considered as "on-topic". The "to-the-point" category implies that a reviewer gives an opinion of the annotation object and then expresses a lot of details of it.

The Replab project developed the *Reputation online*. Spina, Gonzalo, and Amigó (2014) added Twitter signals to content sig-

[8]Latent Dirichlet Allocation is a opinion model used for Opinion Mining

[9]The OpeNER project can be consulted at `http://www.opener-project.eu/`.

nals to improve topic detection and they learnt a similarity function in order to supervise the topic detection clustering process. The last aim of this work is to solve the reputation monitoring problem automatically. They made use of different entities (*Maroon 5, Yamaha, Ferrari*) in the task. The difference with the current work is that their text is unordered (tweets) while ours is ordered (literary criticism).

Our method is a combination of two approaches: based on discourse structure and based on aspect. On the one hand, our approach is based on discourse structure because we will use RST in order to put different weights to Elementary Discourse Units (EDUs) according to their position in a discourse tree. On the other hand, our approximation is also aspect-based because we want to identify the words related with the main topic in order to get better results.

We think that the implementation of discourse structure together with the QWN-PPV method can improve the results. In other words, the polarity of the text will be better assessed. In this paper, we will do a first approach analyzing discourse topic and its influence on structures of attitude. In the following Example (4), we show the results of the QWN-PPV tool on the whole AIZ02 text.

(4) Number of words containing sentiment found: 7
Polarity score: 0.22
Polarity (if threshold > 0.0): positive
Gustura$_{(+)}$ *irakurtzen da nobela, protagonistaren*$_{(+)}$ *joko*$_{(+)}$ *bikoitza nola bukatuko ote den, nahiz eta amaiera horren zantzuak aurretik eskaintzen*$_{(+)}$ *dizkigun*$_{(+)}$ *idazleak. Idazkerak ere laguntzen*$_{(+)}$ *du aurrera plazerez*$_{(+)}$ *egiten.* AIZ02
The novel is read with pleasure$_{(+)}$, how is going to finish the$_{(+)}$ double$_{(+)}$ game$_{(+)}$ of$_{(+)}$ the$_{(+)}$ protagonist$_{(+)}$, although the narrator previously gives$_{(+)}$ us$_{(+)}$ some clues of the ending. Writing also helps to go along with pleasure$_{(+)}$.

The QWN-PPV method determines any positive word with a positive value (+1) and any negative word with a negative value (−1). Then, a polarity score (0.22) is estimated for the text. To do so, both positive and negative words are counted and divided by the number of the words in the text. If there are more positive words, as in Example (4), the polarity score will be higher (0.23) than the threshold (zero) and, therefore, the overall polarity of the text will be positive.

On the other hand, the QWN-PPV method estimates a lower polarity score (0.11) for all the AIZ02 text. So, the example shows how coherence relations related to the discourse topic can contribute to a better assignment of the text polarity.

3 Methodology

We have used the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to achieve the rhetorical information of the text. The main concepts of RST are nuclearity of text spans[10] and coherence relations among text spans. With these two concepts a hierarchical tree structure (RS-tree) of coherence can be build, where all the text spans have a function in the tree, because relations are recursive (one relation can work as one of the spans in another relation). RST relations are hypotactic —hierarchical relations with one nucleus (N) and one satellite (S), e.g.: ELABORATION, JUSTIFY, EVALUATION, CAUSE...— and paratactic —discourse coordination where all the discourse units are nucleus, e.g. CONTRAST, DISJUNCTION, CONJUNCTION...[11]

In a hierarchical RS-tree there is always a Central Unit which is the most salient EDU (Iruskieta, Diaz de Ilarraza, and Lersundi, 2014). For example, in scientific abstracts authors often show explicitly the discourse topic as follows: "the principal aim of this paper is to investigate..." (Paice, 1980).

To show an example of the discourse structure we want to use, a partial RS-tree of AIZ02 in Figure 1 is presented. The central unit of the tree structure is represented with straight vertical lines (the unit 2-2 in the example). The annotator interpreted the RS-tree as follows:

a) PREPARATION for the article, by means of the title ([1−1 > 2−10]).

[10] Discourse structure is recursive so there are formally two different units: *a*) Elementary Discourse Unit and *b*) group of EDUs.

[11] A more detailed explanation of RST can be found at http://www.sfu.ca/rst/ and in Basque at http://www.sfu.ca/rst/07basque/index.html.

b) with the highest EVALUATION linked to the central unit she interprets that it is evaluating all the propositions mentioned before, that can be taken as all the work ([2−14 < 15−20]).

c) with the lowest EVALUATION linked to the central unit, she is evaluating an aspect of the work, only the proposition mentioned in the central unit ([6−6 < 7−7]).

These are the steps taken to carry out this study:

i) Building a corpus. We have collected a corpus of 28 texts, where 19 of them will be used for training and the remaining 9 for testing.[12] The texts are reviews of Basque Literature works. The size of the texts is not uniform: the shortest one has 106 words and the longest one 485 words. A corpus description is presented in Table 1 and the annotated corpus can be consulted in the Basque RST Treebank at `http://ixa2.si.ehu.es/diskurtsoa/en/` (Iruskieta et al., 2013).

Text	Doc.	EDUs	Words
CRITICS	28	1038	8823

Table 1: Corpus description

ii) EDU segmentation of texts. Before preparing the experiment, we have processed the texts. Firstly, we used EusEduSeg (Iruskieta and Zapirain, 2015) a discourse segmenter to segment all the texts automatically.[13] After that, the segmentation has been corrected manually to avoid losing rhetorical information in subsequent phases.

iii) Corpus annotation. After segmentation, we have annotated the most salient EDU or the central unit and after it we have tagged all rhetorical structures of the text with the RSTTool (O'Donnell, 1997) using the Basque extended relations of RST.

iv) Central Unit (CU) and Polarity gold standard. To do so, we have made up a questionnaire based on Google Forms, where 20 annotators participated in the annotation. This was done in order to have a gold standard which we could use to compare the results.
Our gold standard was collected as follows: *i*) the central unit must be selected at least by four participants. If not, we selected the three most voted central units.[14] *ii*) The polarity of each text was conformed with the average of all the annotators, in two ways:
- Polarity 1 (P1): quantitative polarity annotation from 1 to 5.
- Polarity 2 (P2): qualitative polarity description with three values: negative, neutral and positive.

v) Manual extraction of text spans composed with the text of the central unit and the EVALUATION relation. We have manually built different features based on the rhetorical structure tree:
- ALL (F1): the result of QWN-PPV on the full text.
- CU (F2): only the central unit.
- CU-H-EV (F3): the central unit and the highest EVALUATION relation linked to it.
- CU-ALL-EV (F4): the central unit and all the EVALUATION relations linked to it.

Table 2 shows all the glosses we have used to perform the analysis.

Gloss	Meaning
P1	Polarity of five categories
P2	Polarity of three categories
F1	QWN-PPN for all the text
F2	The central unit
F3	The central unit and the highest EVALUATION relation
F4	All the EVALUATION relations of the text

Table 2: Glosses of the predicted variables and features.

vi) Lemmatization and polarity extraction with the QWN-PPV. Before running QWN-PPV, we run Eustagger (Ezeiza et al., 1998) to divide the sentences into unambiguous tokens.[15]
After that, we run the QWN-PPV

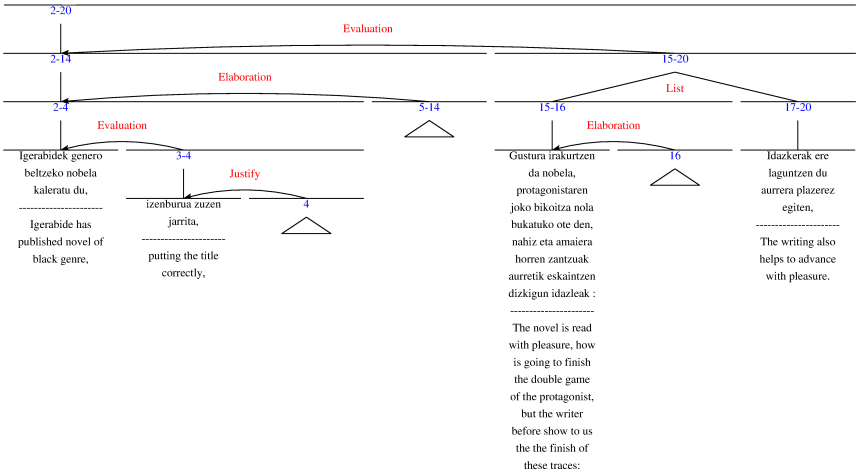
[12] All the texts are available in *Kritiken Hemeroteka* at `http://kritikak.armiarma.eus/`

[13] EusEduSeg can be tested at `http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl`.

[14] Hearst (1997) considered that a subtopic boundary was true if at least three out of seven (42.86%) annotators placed a boundary mark.

[15] Eustagger is a lemmatizer and tagger for Basque based on Stochastic and Rule-Based Methods. Eustagger can be tested at `http://ixa2.si.ehu.es/demo/analisimorf.jsp`.

Figure 1: A partial RS-tree of AIZ02



method (Vicente, Agerri, and Rigau, 2014) and obtained the polarity for each of the four features.

vii) Analysis of results. We have used two methods in order to analyze the results: *Logistic Regression* (LR) and *Sequential Minimal Optimization* (SMO).
They are well known though efficient techniques that have often been used as baseline. The first is adequate for regression problems as in this case, where the P1 class is a numeric polarity from 1 to 5. The second one tackles classification, that is, the class to guess (P2) is nominal and it was obtained by a straightforward discretization mechanism (positive, negative and neutral). Both methods are implemented with open libraries. We calculate percent agreement and precision, recall and f-measure as follows:

$$precision = \frac{correct_{polarity}}{correct_{polarity} + excess_{polarity}}$$

$$recall = \frac{correct_{polarity}}{correct_{polarity} + missed_{polarity}}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

where $correct_{polarity}$ is the number of correct *polarity items*, $excess_{polarity}$ is the number of overpredicted *polarity items* and $missed_{polarity}$ is the number of *polarity items* the system missed to tag.

A summary of the methodology we have employed is presented in Figure 2.

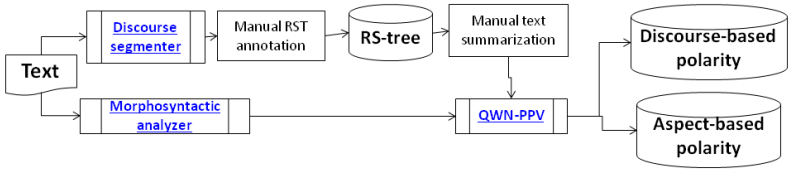
4 Results

Firstly, we present the results of all the features when trying to guess P1 (five categories). The results of each feature and the best combinations are presented in Table 3.

When individual features were considered, F1 with LR obtained the best results (0.37), while F2, F3 and F4 obtain lower results, with F4's contribution near to zero. But when combinations were considered, using features F1, F3 and F4 together (F134) with SMO obtained a better result (0.40). The results show that, in guessing P1, there is a gain employing combinations of features based on discourse structure.

Secondly, we will test the same algorithms to try to guess P2, based on three categories. The results are presented in Table 4. When using individual features, F1 and F2 with LR obtain the best results (0.47). Looking at the combinations, F1234 with SMO gives the

Figure 2: System architecture



Method	Feature	Fm
LR	**F1**	**0.37**
	F2	0.21
	F3	0.32
	F4	-
	F134	0.28
	F1234	0.30
SMO	**F1**	0.23
	F2	0.19
	F3	0.27
	F4	-
	F134	**0.40**
	F1234	0.38

Table 3: Results guessing P1 (five categories, cross-validation on the development set).

Feature	Method	Fm
LR	**F1**	**0.47**
	F2	**0.47**
	F3	0.44
	F4	-
	F134	**0.52**
	F1234	0.39
SMO	**F1**	0.37
	F2	0.36
	F3	0.36
	F4	-
	F134	0.50
	F1234	**0.53**

Table 4: Results on P2 (three categories, cross-validation on the development set).

best results (0.53). Overall, the results show that when using discourse structure (combinations), the results on P2 improve considerably.

To sum up, in the case of P1 with five categories (see Table 3) SMO is the best single algorithm for prediction. In contrast, SMO gave the best results when using a combination of features, with an F-measure of 0.40.

In the case of P2 with three categories (see Table 4), LR continues to be the best method using a single feature with an F-measure of 0.47. In combinations, SMO gives the best results (0.53).

After examining the results on the development set, we test the best methods (LR on the individual features and SMO on the combinations) on the test set. We show the results in Table 5.

	Feature	Method	Fm
P1	**LR**	**F1**	0.09
	SMO	**F134**	0.09
P2	**LR**	**F1**	0.59
	LR	**F134**	**0.84**
	SMO	**F1**	0.40
	SMO	**F1234**	0.73

Table 5: Test set results for P1 and P2

When guessing P1, the best results are obtained with F1 using LR and H134 using SMO. That means that the algorithms based in discourse structure we have used are not able to guess P1 accurately, possibly because the small size of the corpus to deal with five categories. In contrast, the results to guess P2 (three categories) using combinations of features based on discourse structure are considerably better than considering all the text (F1), with 0.84 and 0.73 for LR and SMO, respectively. Therefore, it seems that the implementation of discourse features can improve the results in opinion mining.

Performing a first error analysis, the confusion matrix of F134 guessing P2 with SMO shows that a text with neutral polarity has been classified as having a negative polarity. The error is not specially important, as a matter of fact, the QWN-PPV method puts only positive or negative polarity to the texts. So, the difference is that the method has two main categories and we use three categories.

If we analyze Example (5), we can see that the Central Unit of the text is neutral but the method considers it a negative text.

(5) *Aho gustu*$_{(+)}$ *gazi-gozoa utzi*$_{(*)}$ *dit*$_{(+)}$ *[...]ren bigarren ipuin liburuak.* [...] BER03
The second storybook of [...] has$_{(+)}$ left$_{(*)}$ me a sweet-and-sour taste$_{(+)}$ in the mouth.[...]

In the example, the word *gazi-gozoa* 'sweet-and-sour' is a neutral word but the QWN-PPV does not detect it. Moreover, some words of the remaining text are not tagged with their correct polarity.

5 Conclusions and future work

We have presented a set of algorithms in which we have tried to examine the importance of discourse structure information in Opinion Mining. Firstly, we have concluded that guessing the polarity of the text of literary reviews based on three categories gives better results than the one with five categories for Opinion Mining. This could be due to the fact that the task is easier and also to the reduced size of the training set.

The second conclusion is that combining several discourse structures is the best method for Logistic Regression giving an F-measure of 0.84 (and also for SMO with an F-measure of 0.73). An error analysis has shown that the errors were soft: a text with neutral polarity has been misclassified as having negative polarity. Looking at the importance of each individual feature, we can say that an important weight related to polarity is situated in the EVALUATION discourse relation.

In future works, our aim is to:

- Annotate a bigger corpus. The perliminary experiments performed in this work should be validated using a bigger corpus.
- Implement a full set of experiments on combinations of the central unit and the EVALUATION relation. We want to study if there is any difference with the relations which are attached to the central unit and the relations which are not.
- Test other phenomena of discourse structure such as the nuclearity (satellite vs. nucleus) and INTERPRETATION relations, to check if they have any influence on polarity.
- Automatize all the system, by testing in partial RS-trees, where only the central unit and EVALUATION relations linked to it were considered.
- Study which syntactic and discourse structures are more important (i.e., they change the polarity of lower levels).
- Implement an automatic annotator of word level polarity based on a supervised dictionary, to solve some problems observed in the QWN-PPV.

Bibliografía

[Alistair and Diana2005] Alistair, Kennedy and Inkpen Diana. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*.

[Asher1993] Asher, Nicholas. 1993. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

[Cambria et al.2013] Cambria, Erik, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. In *IEEE Intelligent Systems*, volume 28, pages 15–21, Piscataway, NJ, USA, March.

[Esuli and Sebastiani2010] Esuli, A. and F. Sebastiani. 2010. SentiWordNet: A publibly available lexical resource for opinion mining. In *Proceedings of 5th International Conference on LREC*, pages 417–422.

[Ezeiza et al.1998] Ezeiza, Nerea, Itziar Aduriz, Iñaki Alegria, Jose Mari Arriola, and Ruben Urizar. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *COLING-ACL'98*, volume 1, pages 380–384, Canada, August 10-14.

[Hatzivassiloglou and McKeown1997] Hatzivassiloglou, Vasileios and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for*

computational linguistics, pages 174–181. Association for Computational Linguistics.

[Hearst1997] Hearst, M. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. In *Computational Linguistics*, volume 23, pages 33–64, March.

[Heerschop, Goossen, and Hogenboom2011] Heerschop, B., F. Goossen, and A. Hogenboom. 2011. Polarity Analysis of Texts using Discourse Structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1061–1070, Glasgow, Scotland, UK, October 24–28.

[Iruskieta et al.2013] Iruskieta, Mikel, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.

[Iruskieta, Diaz de Ilarraza, and Lersundi2014] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475. Dublin City University and ACL, Dublin, Ireland.

[Iruskieta and Zapirain2015] Iruskieta, Mikel and Beñat Zapirain. 2015. EusEduSeg: A Dependency-Based EDU Segmentation for Basque. In *SEPNL*, Alicante, September 16-18.

[Lim and Buntine2014] Lim, Kar Wai and Wray Buntine. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1319–1328. ACM.

[Mann and Thompson1987] Mann, William C and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

[Mann and Thompson1988] Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. In *Text*, volume 8, pages 243–281.

[Miltsakaki et al.2005] Miltsakaki, Eleni, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005), Barcelona, Spain, December*.

[O'Donnell1997] O'Donnell, Michael. 1997. Rst-tool: An rst analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, Duisburg, Germany.

[Opener2013] Opener. 2013. OpeNER annotation guidelines. Annotation of Costumer Reviews in the Touristic Domain V5.0. pages 1–19.

[Paice1980] Paice, Chris D. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *3rd annual ACM conference on Research and development in information retrieval*, pages 172–191, Cambridge, June. Butterworth and Co.

[Pang and Liu2008] Pang and N. Liu. 2008. Opinion Mining and Sentiment Analysis. In *Foundations and trends in information retrieval*, volume 2, pages 1–135.

[Pang and Lee2004] Pang, B. and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimun cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271.

[Pang, Lee, and Vaithyanathan2002] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Polanyi and Zaenen2006] Polanyi, Livia and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect*

in text: Theory and applications. Springer, pages 1–10.

[Somasundaran et al.2009] Somasundaran, Swapna, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 170–179. Association for Computational Linguistics.

[Soricut and Marcu2003] Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

[Spina, Gonzalo, and Amigó2014] Spina, Damiano, Julio Gonzalo, and Enrique Amigó. 2014. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 527–536. ACM.

[Taboada, Voll, and Brooke2008] Taboada, M., K. Voll, and J. Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. In *Simon Fraser Univeristy School of Computing Science Technical Report*.

[Taboada and Mann2006] Taboada, Maite and William C Mann. 2006. Applications of rhetorical structure theory. *Discourse studies*, 8(4):567–588.

[Trnavac and Taboada2014] Trnavac, Radoslava and Maite Taboada. 2014. Discourse structure and attitudinal valence of opinion words in sentiment extraction. *LSA Annual Meeting Extended Abstracts*.

[Tumasjan et al.2010] Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. T. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*, number 10, pages 178–185.

[Vanzo, Croce, and Basili2014] Vanzo, Andrea, Danilo Croce, and Roberto Basili. 2014. A context-based model for Sentiment Analysis in Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2345–2354, Dublin, Ireland, August.

[Vicente, Agerri, and Rigau2014] Vicente, Iñaki San, Rodrigo Agerri, and German Rigau. 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, April 26-30.

[Yu and Hatzivassiloglou2003] Yu, Hong and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

[Zhou et al.2011] Zhou, Lanjun, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.

[Zhou2013] Zhou, Yudong. 2013. Fine-grained Sentiment Analysis with Discourse Structure. In *Master Thesis. Saarland University*, October.

4.3 Article 5

Using lexical level information in discourse structures for Basque sentiment analysis

Using lexical level information in discourse structures for Basque sentiment analysis

Jon Alkorta
IXA research group
UPV-EHU
jon.alkorta@ehu.eus

Koldo Gojenola
IXA research group
UPV-EHU
koldo.gojenola@ehu.eus

Mikel Iruskieta
IXA research group
UPV-EHU
mikel.iruskieta@ehu.eus

Maite Taboada
Discourse Processing Lab
Simon Fraser University
mtaboada@sfu.ca

Abstract

Systems for opinion and sentiment analysis rely on different resources: a lexicon, annotated corpora and constraints (morphological, syntactic or discursive), depending on the nature of the language or text type. In this respect, Basque is a language with fewer linguistic resources and tools than other languages, like English or Spanish. The aim of this work is to study whether some kinds of discourse structures based on nuclearity are sufficient to correctly assign positive and negative polarity with a lexicon-based approach for sentiment analysis. The evaluation is performed in two phases: i) Text extraction following some constraints on discourse structure from manually annotated trees. ii) Automatic annotation of semantic orientation (or polarity). Results show that the method is useful to detect all positive cases, but fails with the negative ones. An error analysis shows that negative cases have to be addressed in a different way. The immediate results of this work include an evaluation on how discourse structure can be exploited in Basque. In the future, we will also publish a manually created Basque dictionary to use in sentiment analysis tasks.

1 Introduction

Sentiment analysis is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2012, p. 7).

Automatic sentiment analysis is an area in continuous development. It first started with the identification of subjectivity (Wiebe, 2000) and, after that, polarity identification and measurement of strength have become the center of new developments (Turney, 2002). The objectives of sentiment analysis are evolving as well, as different types of information are used. For instance, initially, entity- and aspect-based information was used (Hu and Liu, 2004) but, later, new types of information, such as discourse structure information, have been used (Polanyi and Zaenen, 2006).[1]

This study is the first work that examines lexical and discourse structure information for sentiment analysis of Basque. The main aim is to evaluate which discourse structures can help in polarity detection following a lexicon-based approach. Our hypothesis is that some discourse structures are more related to opinions than others and we want to identify and study how they can help in a sentiment analysis task.

The paper is organized as follows: Section 2 discusses related works. Section 3 explains the methodology of the study and Section 4 presents the results and error analysis. Finally, conclusions and future work are given in Section 5.

2 Related Work

Various studies from different theoretical approaches analyze the influence of nuclearity and some rhetorical relations in sentiment analysis tasks. For example, Zhou et al. (2011) use discursive in-

[1] See a detailed review of sentiment analysis in Taboada (2016).

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 39–47,
Santiago de Compostela, Spain, September 4 2017.

formation in Chinese to eliminate noise at the intra-sentence level, improving not only polarity classification but also the labeling of rhetorical relations at sentence level.

Wu and Qiu (2012) analyze sentiment analysis based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in Chinese texts. They split texts in segments and, then, they train weights taking into account relations and nuclearity, showing that CONTRAST, CAUSE, CONDITION and GENERALIZATION have a more important role in this task than other discourse relations. Bhatia et al. (2015) use a simpler classification of relations into CONTRAST or NON-CONTRAST, and they show that the distinction improves the results of bag-of-words classifiers using Rhetorical Recursive Neural Networks.

Chardon et al. (2013) rate documents using three approaches: i) bag-of-words, ii) partial discourse information and iii) full discourse information. The discursive approach gives the best result in the framework of Segmented Discursive Representation Theory (SDRT).

Trnavac et al. (2016) propose that a few rhetorical relations have a significant effect on polarity: CONCESSION, CONTRAST, EVALUATION and RESULT. They also conclude that nuclei tend to contain more evaluative words than satellites.

Alkorta et al. (2015) analyze which features perform better in order to detect the polarity of texts using machine learning techniques on Basque texts. Their results show that discourse structure is needed to improve results along with other types of features. They use a dictionary created by automatic means with an unsupervised method (Vicente et al., 2017). The dictionary values of their work are binary (-1 for negative polarity and $+1$ for a positive one).

In this work, we analyze which coherence relations could help to improve lexicon-based sentiment analysis, so that we can assign different weights to discourse structures following Bhatia et al. (2015) when calculating sentiment analysis for a whole text. For this task, we use the RST framework.

The main contributions of this work are: i) A fine-grained dictionary, manually created for Basque with 5 different negative values and 5 different positive ones, ranging from -5 to $+5$. ii) A study of how discourse structure interacts with this polarity lexicon.

3 Methodology

The subsections below detail the main steps followed in the present study.

3.1 Extraction of discourse structures

In the first phase, different discourse structures were compared. They will be used to determine which ones can be helpful in sentiment analysis. To extract as many discourse structures as possible, we use the corpus described in Alkorta et al. (2016), annotated for discourse relations according to RST.

The corpus contains 29 book reviews. Regarding polarity, it is a balanced corpus, with 14 positive reviews and 15 negative ones. The majority of reviews were collected from a website specialized in Basque literary reviews (Kritiken Hemeroteka).[2]

The following subcorpora were created, following some discourse constraints:

- Full text, containing all the RS-tree of the text.
- Texts extracted from central units (CU)[3] of the text.
- Text spans extracted from the CU of the text and from the central subconstituent (CS)[4] of some rhetorical relations (see Table 1).

Relation	CS	Relation	CS
ELABORATION	34	CONCESSION	2
EVALUATION	32	RESTATEMENT	2
PREPARATION	32	SUMMARY	2
BACKGROUND	13	ANTITHESIS	1
CIRCUMSTANCE	8	PURPOSE	1
INTERPRETATION	6	MOTIVATION	1
CAUSE	4	JUSTIFY	1

Table 1: Number of central subconstituents (CS) in the corpus per relation type linked to the CU.

We extracted 139 instances of rhetorical relations from our corpus. For some relations, such as ELABORATION and PREPARATION (66 of 139), we do

[2] http://kritikak.armiarma.eus/.

[3] Central units are defined as the most important EDU (Elementary Discourse Unit), and it is the main nucleus when tree structure is constructed (Iruskieta, 2014).

[4] Central subconstituents are "the most important unit of the modifier span that is the most important unit of the satellite span" (Iruskieta et al., 2015, p. 5).

not expect them to contain important polarity information, because these relations only add extra information to the central unit. In fact, Mann and Thompson (1988, p. 273) mention that in the case of ELABORATION "R(eader) recognized the situation presented in S(atellite) as providing additional detail for N(uclei). R(eader) identifies the element of subject matter for which detail is provided". Similarly, in PREPARATION "R(eader) is more ready, interested or oriented for reading N(uclei)". We did not take into account relations with low frequency (a single instance), such as MOTIVATION, JUSTIFICATION, ANTITHESIS and PURPOSE. Consequently, we will work with a subcorpus containing 69 relations, where almost half of them are central subconstituents of EVALUATION.[5]

3.2 Polarity extraction and evaluation

Polarity was extracted from all the discourse structures using a dictionary (v1.0) of words annotated with their semantic orientation: polarity (positive or negative) and strength (from 1 to 5). To do so, the Spanish SO-CAL dictionary (Taboada et al., 2011) was translated using the Elhuyar (Zerbitzuak, 2013) and Zehazki (Sarasola, 2005) bilingual Spanish-Basque dictionaries. Our dictionary contains information about grammatical categories: nouns, adjectives, verbs and adverbs.

Dictionary	Words	SO(-)	SO(+)
Nouns	2,882	1,635	1,247
Adjectives	3,162	1,733	1,429
Adverbs	652	225	427
Verbs	1,657	1,006	651
Total	**8,353**	**4,599**	**3,754**

Table 2: Characteristics of the Basque dictionary.

As Table (2) shows, the dictionary contains a total of 8,353 words. The majority of words are nouns and adjectives. In terms of polarity, there are more negative words (almost one thousand more).

We created a polarity tagger, based on this dictionary. The polarity tagger used the output of Eustagger (Aduriz et al., 2003), which is a robust and wide-coverage morphological analyzer and a Part-of-Speech tagger (POS) for Basque, to enrich the text with a POS analysis information and to assign polarity to every lemma of the dictionary that matches with the lemma and category of the text. With the aim of comparing the results of the system, a linguist annotated the polarity (positive, negative or neutral) of all the discourse structures described in Section (3.1).

Figure 1 shows a portion of the RST tree of one text (LIB28).[6] After the full RST analysis was performed for each text, we extracted the following discourse structures: i) the text of the central unit (EDU_2), as shown in Example (1), and ii) the central subconstituent of the EVALUATION ($EDU_{21,22,23,25}$), in Example (2).

(1) XIX. mendean Gasteiz inguruak izutu$_{(-3)}$ zituen Juan Diaz de Garaio Sacamantecas pertsonaia hartu$_{(+2)}$ du Aitor Aranak (Legazpi, 1963) bere azken eleberrian$_{(+2)}$. (LIB28_CU)
English: Aitor Arana (Legazpi, 1963) has taken$_{(+2)}$ in his last novel$_{(+2)}$ the character Juan Diaz Garaio Sacamantecas who scared$_{(-3)}$ the surroundings of Gasteiz in the 19th century.

(2) Hala ere, nahiko$_{(+2)}$ planoa da nobela$_{(+2)}$, erritmoa falta$_{(-1)}$ zaio eta bortxaketen kontaketak aspergarriak$_{(-3)}$ ere bihurtzen$_{(-2)}$ dira, Bestalde, alabaren ikuspuntua$_{(+2)}$ ez da batere argi geratzen$_{(-2)}$, (...). (LIB28_EVA)
English: However, the novel$_{(+2)}$ is fairly$_{(+2)}$ flat, it lacks$_{(-1)}$ rhythm, and the stories of rapes also become$_{(-2)}$ boring$_{(-3)}$. On the other hand, the point of view$_{(+2)}$ of the daughter is not clear$_{(-2)}$ (...)

The classifier then assigns polarity to each word in the dictionary, as shown in Table 3 and in examples (1) and (2). The table shows that the semantic

[5] All the reviews of the corpus were coded, assigning the domain LIB (for literature review) and a number, and each discourse structure extracted from them was also coded: CU stands for text that only contains the central unit of the text, CAUS for texts that contain CAUSE relation, INT for INTERPRETATION, ELAB for ELABORATION, CIR for CIRCUMSTANCE, BACK for BACKGROUND and finally, EVA for EVALUATION. In addition, if the same relation appears more than once in each text, we added letters (e.g., a, b, c) to each relation, to indicate their order of appearance.

[6] Size constraints prevent us from showing the entire tree.

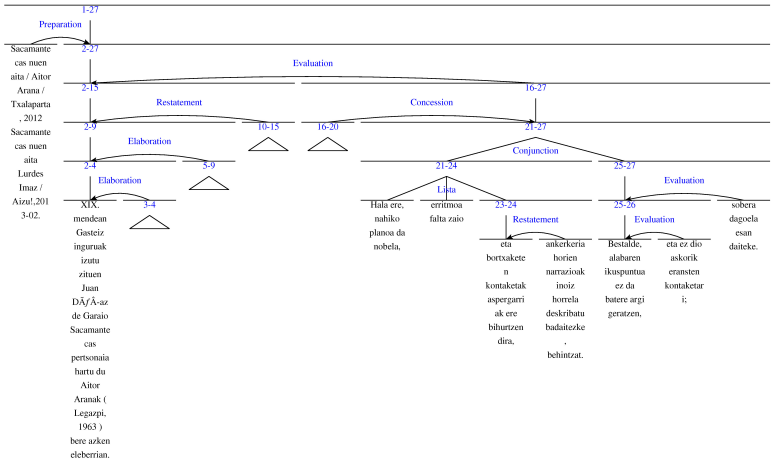


Figure 1: Central unit and the central subconstituent of EVALUATION in text LIB28

orientation of the central unit (LIB28_cu) is positive, while the semantic orientation of the central subconstituent (LIB28_EVA) is negative.

Ex.	CS ID	Classifier	SO	Manual
1	LIB28_cu	$-3+2+2$	$+1$	Neutral
2	LIB28_EVA	$+2+2-1-3-2$	-2	Negative

Table 3: Semantic orientation of LIB28_cu and LIB28_EVA: results of the classifier and of the manual annotation.

3.3 Normalization of semantic orientation results

We normalized the results obtained with the classifier to compare the different discourse structures, as in the following examples:

(3) Gure izaeraz$_{(+3)}$ hausnartzeko$_{(+1)}$ manual gisa eta, etxetik ibiltzeko$_{(+2)}$ dosi psikoanalitiko ttipi$_{(-1)}$ moduan$_{(+1)}$ hautematen$_{(+4)}$ dut nik. (LIB26_INT)
English: I consider$_{(+4)}$ it is like a manual with a small$_{(-1)}$ dose of psychoanalysis, a domestic$_{(+2)}$ consideration$_{(+1)}$ to reflect about$_{(+1)}$ our being$_{(+3)}$.

(4) Nolanahi$_{(-2)}$ den dela, saihestezina da gatazka$_{(-4)}$. (LIB13_CIR)
English: In any case$_{(-2)}$, the conflict$_{(-4)}$ is inevitable.

The results obtained by the classifier are $+112$ (LIB10),[7] $+10$ (LIB26_INT) and -6 (LIB23_CIR), as shown in Table 4. To compare those results among them, we normalized the frequencies dividing these results by the number of the words in each discourse structure. We show the normalized frequencies in Table 4.

Ex.	CS ID	SO	Words	NV
	LIB10	$+112$	418	$+0.27$
3	LIB26_INT	$+10$	17	$+0.59$
4	LIB13_CIR	-6	8	-0.75

Table 4: Examples of semantic orientation results after normalization (NV = Normalized Value).

Table 4 shows how normalization helps to better adjust the weight of the automatically assigned polarities. As a matter of fact, the values are adjusted

[7] Remember that this notation, LIB10, represents the entire text.

to a smaller range and, therefore, they are more easily comparable.

4 Results and error analysis

The results show that using a simple classifier with a manually built dictionary, along with different rhetorical structures, helps to identify the strength of such structures. For example, the result obtained in the central subconstituent of EVALUATION is strong.

(5) Guztiz$_{(+3)}$ gomendagarria$_{(+3)}$. (LIB26_EVA)
English: Highly$_{(+3)}$ recommended$_{(+3)}$.

(6) Liburu$_{(+5)}$ sano gomendagarria$_{(+3)}$ da, (LIB23d_EVA)
English: It's a very recommendable$_{(+3)}$ book$_{(+5)}$,

In Examples (5) and (6) the strength is higher than 1: $+2$ ($+6$ / 3 = 2) and $+1.6$ ($+8$ / 5 = $+1.6$), respectively, while the strength in other relations is lower.

(7) Izugarri$_{(+5)}$ gustura irakurri dut Bertol Arrietaren Alter ero narrazio bilduma. (LIB26_CAUS)
English: I have read very$_{(+5)}$ comfortable the Alter ero narration collection of Bertol Arrieta.

(8) Udako giro$_{(-2)}$ sapa horretan gertatzen diren kontakizun xumeak$_{(+3)}$ ekarriko dizkigu idazleak. (LIB15_CIR)
English: The writer will bring us the common$_{(+3)}$ stories that happen in that sticky atmosphere$_{(-2)}$ of summer.

The strength of CAUSE shows in Example (7) a value lower than 1 ($+5$ / 11 = $+0.45$). In Example (8) the central subconstituent of INTERPRETATION shows a value lower that 1 with a value of $+0.08$ ($+1$ / 12 = $+0.08$) and lower value than in Example (7).

We have analyzed the discourse structure with the aim of determining the strongest discourse structures of our corpus and therefore the structures that contribute most to improving sentiment labeling. Most of the values are between -1 and $+1$, but in 11.59% of the relations (8 of 69 relations), the values are higher than one (see Table 5).

RR	**Total**	**Total (<1)**	**%**
EVALUATION	32	6	18.75
INTERPRETATION	6	1	16.67
BACKGROUND	13	1	7.69
Others	18	0	0.00
Total	69	8	11.59

Table 5: Polarity strength ($< +1$ and > -1) of central subconstituents.

The most frequent and strongest value is obtained in EVALUATION (18.75%, 6 of 32). After that, the second strongest relation is INTERPRETATION with 16.67% (1 of 6). And, finally, BACKGROUND is once above one (7.69%, 1 of 13).

As examples (9, 10, 11, 12, 13) show, these relations have similar characteristics: short central subconstituents with many and strong evaluative words.

(9) berriz, zuzenean$_{(+3)}$ egin$_{(+2)}$ dut. (LIB14a_EVA)
English: whereas, I have done$_{(+2)}$ it directly$_{(+3)}$.

(10) Abentura$_{(+2)}$ liburu$_{(+5)}$ ederra$_{(+3)}$ iruditu$_{(+1)}$ zait, eta erremate paregabea$_{(+4)}$ trilogiarentzat. (LIB14b_EVA)
English: It seemed$_{(+2)}$ to me a beautiful$_{(+3)}$ adventure$_{(+2)}$ book$_{(+5)}$, and extraordinary$_{(+4)}$ finish for the trilogy.

(11) izenburua zuzen$_{(+3)}$ jarrita$_{(+1)}$, (LIB29a_EVA)
English: the title set$_{(+1)}$ correctly$_{(+3)}$,

(12) Intrigazko$_{(+2)}$ argumentua garatu$_{(+1)}$ nahi$_{(+3)}$ da. (LIB01b_EVA)
English: You want$_{(+3)}$ to develop$_{(+1)}$ an argument of intrigue$_{(+2)}$.

(13) Folklorean ikusi$_{(+4)}$ nahi$_{(+3)}$ ditu idazleak komunitate$_{(+1)}$ baten bizi$_{(+2)}$ nahi$_{(+3)}$ eta indarra$_{(+3)}$. (LIB35_INT)
English: The author wants$_{(+3)}$ to see$_{(+4)}$ in the folklore the strength$_{(+3)}$ and the desire$_{(+3)}$ to live$_{(+2)}$ of one community$_{(+1)}$.

Consequently, their value is higher than one, as shown in Table (6).

Ex.	**CS ID**	**NV**
9	LIB14a_EVA	1
10	LIB14b_EVA	1.36
11	LIB29a_EVA	1
12	LIB01b_EVA	1
13	LIB35_INT	1.33

Table 6: Central subconstituents and their value ($<+1$).

In contrast, we did not see any case of other central subconstituents with a value higher than one. If we compare partial discourse structures with the results obtained with all words of a text, the strength is lower in all cases. This is because polarity words do not have the same frequency in other rhetorical relations and, as a consequence, the concentration of words with semantic orientation is smaller. The highest value across the texts is $+0.50$ (LIB35), and the lowest value is -0.1 (LIB28).

These results suggest that opinions and, consequently, words with semantic orientation, are mainly found in the central subconstituent of EVALUATION, INTERPRETATION and BACKGROUND.

Apart from helping to identify the strongest central subconstituents, we have observed that the dictionary together with some central subconstituents can help in sentiment analysis. In fact, assigning a weight to some CSs could help to improve sentiment analysis results, as in text LIB34.

(14) "Behi eroak$_{(-3)}$" bilduman, ordea, egileak aurrekoan izan zituen arazoak$_{(-1)}$ konpondu$_{(+3)}$ ditu. Zoritxarrez$_{(-4)}$ bilduma honek batzuetan xelebrekeria$_{(-1)}$ merketik$_{(+3)}$ badu nahiko$_{(-2)}$. (LIB34b_EVA)
English: However, in "Behi eroak$_{(-3)}$" collection, the author has solved$_{(+3)}$ the problems$_{(-1)}$ that he had before. Unfortunately$_{(-4)}$, this collection has enough$_{(-2)}$ cheap$_{(+3)}$ eccentricity$_{(-1)}$.

The human annotator marked LIB34 as a negative review and the system assigns a value of $+0.15$ for the entire text, but a negative value of -0.2 (-5/25=-0.2) for LIB34b_EVA, Example (14). If the proper weight was assigned to this CS (LIB34b_EVA), the semantic positive orientation of the entire text (LIB34) would be corrected and tagged as negative.

We analyzed the previous finding in all the CSs of EVALUATION, but taking the results of the human annotator, instead of the classifier. In total, in 29 texts, there are 32 CSs of EVALUATION and in 24 of them, the human annotation of polarity of CSs and texts agree. So, the agreement happens in 75% of CSs and 86.20% of texts (25 texts).

Even though most of the times there is agreement between the annotated polarity of CSs and texts, this does not happen in all cases. For example, in other cases, the same text has one positive central subconstituent and another negative central subconstituent of EVALUATION. These cases are 12.50% of central subconstituents and 6.89% of texts (LIB03ab and LIB12ab).

Finally, there are two cases in which the polarity of the central subconstituent of EVALUATION and the polarity of all text are the opposite (LIB02ab and LIB19ab).

(15) eta apustu ausarta$_{(+3)}$ egin$_{(+2)}$ du bertan. (LIB19a_EVA)
English: and has made$_{(+2)}$ a strong$_{(+3)}$ bet there.

(16) Batetik, idazleak goi-literaturaren jokalekua hautatu duelako —liburuaren$_{(+5)}$ erlazio estratestualak eta baliatutako$_{(+1)}$ errekurtso andana$_{(-1)}$ lekuko—. Bestetik, borgestarretik asko duen jokoa$_{(-4)}$ delako liburuan$_{(+5)}$ dagoena. (LIB19b_EVA)
English: On the one hand, because the writer has chosen a scene from high literature —extratextual relations and a lot$_{(-1)}$ of resources used$_{(+1)}$ in the book$_{(+5)}$ as proof—. On the other hand, because there is a game$_{(-4)}$ that has a lot of Borges in the book$_{(+5)}$.

In this case, the text LIB19 is negative, whereas examples (15) and (16) are positive. We observe that the change of polarity happens in the EVALUATION situated inside an ELABORATION coherence relation.

(17) Baina, horiek horrela izanik ere, emaitza$_{(+1)}$ zalantzagarria$_{(-1)}$ da. Izan ere, liter-

aturan, baliabide$_{(+2)}$ orok medio izan behar$_{(-1)}$ du, eta irakurleak ikusi$_{(+4)}$ behar$_{(-1)}$ du errekurtsoak literaturaren mesedetan$_{(+3)}$ daudela "baita metaliteraturaz ari$_{(+2)}$ garenean ere". Hemen, ordea, medioak emaitza$_{(+1)}$ estaltzen$_{(-2)}$ du maiz$_{(+1)}$: literaturaren mekanismoekin egindako$_{(+2)}$ jokoek$_{(-4)}$ ipuinetan$_{(+2)}$ dauden istorioak$_{(-1)}$ indartu$_{(+1)}$ beharrean$_{(-1)}$, higatu$_{(-2)}$ egiten$_{(+2)}$ dituzte. Aldamioa oso$_{(+1)}$ nabarmena$_{(+4)}$ da, idazle askok beretzat nahi$_{(+3)}$ lukeen ahalmenez$_{(+2)}$ jasoa$_{(+2)}$. Haatik, hartatik sortzen$_{(+2)}$ den literatura ez da hain ikusgarria$_{(+4)}$. (LIB19_ELAB)
English: But, they being so, the result$_{(+1)}$ is doubtful$_{(-1)}$. In fact, in the literature, all resources$_{(+2)}$ need$_{(-1)}$ to be the medium, and the reader needs$_{(-1)}$ to see$_{(+4)}$ that resources are in favor$_{(+3)}$ of literary, "also when we are talking$_{(+2)}$ about metaliterature." But here, the medium hides$_{(-2)}$ the result$_{(+1)}$ in many times$_{(+1)}$: games$_{(-4)}$ made$_{(+2)}$ by literary devices wear away$_{(+2)(-2)}$ the tales$_{(-1)}$ of the stories$_{(+2)}$ instead$_{(-1)}$ of strengthening$_{(+1)}$ them. The scaffolding is very$_{(+1)}$ evident$_{(+4)}$, built$_{(+2)}$ with capacity$_{(+2)}$ as many writers would like$_{(+3)}$. However, the literature created$_{(+2)}$ is not very impressive$_{(+4)}$.

In Example (17), there are some discourse markers (*but, however*) and words (*doubtful, wear away, not very impressive*) that suggest a change of polarity that affects all text. Consequently, this example shows that, apart from central constituents of EVALUATION, a deeper analysis of nuclearity assigning different weighs could be necessary in order to improve sentiment analysis.

4.1 Error analysis

In this section, we will analyze the errors that can affect accurate detection of sentiment analysis, and specially the ones that were relevant in this study: i) errors in negative reviews, and ii) errors related to syntax.

4.1.1 Errors in negative reviews

Brooke et al. (2009) mention that lexicon-based sentiment classifiers show a positive bias because humans tend to use positive language (see also Taboada et al. (2017)). We also found this problem by examining the results of the classifier.

As Table (2) shows, the majority of the words in the dictionary are negative. Therefore, it is expected that we will detect more negative words in the texts. However, the results of the classifier with our dictionary show a tendency to classify texts as positive in different discourse structures of the texts.

For example, this tendency is observed in results of the CS of EVALUATION[8] (see Table 7).

CS of EVALUATION	Total	Guess	%
Positive	20	19	95.00
Negative	11	4	36.36
Neutral	1	0	0.00
Total	32	23	71.88

Table 7: Positive polarity tendency in central subconstituents of EVALUATION.

Table 7 demonstrates that the classifier tends to consider as positive the majority of central subconstituents of this rhetorical relation. In fact, 26 of 32 central subconstituents have been classified as positive. Consequently, the correct guess rate in CSs is higher in positive (95%) versus negative (36.36%).

A tendency to positive semantic orientation is higher if we analyze the results of all texts instead of just central subconstituents of EVALUATION as shown in Table 8.

Texts	Total	Guess	%
Positive	14	14	100
Negative	15	1	6.67
Total	29	15	51.72

Table 8: Positive polarity tendency in texts of the corpus.

As a consequence of this positive bias, our classifier guesses easily the texts with positive polarity and the correct guess rate is 100%. In contrast, the rate is very low in negative texts, as a matter of fact, there is only one right guess in text LIB28 (-0.1) and consequently, the correct guess rate is 6.67%.

[8]We have analyzed this relation and not others because it accounts for almost half of all the studied rhetorical relations.

However, if we compare the results of central subconstituents and texts, we can observe another tendency. The rate of correct assignments in positive texts is higher (95% vs. 100%) on the full texts (long text), while for negatives it is higher (36.36% vs. 6.67%) in central subconstituents (short text). This suggests that the tendency to positive semantic orientation is stronger using our dictionary as a bag-of-words approach as the text is longer.

In summary, the dictionary classifier shows the same problem already described in previous research, as there is a strong tendency towards positive semantic orientation, which increases as the text is longer.

4.1.2 Errors related to syntax

As we mentioned in Section 4.1.1, there is a tendency towards positive polarity caused by the use of positive language and, for that reason, the correct guess rate is lower in negative texts. However, it is not the only reason, and information at the syntactic level also affects the results. As an example, we will discuss one particular problem, negation. Due to negation, the polarity of a sentence is changed and it is necessary to take this characteristic into account in sentiment analysis.

(18) (...) narrazioak ere ez du arretarik bereganatzen$_{(+4)}$ (...) (LIB18_EVA).
English: (...) the narration also does not get attention$_{(+4)}$ (...)

In Example (18), the semantic orientation of the sentence would be negative but our classifier regards it as positive. The classifier has detected *bereganatu* 'to get hold of' as a positive word ($+4/7=+0.57$). But, in this case, a correct analysis should assign it a negative value.

In a first study of our subcorpus of CSs of different rhetorical relations, we estimate that this affects to 11.43% of the constituents, since 8 of 70 CSs have some type of negation.

5 Conclusions and future work

This study has analyzed whether combining a semantic oriented dictionary with some discourse structure constraints is helpful in sentiment analysis of Basque.

The results show that i) the central subconstituents (CS) of EVALUATION, INTERPRETATION and BACKGROUND are the units with the strongest semantic orientation, and ii) the CSs of EVALUATION could help in improving semantic orientation of the texts, given that the results of the human annotation of polarity of CSs and the full text text agree in 75% of the cases.

On the other hand, error analysis has shown that there are some aspects that should be addressed: i) a tendency to positive semantic orientation, and ii) sentence and more discourse level constraints are needed.

In the near future, we plan to pursue the following aspects:

i) Do reviews have a specific discourse structure? We hypothesize that reviews have a specific structure and, consequently, the same discourse relations will be repeated with high frequency, and they will appear in the same place.

ii) How we can weigh properly the central subconstituents of EVALUATION and INTERPRETATION, and neutralize the positive tendency, to improve the results for negative reviews?

iii) Are other CSs not linked to the CU important for sentiment analysis?

Acknowledgments

We thank Arantxa Otegi for assistance with the lexicon-based polarity tagger. Jon Alkorta's work is funded by a PhD grant (PRE_2016_2_0153) from the Basque Government and Mikel Iruskieta's work is funded by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish *Ministerio de Economia, Comercio y Competitividad*.

References

[Aduriz et al.2003] Itziar Aduriz, Izaskun Aldezabal, Inaki Alegria, J Arriola, Arantza Dıaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite state applications for basque. In *EACL2003 Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

[Alkorta et al.2015] Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, and Alicia Prez. 2015. Using rela-

tional discourse structure information in basque sentiment analysis. In *SEPLN 5th Workshop RST and Discourse Studies. ISBN: 978-84-608-1989-9. https://gplsi.dlsi.ua.es/sepln15/en/node/63*.

[Alkorta et al.2016] Jon Alkorta, Koldo Gojenola, and MIkel Iruskieta. 2016. Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela , September 29 th - 30 th . Extended Abstracts. ISBN: 978 - 84 - 608 - 9305 - 9*.

[Bhatia et al.2015] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

[Brooke et al.2009] Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50–54.

[Chardon et al.2013] Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–37. Springer.

[Hu and Liu2004] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Iruskieta et al.2015] Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

[Iruskieta2014] Mikel Iruskieta. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea*.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Mann and Thompson1988] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[Polanyi and Zaenen2006] Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

[Sarasola2005] Ibon Sarasola. 2005. *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

[Taboada et al.2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

[Taboada et al.2017] Maite Taboada, Radoslava Trnavac, and Cliff Goddard. 2017. On being negative. *Corpus Pragmatics*, 1(1):57–76.

[Taboada2016] Maite Taboada. 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

[Trnavac et al.2016] Radoslava Trnavac, Debopam Das, and Maite Taboada. 2016. Discourse relations and evaluation. *Corpora*, 11(2):169–190.

[Turney2002] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

[Vicente et al.2017] Inaki San Vicente, Rodrigo Agerri, and German Rigau. 2017. Q-wordnet ppv: Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *arXiv preprint arXiv:1702.01711*.

[Wiebe2000] Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

[Wu and Qiu2012] Fei Wang1 Yunfang Wu and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. In *24th International Conference on Computational Linguistics*, page 1311.

[Zerbitzuak2013] Elhuyar Hizkuntza Zerbitzuak. 2013. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. usurbil: Elhuyar.

[Zhou et al.2011] Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.

4.4 Article 6

Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus

Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus

Jon Alkorta Ixa Group / UPV/EHU
Koldo Gojenola Ixa Group / UPV/EHU
Mikel Iruskieta Ixa Group / UPV/EHU
{jon.alkorta, koldo.gojenola, mikel.iruskieta}@ehu.eus

Abstract

Discourse information is crucial for a better understanding of the text structure and it is also necessary to describe which part of an opinionated text is more relevant or to decide how a text span can change the polarity (strengthen or weaken) of other span by means of coherence relations. This work presents the first results on the annotation of the Basque Opinion Corpus using Rhetorical Structure Theory (RST). Our evaluation results and analysis show us the main avenues to improve on a future annotation process. We have also extracted the subjectivity of several rhetorical relations and the results show the effect of sentiment words in relations and the influence of each relation in the semantic orientation value.

1 Introduction

Sentiment analysis is a task that extracts subjective information for texts. There are different objectives and challenges in sentiment analysis: i) document level sentiment classification, that determines whether an evaluation is positive or negative (Pang et al., 2002; Turney, 2002); ii) subjectivity classification at sentence level which determines if one sentence has subjective or objective (factual) information (Wiebe et al., 1999) and iii) aspect and entity level in which the target of one positive or negative opinion is identified (Hu and Liu, 2004).

In order to attain those objectives, some resources and tools are needed. Apart from basic resources as a sentiment lexicon, a corpus with subjective information for sentiment analysis is indispensable. Moreover, such corpora are necessary for two approaches to sentiment analysis. One approach is based on linguistic knowledge, where a corpus is needed to analyze different linguistic phenomena related to sentiment analysis. The second approach is based on statistics and, in this case, the corpus is useful to extract patterns of different linguistic phenomena.

The aim of this work is to annotate the rhetorical structure of an opinionated corpus in Basque to check out the semantic orientation of rhetorical relations. This annotation was performed following the *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988). We have used the Basque version of SO-CAL tool to analyze the semantic orientation of this corpus (Taboada et al., 2011).

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the theoretical framework, the corpus for study and the methodology of annotation as well as the analysis of the corpus carried out. Then, Section 4 explains the results of the annotation process, the inter-annotator agreement and the results with regard to analysis in the subjectivity of the corpus. After that, Section 5 discusses the results. Finally, Section 6 concludes the paper, also proposing directions for future work.

2 Related work

The creation of a specific corpus and its annotation at different linguistic levels has been very a common task in natural language processing. As far as a corpus for sentiment analysis is concerned, information related to subjectivity and different grammar-levels has been annotated in different projects.

Refaee and Rieser (2014) annotate the Arabic Twitter Corpus for subjectivity and sentiment analysis. They collect 8,868 tweets in Arabic by random search. Two native speakers of Arabic annotated the tweets. On the one hand, they annotate the semantic orientation of each tweet. On the other hand, they also annotate different grammatical characteristics of tweets such as syntactic, morphological and semantic features as well

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 144–152
Minneapolis, MN, June 6, 2019.

as stylistic and social features. They do not annotate any discourse related feature. They obtain a Kappa inter-annotator agreement of 0.84.

The majority of corpora for sentiment analysis are annotated with subjectivity information. There are fewer corpora annotated with discourse information for the same task. Chardon et al. (2013) present a corpus for sentiment analysis annotated with discourse information. They annotate the corpus using Segmented Discourse Representation Theory (SDRT), creating two corpora: i) movie reviews from *AlloCinéf.fr* and ii) news reaction from *Lemonde.fr*. They collect 211 texts, annotated at EDU and document level. At the EDU level, subjectivity is annotated while at the document level, subjectivity and discourse relations are annotated. Results in subjectivity show that, at EDU level, Cohen's Kappa varies between 0.69 and 0.44 depending on the corpus and, at the document level, Kappa is between 0.73 and 0.58, respectively. They do not give results regarding the annotation of discourse relations.

Asher et al. (2009) create a corpus with discourse and subjectivity annotation. They categorize opinions in four groups (REPORTING, JUDGMENT, ADVISE and SENTIMENT), using SDRT as the annotation framework for discourse. Exactly, they use five types of rhetorical relations (CONTRAST/CORRECTION, EXPLANATION, RESULT and CONTINUATION). They collect three corpora (movie reviews, letters and news reports) in English and French. 150 texts are in French and 186 texts in English. According to Kappa measure, in opinion categorization, the inter-annotator agreement is 95% while in discourse segmentation it is 82%.

Mittal et al. (2013) follow a similar methodology. By the annotation of negation and discourse relations in a corpus, they measure the improvement made in sentiment classification. They collect 662 reviews in Hindi from review websites (380 with a positive opinion and 282 with a negative one). Regarding discourse, they annotate violating expectation conjunctions that oppose or refute the current discourse segment. According to their results, after implementing negation and discourse information to HindiSentiWordNet (HSWN), the accuracy of the tool increases from 50.45 to 80.21. They do not mention the inter-annotating agreement of violating expectation conjunctions.

To sum up, this section gives us a general overview about discourse-based annotated corpora for sentiment analysis. Corpora have been made for specific aims, annotating only some characteristics or features related to discourse and discourse relations. This situation differs from our work, because our work describes the annotation process of the relational discourse structure and how the function in the rhetorical relation affect to the analysis in the semantic orientation.

3 Theoretical framework and methodology

3.1 Theoretical framework: Rhetorical Structure Theory

We have annotated the opinion text corpus using the principles of *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988; Taboada and Mann, 2006), as it is the most used framework in the annotation of discourse structure and coherence relations in Basque where there are some tools (Iruskieta et al., 2013, 2015b) to study rhetorical relations. According to this framework, a text is coherent when it can be represented in one discourse-tree (RS-tree). In a discourse-tree, there are elementary discourse units (EDU) that are interrelated. The relations are called *coherence relations* and the sum of these coherence relations forms a discourse-tree. Moreover, the text spans present in a discourse relation may enter into new relations, so relations can form compound and recursive structures.

Elementary discourse units are text spans that usually contain a verb, except in some specific situations. The union of two or more EDUs creates a coherence relation. There are initially 25 types of coherence relations in RST. In some cases, one EDU is more important than other one and, in this case, the most important EDU in the relation is called *nucleus*-unit (basic information) while the less important or the auxiliary EDU is called *satellite*-unit (additional information). Coherence relations of this type are called *hypotactic relations*. In contrast, in other relations, EDUs have the same importance and, consequently, all of them are nucleus. The relations with EDUs of same rank are called *paratactic relations*. The task that selects the nucleus in a relation is called *nuclearity*.

Hypotactic relations are also divided into two groups according to their effect on the reader. Some relations are *subject matter* and they are re-

lated to the content of text spans. For example, CAUSE, CONDITION or SUMMARY are subject matter relations. On the other hand, the aim of other relations is to create some effect on the reader. They are more rhetorical in their way of functioning. EVIDENCE, ANTITHESIS or MOTIVATION belong to this group.

Figure 1 presents a partial discourse-tree of an opinion text (tagged with the code LIB29). The text is segmented and each text span is a discourse unit (EDU). The discourse units are linked by different types of rhetorical relations. For instance, the EDUs numbered with 15 and 16 are linked by an ELABORATION relation and the EDUs ranging from 15 to 20 are linked by LIST (multinuclear relation). On the other hand, the EDU numbered 2 is the central unit of this text because other relations in the text are linked to it and this text span is not attached to another one (with the exception of multinuclear relations).

According to Taboada and Stede (2009), there are three steps in RST-based text annotation:

1- Segmentation of the text in text spans. Spans are usually clauses.

2- Examination of clear relations between the units. If there is a clear relation, then mark it. If not, the unit belongs to a higher-level relation. In other words, the text span is part of a larger unit.

3- Continue linking the relations until all the EDUs belong to one relation.

Following Iruskieta et al. (2014) we think that it is recommendable, after segmenting the corpus, to identify first the central unit, and then mark the relations between different text spans.

3.2 The Basque Opinion Corpus

The corpus used for this study is the *Basque Opinion Corpus* (Alkorta et al., 2016). This corpus has been created with 240 opinion texts collected from different websites. Some of them are newspapers (for instance, Berria and Argia) while others are specialized websites (for example, Zinea for movies and Kritiken Hemeroteka for literature).

The corpus is multidomain and, in total, there are opinion texts of six different domains: sports, politics, music, movies, literature books and weather. The corpus is doubly balanced. That is, each domain has the same quantity of opinion texts (40 per domain) and each semantic orientation (positive or negative subjectivity) has the same quantity of opinion texts per each domain (20 positive and 20 negative texts per domain). We extract preliminary corpus information using the morphosyntactical analysis tool Analhitza (Otegi et al., 2017): 52,092 tokens and 3,711 sentences.

We made preliminary checks to decide whether the corpus is useful for sentiment analysis. The opinion texts are subjective, so the frequency information of the first person should be high. The results show that the first person appearance is of 1.21% in a Basque objective corpus (Basque Wikipedia) whereas its appearance is of 8.37% in the Basque Opinion Corpus. As far as the presence of adjectives is concerned, both corpora show similar results. From all the types of grammatical categories, 8.50% of the words correspond to adjectives in Basque Wikipedia and 9.82% in the corpus for study. Other interesting features for sentiment analysis, such as negation, *irrealis blocking* and discourse markers, have also been found in the corpus.

3.3 Methodological steps

We have followed several steps to annotate the Basque Opinion Corpus using the RST framework:

	A1	**A2**	**Total**
Movie	21 + 9	9	30
Weather	10 + 5	5	15
Literature	5	20 + 5	25
Total	50	39	70

Table 1: Number of texts annotated by two annotators. The number after the sum sign indicates the quantity of texts with double annotation.

1- Limiting the annotating work. Annotating 240 texts needs a lot of work and time. For that reason, we have thought to annotate some part of the corpus initially and, if the results of the annotation are acceptable, continue with the work. Taking into account the previously described data, both annotators have worked with 70 texts (29.16%) of three different domains. 21 texts from the movie domain have been annotated by one annotator and other 9 texts have been annotated by the two annotators. 10 texts from weather have been annotated once and other 5 texts of the same domain by two annotators. Finally, 25 texts

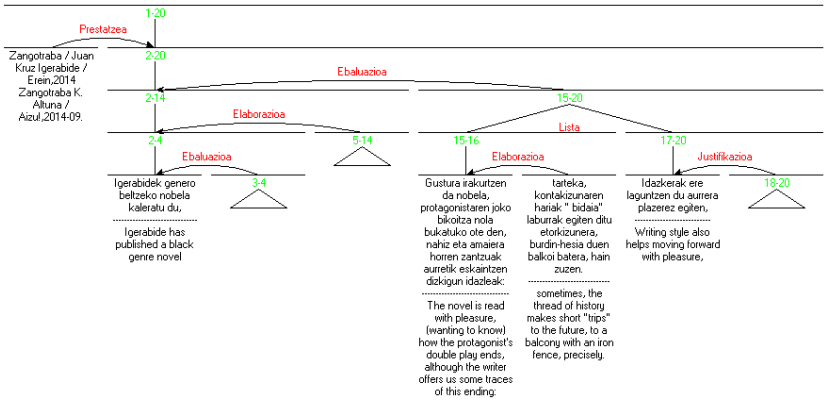


Figure 1: Part of a discourse-tree of the LIB29 review annotated with the RST framework.

of literature reviews have been annotated by one annotator and other 5 texts from the same domain by two. In total, 19 texts from 70 (27.14%) have been annotated by two annotators.

2- Annotation procedure and process. We decided to follow the annotation guidelines proposed by Das and Taboada (2018). Each person annotated four or five texts per day during two or three weeks. The time to annotate documents varied according to the domain. The texts corresponding to the weather domain are shorter and, consequently, easier to annotate while texts about movies as well as those of the literature domain are more difficult because their writing style is more implicit (less indicators and relation signals) and complex (longer at least). Approximately, each weather text was annotated in 20 minutes while movie and literature texts were annotated in one hour.

3- Measurement of inter-annotator agreement. In order to check the quality of the annotation process, inter-annotator agreement was measured. This was calculated manually following the qualitative evaluation method (Iruskieta et al., 2015a) using F-measure. In this measurement, in contrast with the automatic tool, the central subconstituent factor was not taken into account.

4- Semantic orientation extraction. Using the Basque version of the SO-CAL tool (Taboada et al., 2011), we have extracted the subjective information of rhetorical relations in the three domains of the corpus in order to check how the type of rhetorical relation affects their sentiment valence. SO-CAL needs a sentiment lexicon where words have a sentiment valence between -5 and $+5$. The Basque version of the sentiment lexicon contains 1,237 entries.

We have extracted the sentiment valence of 75 instances if CONCESSION and EVALUATION relations. From the 75 CONCESSION relations, 16 come from the weather domain, 34 from literature and 25 from movies. In the case of EVALUATION, 19 come from weather, 31 from literature and 25 from weather.

5- Results. On the one hand, we have calculated the percentage of rhetorical relations with the same label annotated by two persons. On the other hand, we have measured accumulated values of sentiment valences in nuclei and satellites in texts of different domains.

4 Results

4.1 Inter-annotator agreement

Table 2 shows the inter-annotator agreement of rhetorical relations (RR) between both annotators. This agreement was calculated following the qualitative method (Iruskieta et al., 2015a). According to these results, the highest agreement has been reached in the domain of weather where 17 of 39 relations (43.59%) have been annotated with the same relation label. After that, inter-annotator agreement in literature is 41.67% (70 from 168). Finally, the domain of movies obtained the lowest results, since the agreement is 37.73% (83 of

220). Taking all domains into account, 39.81% of the rhetorical relations have been annotated in the same way (170 relations of 427). The disagreements are due to different reasons: i) both annotators have to train more to reach a higher agreement and to obtain better results. ii) opinionative texts are more open than news or scientific abstracts. Therefore, there is more place for different interpretations.

Domain	Agreement (%)	Agreement (RR)
Weather	43.59	17 of 39
Literature	41.67	70 of 168
Movies	37.73	83 of 220
Total	**39.81**	**170 of 427**

Table 2: Inter-annotator agreement in different domains of the corpus measured by hand.

4.2 Subjectivity extraction from rhetorical relations

The annotation of the corpus using Rhetorical Structure Theory allows us to check the usefulness of the corpus. We have extracted the subjectivity from different types of rhetorical relations using the Basque version of the SO-CAL tool and we have been able to check the distribution of words with sentiment valence in each type of rhetorical relation and domain.

We have analyzed how words with sentiment valence appear in nuclei as well as satellites of CONCESSION and EVALUATION[1] in three domains. The results[2] are presented in Table 3. In the case of CONCESSION, the presence of words with sentiment valence in nuclei (47.21%) and satellites (52.79%) is similar in the three domains, although satellites show a higher proportion. In contrast, in the case of EVALUATION, words with sentiment valence are more concentrated on satellites (55.00%) in comparison with nuclei (45.00%). The only exception is weather, where nucleus prevail over satellites as far as the concentration of words with sentiment valence is concerned[3].

This information contrast between discourse and sentiment analysis provides us the option to understand what happens there. For example, in **CONCESSION**, the nucleus presents a situation affirmed by the author and the satellite shows a situation which is apparently inconsistent but also affirmed by the author (Mann and Taboada, 2005). In other words, the probability of an opinion appearance is similar in both. The sentiment valence of the nucleus prevails over the satellite but the application of Basque SO-CAL does not give the correct result because the tool does not apply any discourse processing and, consequently, in this CONCESSION relation, nuclei as well as satellite are given the same weight.

(1) [S[Puntu ahulak izan arren,]$_{-1.5}$ N[film erakargarri eta berezia da Victoria.]$_{+6}$]$_{+4.5}$ (ZIN19)
[S[Although it has weak points,]$_{-1.5}$ N[Victoria is an entertaining and special movie.]$_{+6}$]$_{+4.5}$

(2) [N[Joxek emaztea eta lagunak dauzka,]$_{-1.5}$ S[gaizki tratatzen baditu ere.]$_{-4.5}$]$_{-2.5}$ (SENTAIZ02)
[N[Joxe has a wife and friends,]$_{+2}$ S[although he treats them badly]$_{-4.5}$]$_{-2.5}$

(3) [S[Eta Redmaynen lana oso ona bada ere,]$_{+1}$ N[Vikanderrena bikaina da.]$_{+5}$]$_{+6}$ (ZIN15)
[S[Although Redmayn's work is very good]$_{+1}$, N[Vikander's is excellent.]$_{+5}$]$_{+6}$

In Example (1), the semantic orientation of the nucleus is positive while the semantic orientation of the satellite is negative. The sum is positive and, in this case, SO-CAL correctly assigns the semantic orientation of the overall rhetorical relation. In contrast, in Example (2), according to SO-CAL, the sentiment orientation of the relation is negative but it should be positive, because the semantic orientation of the nucleus is positive. This example clarifies how discourse information is needed in lexicon-based sentiment classifiers such as SO-CAL. Finally, in Example (3), the nucleus as well as the satellite and the rhetorical relation have positive semantic orientation and SO-CAL assigns correctly the semantic orientation.

Another type of rhetorical relation is **EVALUATION**, where the satellite makes an evaluative comment about the situation presented in the nucleus (Mann and Taboada, 2005). That means that the words with subjective information are more likely to appear in the satellite.

[1] We decide to choose these rhetorical relations, because we think they are more related to opinions and emotions.

[2] In order to measure the presence of words with subjectivity, we have calculated the sum of all the sentiment valences without taking into account their sign.

[3] In the weather domain, one of rhetorical relations has a very long nucleus compared to satellite. This situation may have influenced the results. In other cases, the length of nucleus and satellites has been similar.

Sum of sentiment valences	CONCESSION		EVALUATION	
	Nucleus	**Satellite**	**Nucleus**	**Satellite**
Weather	39.41	39.75	49.86	33.35
Literature	61.02	68.73	53.13	80.30
Movies	13.98	19.45	26.01	45.58
Total	114.41 (47.21 %)	127.93 (52.79 %)	128.99 (45.00%)	159.23 (55.00%)

Table 3: Accumulated values of sentiment valences in nuclei and satellites for each domain.

(4) [N[Arrate Mardarasek bere lehen liburua argitaratu du berriki, Pendrive,]$_0$ S[eta apustu ausarta egin du bertan.]$_{+3}$]$_{+3}$ (SENTBER04)
[N[Arrate Mardaras has published her first book recently, Pendrive,]$_0$ S[and she has made a daring bet there.]$_{+3}$]$_{+3}$

(5) [N[Bada, erraz ikusten den filma da "The danish girl".]$_{+1}$ S[Atsegina da, hunkigarria, entretenigarria]$_{+6}$]$_{+7}$ (ZIN15).
[N[So, "The danish girl" is a film easy to watch.]$_{+1}$ S[It is nice, touching, entertaining.]$_{+6}$]$_{+7}$

(6) [N[Talde lana izatetik pertsonaia bakarraren epika izatera pasako da erdialdetik aurrera]$_{+0.5}$ S[eta horretan asko galduko du filmak.]$_{-3.9}$]$_{-3.4}$ (ZIN39)
[N[It is going to pass from being team work to epic of one person]$_{+0.5}$ S[and in that, the film will lose a lot.]$_{-3.9}$]$_{-3.4}$

Here, we can see some specific characteristics of each rhetorical relation. Unlike CONCESSION, there is a concentration of words with sentiment valence in the satellite while words with sentiment valence have little presence in the nucleus. In fact, the sentiment valence of nuclei is never higher than $+1$ whereas satellites have a higher sentiment valence than ± 3 in all the cases. In these three Examples (4, 5 and 6), the Basque version of the SO-CAL tool guesses correctly the semantic orientation of rhetorical relations. For example, in Example (6), the semantic orientation of nucleus is positive and of satellite is negative. The sum of the two EDUs is negative and SO-CAL correctly assigns a -3.4 sentiment valence. This does not happen in all cases because the tool has not implemented any type of discourse information processing. Anyway, the tool provides information about semantic orientation that is necessary to study the relation between sentiment analysis and rhetorical relations.

5 Discussion

5.1 Inter-annotator agreement

Regarding inter-annotator agreement (Table 2), the agreement goes from 37.73% to 43.59%. However, some domains do not show regularity regarding agreement. For example, in the case of reviews (domain of literature), inter-annotator agreement is situated between 38% and 48%, except in two texts where the agreement is lower (26% and 30%). In the same line, in the weather domain, some texts show higher agreement than the average in the domain.

If we evaluate this doubly annotated corpus by automatic means in a more strict scenario (if and only if the central subconstituent is the same) following Iruskieta et al. (2015a), we can observe and evaluate other aspects of rhetorical structure, such as:

- **Constituent (C)** describes all the EDUs that compose each discourse unit or span.
- **Attachment point** is the node in the RS-tree to which the relation is attached.
- **N-S** or nuclearity specifies if the compared relations share the same direction (NS, NS or NN).
- **Relation** determines if both annotators have assigned[4] the same type of rhetorical relation to the attachment point of two or more EDUs in order to get the same effect.

Another aspect to take into consideration is that the manual and automatic evaluation does not show the same results with regard to inter-annotator agreement of the type of relation. According to a manual evaluation, inter-annotator

[4]If the central subconstituent is not described with the same span label and compared position (NS or SN), there is no possibility of comparing relations.

	Constituent		Attachment		N-S		Relation	
Domain	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**
Weather	20 of 37	0.54	9 of 37	0.24	22 of 37	0.59	15 of 37	0.41
Literature	84 of 155	0.54	67 of 155	0.43	105 of 155	0.68	48 of 155	0.31
Movies	112 of 221	0.56	88 of 221	0.40	147 of 221	0.67	68 of 221	0.31
Total	**216 of 413**	**0.52**	**164 of 413**	**0.40**	**274 of 413**	**0.66**	**131 of 413**	**0.32**

Table 4: Inter-annotator agreement results given by the automatic tool.

agreement is 39.81% while the automatic evaluation shows an agreement of 31.72%. As we have noted before, this difference comes due to the fact that the automatic comparison is made in a strict scenario and some relations are not compared, because the description of the central subconstituent of such relations is slightly different.

The inter-annotator agreement results given by the automatic tool offer complementary information related to the annotation of the corpus. As Table 4 shows, the inter-annotator agreement is low in the case of type of relation but the results are better in other aspects of rhetorical relations such as constituent and nuclearity. The agreement in attachment point achieves 0.40 that is low still but constituent as well as nuclearity have achieved the inter-annotator agreement of 0.52 and 0.66, respectively.

On the other hand, another interesting aspect is that there is no difference between domains as far as the agreement of different aspects related to writing style is concerned. It is surprising because the type and the way to express opinions are very different for each domain. In the weather domain, texts are short and clear and the language is direct. In contrast, in literature and movies, texts are longer, more diffuse and they use figurative expression many times. Even so, the weather domain obtains lowest results in three aspects mentioned in Table 4 but the type of relation obtains a better result compared to other domains.

The interpretation of inter-annotator agreement suggests that in the evaluation of some rhetorical relations the agreement is lower while other aspects related to rhetorical relations like constituent and nuclearity obtain a better agreement. We have also discovered that specially ELABORATION, EVALUATION and some multinuclear relations show higher disagreement.

5.1.1 Relevant RR disagreement: confusion matrix

In order to know the differences of these disagreements, we have also measured the type of rhetorical relations with the highest disagreement. With that aim, we have calculated a confusion matrix, and then we have identified the most controversial rhetorical relations. Results are shown in Table 5.

A1	**A2**		
RRs		**#**	**Total**
ELABORATION	MOTIVATION	9	19
ELABORATION	INTERPRETATION	6	
RESULT	ELABORATION	4	
INTERPRETATION	JUSTIFICATION	4	4
CONCESSION	CONTRAST	6	14
EVALUATION	CONTRAST	4	
LIST	CONJUNCTION	4	

Table 5: Disagreement in rhetorical relations.

According to Table 5, ELABORATION has been used by one annotator whereas the other has employed a more informative relation. In two cases, the first annotator (A1) has annotated an EVALUATION relation while the other annotator (A2) has annotated MOTIVATION and INTERPRETATION. In other case, A2 has annotated ELABORATION whereas A1 has tagged RESULT. In total, there are 19 instances in which ELABORATION has been annotated by one of the annotators. Moreover, there are 4 instances of disagreement between INTERPRETATION and JUSTIFICATION. Finally, there are also disagreements in multinuclear relations. While A2 has annotated CONTRAST in 10 relations, A1 has employed CONCESSION and EVALUATION. There are also 4 instances of disagreement between LIST and CONJUNCTION.

Our interpretation of this results is that one annotator (A1) tends to annotate more general rhetorical relations (e. g. ELABORATION) while other annotator (A2) annotates more precise relations. When it comes to multinuclear relations, it seems that A1 annotator has a tendency to not an-

notate multinuclear relations.

5.2 Checking the usefulness of the corpus for sentiment analysis

The second aim of this work has been to check the usefulness of the corpus for sentiment analysis. Firstly, the results have shown that in some cases the Basque version of SO-CAL does not assign a suitable semantic orientation to all the rhetorical relations, even when the semantic orientation of EDUs of the relation is correct. This means that the information of rhetorical relations would be needed in order to make a lexicon-based sentiment classification. In other words, this suggests that it would be recommendable to assign weights to EDUs of rhetorical relations to model their effect on sentiment analysis. Each type of rhetorical relation has different characteristics and, consequently, the way to assign weights to EDUs in each relation must be different.

For that reason, we have made a preliminary study with the purpose of checking how different types of rhetorical relations present a semantic orientation and what is the distribution of words with sentiment valence in rhetorical relations. The study of CONCESSION has shown that *i)* the probability of sentiment words appearing in nuclei as well as satellites is similar, and that *ii)* nucleus always prevails over the satellite and, consequently, the semantic orientation of nucleus must be the semantic orientation of all the rhetorical relation. However, the semantic orientation of the satellite must be also taken into consideration in the semantic orientation of all the rhetorical relation. Although comparing with nucleus, satellite has to be less important.

The opposite situation happens in EVALUATION. Here, we can see that words with sentiment valence concentrate more on the satellite while there are fewer words with sentiment valence in the nucleus. That means that the weight must be assigned to the satellite because that part of the relation is more important from the point of view of sentiment analysis.

This interpretation of the results suggests that the Basque Opinion Corpus annotated using RST can be useful for different tasks of sentiment analysis, in fact, the preliminary analysis made with rhetorical relations shows some characteristics and differences that are related to rhetorical relations.

6 Conclusion and Future Work

In this work, we have annotated a part of the Basque Opinion Corpus using Rhetorical Structure Theory. Then, we have measured inter-annotator agreement. The manual evaluation of the results shows that the inter-annotator agreement of the type of rhetorical relations is 39.81%. On the other hand, using an automatic tool we have obtained more fine-grained results regarding aspects of relations and attachment, as well as nuclearity, with an inter-annotator agreement higher than 0.5. We have also identified that ELABORATION, EVALUATION and some multinuclear relations show the highest disagreement.

On the other hand, we have also checked the usefulness of this annotated corpus for sentiment analysis and the first results show that it is useful to extract subjectivity information of different rhetorical relations. In CONCESSION relations, the semantic orientation of the nucleus always prevails but the valence of the satellite must also be taken into consideration. In EVALUATION relations, words with sentiment valence concentrate on satellite.

In future, firstly, we plan to build extended annotation guidelines to annotate the corpus with more reliability. This would be the previous step before annotating the entire corpus. On the other hand, we would like to continue analyzing how the subjective information is distributed in relations.

Acknowledgments

This research is partially supported by a Basque Government scholarship (PRE_2018_2_0033), the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) project PROSA-MED (TIN2016-77820-C3-1-R), University of the Basque Country project UPV/EHU IXA Group (GIU16/16) and project Procesamiento automático de textos basado en arquitecturas avanzadas (PES18/28).

References

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2016. Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis*, pages 58–62.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of opinion expressions in

discourse. *Lingvisticæ Investigationes*, 32(2):279–292.

Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–37. Springer.

Debopam Das and Maite Taboada. 2018. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Mikel Iruskieta, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015a. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Mikel Iruskieta, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 466–475.

Mikel Iruskieta, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2015b. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 11(2):303–334.

William C Mann and Maite Taboada. 2005. RST web site.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. 2013. Sentiment Analysis of Hindi Reviews based on Negation and Discourse Relation. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 45–50.

Arantxa Otegi, Oier Imaz, Arantza Diaz de Ilarraza, Mikel Iruskieta, and Larraitz Uria. 2017. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In *LREC*, pages 2268–2273.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Maite Taboada and Manfred Stede. 2009. Introduction to RST (Rhetorical Structure Theory). *ESSLLI2016*.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 246–253.

4.5 Article 7

The role of hierarchical structure and coherence relations in sentiment analysis: a study in Basque (preprinted)

The role of hierarchical structure and coherence relations in sentiment analysis: A study in Basque

Jon Alkorta
UPV/EHU

Maite Taboada
Simon Fraser University

Koldo Gojenola
UPV/EHU

Mikel Iruskieta
UPV/EHU

Sentiment analysis is now a mature area of study, with a number of tools and dictionaries in multiple languages. The goal of most automated systems for sentiment analysis is to determine whether a text (broadly construed, i.e., a review, a headline, a tweet) expresses positive or negative opinion towards its subject matter. The majority of the existing sentiment analysis systems follow a bag-of-words approach and, until now, syntactic and discourse constraints have been excluded, because it has been assumed that this information was not necessary or it was too complex to process adequately. We believe that contextual discourse information is crucial to accurate sentiment analysis and, to this end, the main aim of this work is to study the most relevant features of discourse relations vis-à-vis sentiment analysis. In order to add discourse features, we need to know how relational discourse structure phenomena affect or modify the polarity of text spans (what we define as semantic orientation of text spans), and then add this information to a sentiment analysis system.
Specifically, we identify the discourse structures that have a role in the interpretation of sentiment and opinion in text. We will base our study on Basque, a language that has not been widely addressed in sentiment analysis. We have annotated a corpus of book reviews in Basque with two different types of features: a) the relational structure of texts following Rhetorical Structure Theory (RST); and b) the semantic orientation (positive or negative polarity) of the annotated discourse relations. We analyze how the semantic orientation of such discourse relations interacts with different aspects of the relation (relation type, nuclearity) and with the overall text structure (position, distance from the text's central unit). The results suggest that there is evaluative prosody at the local level of discourse structure: The semantic orientation of a rhetorical relation tends to match that of neighbouring nuclei (as opposed to satellites) and it also tends to match the semantic orientation of the following spans (as opposed to previous spans).
Moreover, discourse relations near the central unit match the semantic orientation of the text overall more often than those that are farther away from the central unit, whether before or after the central unit. Finally, the results show that discourse relations are not evenly distributed across the text, but rather tend to appear in specific places in discourse trees and are sensitive to the semantics of the discourse relation in question.

Keywords: discourse relations, evaluative text, sentiment analysis, Rhetorical Structure Theory, Basque

Introduction: Sentiment analysis and discourse

Sentiment analysis aims at extracting subjectivity and opinion from language, often expressed as polarity, that is, whether the semantic orientation of a text, sentence or word is positive or negative. The most basic approach to sentiment analysis consists of using a sentiment lexicon considering the texts as 'bags-of-words' or unordered sequences of words, without taking into account syntactic or discursive features of the text (Taboada, 2016; Benamara, Taboada, & Mathieu, 2017). This technique is applied to various types of texts, such as tweets, headlines, news articles, on-line comments, reviews or performance evaluations. For instance, the polarity of the Basque sentence in Example (1) can be determined by first assigning a polarity or semantic orientation to each of the words, as shown in Example (2). The values of each of the words are extracted from a Basque sentiment lexicon (Alkorta, Gojenola, & Iruskieta, 2018).[1]

(1) Irabazi ezinik jarraitzen du Eibarrek. (KIR17)
win cannot continue AUX Eibar.
"(The soccer team) Eibar continues without winning."

(2) Ezin_0; Eibar_0; Irabazi_{+2}; Jarraitu_0
Cannot_0; Eibar_0; Win_{+2}; Continue_0

If we add up all the values of Example (2), the sentiment value of the sentence would be 0.5 (+2 divided by 4 words). In this case, the semantic orientation given by the bag-of-words approach is not completely right. The sentiment orientation of the sentence should be negative because the soccer team has not won the match. This change in the semantic orientation of the word *irabazi* +2 ('to win') is due to the negation marker *ezin* ('can not'), a contextual valence shifter (Polanyi & Zaenen, 2006) that changes the semantic orientation of words, phrases or sentences. There are many contextual valence shifters in Basque, in addition to negation markers. Some of them operate at the discourse level, whereas some others operate at the morphological or even phonological level. For example, *gozo* means "gently" referring to people which has a positive semantic orientation. Its form with expressive palatalization is *goxo* (change from [z] to [x]). It is a hypocoristic form, changing the semantic meaning and strengthening the semantic orientation because the meaning has more intensity.

The goal of this work is to describe the role of contextual discourse valence shifters in a Basque corpus annotated with discourse relations and their semantic orientation. Specifically, we focus on the role of discourse relations in the interpretation of opinion. Discourse relations hold between two non-overlapping text spans, indicating a connection between the propositions in those spans. Possible labels for those connections include Condition, Elaboration, Cause or Evaluation.

(3) [[Zahartuz zihoala, gero eta zailagoa egiten zitzaion basatikeriak egitea,]$_N$ → [[*NEG*]
[eta horri esker harrapatu zuten.]$_S$] → [POS]POS] (SENTAIZ03)
[[As he got older, it became more and more difficult for him to do wild things,]$_N$ →[[POS]
[and by that he was caught.]$_S$] → [POS]POS]

Example (3) shows a CAUSE rhetorical relation extracted from the opinion text SENTAIZ03. In the first line of the example, the nucleus-unit (N) of the relation presents a negative semantic orientation. On the other hand, the satellite-unit (S) of the relation appears in the second line with a positive semantic orientation. The union of both discourse units is made by a CAUSE rhetorical relation with positive ~~semanti~~c orientation.

[1] In the examples, the code in parentheses represents the domain (KIR = sport) and the text's number. The texts are part of a Basque corpus (Alkorta, Gojenola, & Iruskieta, 2016). Underlined words are valence shifters, i.e., expressions that can change the semantic orientation of the words in their scope.

Previous work has shown that nuclearity (the relative status of spans in a relation), the type of relation or the position of the relation in the text have an effect on the evaluative interpretation of that relation, and also on the overall polarity for the text (Taboada, Voll, & Brooke, 2008; Chardon, Benamara, Mathieu, Popescu, & Asher, 2013); see also next section. Following this work, and applying it to our corpus of Basque texts, we formulated the following research questions, which link the most significant discourse structure phenomena to their role in sentiment analysis:

1. Related to relations: Which span in the discourse relation tends to convey the semantic orientation for the entire relation? Spans in most relations are asymmetric, one being more important or salient (the nucleus) and the other one contributing to the information in the nucleus (the satellite). In other words, does the semantic orientation of the entire relation tend to match the semantic orientation of the nucleus or the satellite?

2. Related to the text: Which part of the text tends to convey the semantic orientation for the entire text? Do relations at different levels of the discourse tree contribute more or less to the semantic orientation for the entire text?

3. Related to the text: What is the relationship between central unit for the text and semantic orientation?

The work is organized as follows: after discussing related work, we describe data and methodology. We then proceed to present and discuss the results of the analysis, concluding with possible directions for future work.

Related work

Polanyi and Zaenen (2006) proposed that contextual information plays a role in the interpretation of evaluative words when performing automatic sentiment analysis. Their proposal emphasized phrase-level information, such as intensifiers (*great* in *great movie*) or negation (*not such a great movie*), but also considered conjunctions and discourse adverbs such as *although* or *however*. This naturally leads to considering the entire discourse relation that those connectives signal, and how it affects the meaning of evaluative words in the relation. For instance, Trnavac and Taboada (2012, p. 305) discuss Example (4), pointing out that the positive meaning conveyed by *fun* and *good* is tempered, as it were, by the conditional relation those words are part of.

(4) Fun is good, but only if you know when to stop.

Continuing this line of research, Trnavac, Das, and Taboada (2016) explore not only the types of relations that play a role in changing the polarity of the words they contain, but also which part, nucleus or satellite, most directly conveys evaluative meaning in discourse relations. They found that there is a group of relations that tend to intensify the polarity of evaluative words, namely Concession, Elaboration, Evaluation, Evidence and Restatement. They also observed a strong tendency for evaluative words to be found in the nucleus of a discourse relation.

Other research in sentiment analysis has made use of either manually annotated or automatically extracted discourse relations, to ascertain whether incorporating discourse information improves the results of sentiment analysis systems. Different discourse relation frameworks have been deployed in this work.

Some of this research has followed Rhetorical Structure Theory (Bhatia, Ji, & Eisenstein, 2015; Heerschop et al., 2011; Zirn, Niepert, Stuckenschmidt, & Strube, 2011), whereas a different strand employs Segmented Discourse Representation Theory (SDRT) for annotation and use in sentiment analysis systems (Chardon et al., 2013; Asher, Benamara, & Mathieu, 2008). Table 1 summarizes some of this work, by framework, type of discourse information annotated and granularity of the sentiment annotation.

Authors	Approach	Discourse information	Sentiment information
Mukherjee and Bhattacharyya (2012)	-	Discourse markers, connectives and Conditional discourse relations	Semantic orientation of tweets
Heerschop et al. (2011)	RST	Nuclearity: nucleus / satellite, RST relation types	-
Chardon et al. (2013)	SDRT	Partial discourse information, full discourse information	Semantic orientation at discourse unit and document level
Bhatia et al. (2015)	RST	Discourse unit position, Recursive neural networks, RST parser	Semantic orientation at discourse unit and document level
Asher et al. (2008)	SDRT	Relations: CONTRAST, CORRECTION, SUPPORT, RESULT and CONTINUATION	Four groups: reporting, judgment, advise and sentiment
Somasundaran, Namata, Wiebe, and Getoor (2009)	Dialog Acts	Types of frame relations (reinforcing and non-reinforcing relations)	Polarity (positive, negative, neutral) of dialogue units
Taboada et al. (2008)	RST	Nuclei, topic sentences	Semantic orientation of nuclei and texts
Zirn et al. (2011)	RST	Relations: CONTRAST (CONCESSION and CONTRAST), NON-CONTRAST.	Semantic orientation of discourse relations
Zhou, Li, Gao, Wei, and Wong (2011)	-	A set of cue-phrase-based patterns	-

Table 1
Previous work on discourse-based sentiment analysis.

Similarly, corpora used in these projects range in genre or text type, size and type of annotation. The most common text type is reviews, given the natural interest in extracting sentiment from online reviews. Other text types include tweets (Mukherjee & Bhattacharyya, 2012) or news (Chardon et al., 2013). Regarding the size of corpora, it ranges from 120 reviews in Zirn et al. (2011) to 1,000 reviews in Heerschop et al. (2011). For shorter texts, the corpus presented in Bhatia et al. (2015) consists of 32,000 tweets. Finally, some corpora are polarity-balanced, with the same number of positive and negative texts (Heerschop et al., 2011; Zirn et al., 2011).

With respect to methodology, there is a great amount of variation. Most of these projects annotate the corpus under consideration following annotation guidelines, leading to a gold standard that was created by humans (Chardon et al., 2013). Some research, on the other hand, makes use of automatic or semi-automatic methods to perform annotation. For instance, Heerschop et al. (2011) and Zirn et al. (2011) use sentiment lexicons to assign sentiment valence to words in the text. For annotation of discourse phenomena, automatic methods include discourse reweighing and Rhetorical Recursive Neural Networks (Bhatia et al., 2015), discourse-based bag-of-words models (Mukherjee & Bhattacharyya, 2012) and Sentiment Classifier Discourse Parser (Zirn et al., 2011).

Once a corpus is annotated with both sentiment and discourse information, the crucial question is whether the discourse information helps in automatically determining the sentiment. This is the evaluation process. In some cases, the evaluation measures the effects of linguistic phenomena. For instance, Chardon et al. (2013) use a bag-of-segments and discourse information, Mukherjee and Bhattacharyya (2012) evaluate a lexicon-based classification with and without discourse information, Heerschop et al.

(2011) evaluate the effect of word position and, finally, Zirn et al. (2011) evaluate sentence classification with nucleus or satellite information with RST relations. On the other hand, statistical approaches have used Support Vector Machine models (Mukherjee & Bhattacharyya, 2012) and Markov logic networks (Zirn et al., 2011). In most cases, the use of discourse information leads to positive results, i.e., an increase of accuracy for sentiment analysis.

This work presents similarities in several aspects with Bhatia et al. (2015) and Heerschop et al. (2011) in the sense that, like us, they use discourse information to detect the most important text spans and apply different treatments to them. The main difference lies in the main objectives of each work. Bhatia et al. (2015); Heerschop et al. (2011) use this approximation to check if their results improve upon previous work. In contrast, our aim is to analyze the interaction between rhetorical relations and sentiment analysis. In other words, we want to check whether some discourse structures affect the sentiment valence or semantic orientation of discourse relations or what the effect of nuclearity is in different text spans. This information can, naturally, be deployed in sentiment analysis, but our first goal is to explore this relationship, before we can decide how best to implement this knowledge.

Our work is unique in that it explores the interaction of sentiment and discourse for Basque, a relatively understudied language. Existing work has approached Basque from the point of view of RST (Iruskieta, 2014) and, separately, from the point of view of sentiment analysis (San Vicente, Saralegi, & Agerri, 2015; Alkorta et al., 2018). Some of our previous work has combined the RST framework and sentiment analysis (Alkorta, Gojenola, Iruskieta, & Pérez, 2015; Alkorta, Gojenola, Iruskieta, & Taboada, 2017). In this paper, we pursue this relationship more rigorously.

Data and methodology

In this section, we will present: 1) a description of the corpus and the annotation of discourse relations, 2) a parameter analysis, 3) the method for comparison of parameters and 4) the reliability and evaluation measures.

1) **The corpus and annotation of discourse relations.** The study of discourse relations and sentiment in Basque is based on 28 reviews about literature. These reviews are part of the Basque Opinion Corpus (Alkorta et al., 2016) and they are also available in the Basque RST Treebank[2] (Iruskieta et al., 2013).

Discourse relation	**Instances**
ENABLEMENT/MOTIVATION	6
CONDITIONAL subgroup	18
CONTRAST	26
EVIDENCE/JUSTIFY	53
CONCESSION/ANTITHESIS	70
CAUSE subgroup	71
EVALUATION/INTERPRETATION	140
Total	**384**

Table 2
Discourse relations of the study.

[2] http://ixa2.si.ehu.es/diskurtsoa/index.php.

We first extracted the discourse relations in Table 2 from the reviews. In total, 384 discourse relations of seven different types were extracted. According to the Classical Mann and Thompson extended classification (Mann & Thompson, 1988), there are 12 types of discourse relations[3] but we decided to take seven of them into account. The rationale behind this decision was that the selected discourse relations were more likely to have subjective content in comparison with others. Moreover, Trnavac et al. (2016) show that CONCESSION, ELABORATION, EVALUATION, EVIDENCE and RESTATEMENT most frequently intensify the polarity of the opinion words they contain. After the extraction, two linguists assigned the semantic orientation of these reviews, the text spans of the discourse relations and the discourse relations.

2) **Measuring the inter-annotator agreement of the semantic orientation.** The semantic orientation of discourse relations and their spans were annotated by one person. We did, however, take a subset of the corpus and calculated Cohen's kappa coefficient between two linguists, to test the level of reliability that could be achieved in such annotation.

For the reliability study, two linguists assigned semantic orientation to discourse relations and their spans. This involved three different types of semantic orientation per discourse relation: i) semantic orientation of the discourse relation as a whole, ii) semantic orientation of the nucleus, and iii) semantic orientation of the satellite (except in multinuclear relations). In total, 40% of the corpus was annotated by both linguists. The level of inter-annotator agreement, measured using Cohen's kappa, was 0.58, which is considered as 'moderate agreement' according to Landis and Koch (1977).

		A2			
		NEG	**NEU**	**POS**	**Grand Total**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Grand Total	92	120	181	393

Table 3
Contingency table of two annotators regarding the semantic orientation annotation of discourse relations.

In Table 3 we show a contingency table of confusion across the two annotators. We can see that the largest differences are related to the neutral semantic orientation. When one annotator (A2) assigned a neutral semantic orientation, the other annotator (A1) assigned a positive or negative one. This occurs in 120 instances. The opposite case also occurred, but with less frequency (98 instances). This is somewhat expected, as we have a tendency to conflate subjectivity (i.e., semantic orientation) with either positive or negative polarity. Neutral semantic orientation is more difficult to identify accurately (see Example (5)).

(5) [Idazleak berak esan bezala, "gertaera xumeak" dira,]$_1$
[baina hala ere, arazoak euren onetik ateratzen ditu pertsonaiak.]$_2$
[As the writer himself says, these are "humble facts."]$_1$
[but even so, problems drive people up the wall.]$_2$

[3] The 12 types of discourse relations according to Rhetorical Structure Theory (RST) are the following: CIRCUMSTANCE, SOLUTIONHOOD, ELABORATION, BACKGROUND, ENABLEMENT and MOTIVATION group, EVIDENCE and JUSTIFY group, relations of CAUSE, ANTITHESIS and CONCESSION, CONDITION and OTHERWISE, INTERPRETATION and EVALUATION, RESTATEMENT and SUMMARY, and SEQUENCE and CONTRAST.

In Example (5), one annotator has considered all the rhetorical relation and its spans neutral while the other annotator does not consider it. According to one annotator, the semantic orientation is positive and negative for the first and second discourse units, respectively. Besides, the semantic orientation of all the rhetorical relations as a whole is negative. In contrast, the second annotator has considered the semantic orientation of the first and second discourse units and all the rhetorical relation neutral.

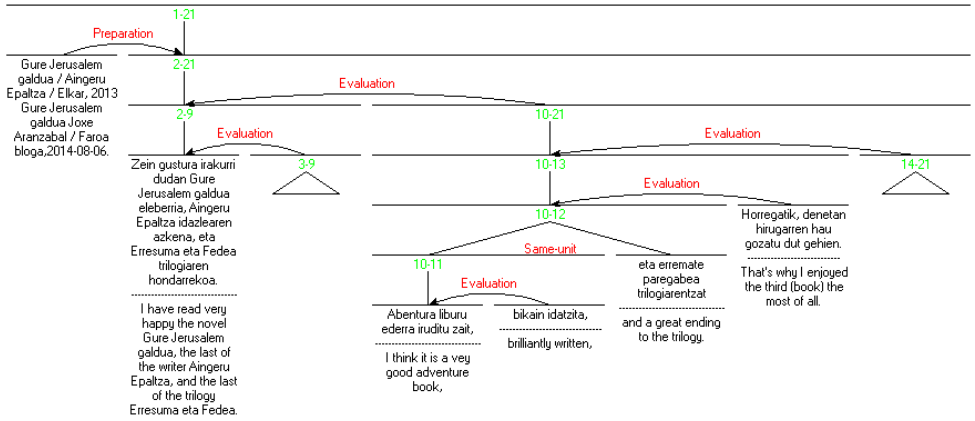


Figure 1. RST-tree for text SENTFAR-01.

3) **Analysis of parameters.** We use Figure 1 and the EVALUATION relation (EDU 10-13) as an example for illustration purposes. After the extraction process, discourse relations were analyzed with the following parameters:

   i) Nuclearity. The discourse relation can be mononuclear or multinuclear. In the first case, one of the units of the discourse relation is more important than the other, in the sense that it contributes more to the effect that the relation intends to achieve on the listener/reader. The smallest units of discourse that enter into a discourse relation are usually referred to as Elementary Discourse Units (EDUs). The most important EDU is labelled as nucleus, with the other unit receiving the satellite label. Relations that are composed of a nucleus and satellite can appear in two different patterns: Nucleus-Satellite or Satellite-Nucleus. In addition, some relations present two or more EDUs as having the same importance; those are referred to as multinuclear relations. In Figure 1, the span 11 is linked with a EVALUATION relation to the span 10 and follows the N(ucleus)-S(atellite) pattern. The span 13 follows the same structure and it is linked to span 10-12 with an EVALUATION relation.

   ii) Semantic orientation to text spans and discourse relations. We assigned a semantic orientation[4] to different text spans and discourse relations. First, we assigned the semantic orientation to text spans. In the example, the text spans 10-12 and 13 both have a positive semantic orientation. This means that the EVALUATION relation links two positive discourse units

[4] We assigned three types of semantic orientation to discourse units and discourse relations: POS, positive; NEG, negative and NEU, neutral.

(10-12 and 13). Then, we assigned the semantic orientation to the entire discourse relation (positive).

iii) Distance to central unit. The central unit is the most important nucleus in the text, one that most directly contributes to the intention or effect that the text intends to achieve (Iruskieta, da Cunha, & Taboada, 2015; Marcu, 2000). We measured the distance between discourse relations and central unit counting the number and types (NN or NS) of discourse relations between the central unit and the current discourse relation (horizontal distance). As far as the span 13 is concerned, the distance in discourse relations is +2 in Figure 1, because there are 2 NS discourse relations (2 NS EVALUATIONs) between span 13 and span 2, which is the central unit, we give a distance of +2.

iv) Type of discourse relation. The last parameter used in our study was the type of relation. As Table 2 shows, in total seven types of discourse relations were considered.

The EVALUATION discourse relation joins the 10-12 and 13 EDUs. That means that the 10-12 EDU explains a situation and EDU 13 makes an evaluative comment about it.

Results

In this section, we present the results trying to address our main research questions. We first explore our conclusions regarding nuclearity and sentiment analysis, to then approach the semantic orientation in texts. We end this section with a discussion of the position of discourse relations in opinion texts.

Semantic orientation, nuclearity and position in discourse relations

Tables 4 and 5 present the breakdown of semantic orientation with regard to nuclearity and position. Our first question is whether the semantic orientation of the entire relation is more often that of the nucleus or the satellite. For instance, if the relation is positive overall, is it more often the case that the nucleus is also positive, or the satellite? Table 4 shows that, compared to satellites, the nuclei coincides in more cases in semantic orientation with the whole discourse relation they belong to (0.71 in the case of nuclei and 0.64 in the case of satellites). However, this tendency does not happen in all types of discourse relations. In the case of EVALUATION, the semantic orientation of the rhetorical relation is more often the same as the semantic orientation of the satellite (0.82 for the satellite vs. 0.69 for the nucleus).

(6) [[Igerabideren estiloa aberatsa baita,]$_N$ → [[POS]
[egoerari eta pertsonaiei ederki egokitua,]$_S$] → [POS]POS] (SENTAIZ02)
[[The style of Igerabide is rich,]$_N$ →[[POS]
[beautifully adapted to the situation and the characters.]$_S$ → [POS]POS]

In Example (6), there is an annotation of the EVALUATION rhetorical relation in the text SENTAIZ02. In the first line of the example, there is a nucleus-unit with a positive semantic orientation. In the second line, the semantic orientation is also positive in satellite-unit and consequently the semantic orientation of the rhetorical relation is positive.

By contrast, the CONCESSION group shows the largest difference between nuclei and satellites, with a 0.70 match for nuclei, 0.33 for satellites. Another aspect is where the highest SO match between discourse relations and nuclei or satellites appears, the results showing that nuclei in CAUSE and the EVIDENCE group match most (0.83). In the case of the CONTRAST relation, the semantic orientation of the discourse relation with one nuclei is high (0.73) but much less frequent with both nuclei (0.31).

Relations	Instances	N	S
CAUSE	71	0.83	0.66
CONCESSION group	70	0.70	0.33
CONDITIONAL	18	0.61	0.56
ENABLEMENT group	6	0.60	0.5
EVALUATION group	140	0.69	0.82
EVIDENCE group	53	0.83	0.64
CONTRAST*	26	0.73	0.31
Total	384	0.73	0.65

Table 4
Agreement of the semantic orientation of discourse relations with nucleus and satellites.

Relations	Instances	First span	Last span
CAUSE	71	0.66	0.83
CONCESSION group	70	0.30	0.73
CONDITIONAL	18	0.28	0.89
ENABLEMENT group	6	0.67	0.50
EVALUATION group	140	0.68	0.83
EVIDENCE group	53	0.79	0.68
CONTRAST	26	0.35	0.69
Total	384	0.58	0.78

Table 5
Agreement of the semantic orientation of discourse relations and position (with the first or last span).

Whereas our first comparison relates to the hierarchical structure of discourse, in terms of nuclei and satellites, we then consider linear structure. We are interested in whether the semantic orientation of the discourse relation as a whole tends to match that of the EDU to its most immediate right or the EDU to the immediate left.

When we compare the semantic orientation of the discourse relation and position (first or last span), the results are clearer. Table 5 shows that the last span fits more (0.78) with the discourse relation in comparison with the first span (0.58). However, not all the discourse relations show the same tendency. ENABLEMENT and EVIDENCE have the same semantic orientation as that of the first span more often. In the case of ENABLEMENT, the score with the first and last spans is 0.67 and 0.50, respectively, while the scores are 0.79 and 0.68 for the EVIDENCE relation. The highest difference in score appears in the CONDITIONAL relation, because the first span has a low agreement in semantic orientation (0.28) and a high one (0.89) with the last span. Finally, the last spans of EVALUATION and CAUSE matches more often (0.83) the semantic orientation of the discourse relation as a whole.

In summary, regarding the semantic orientation of discourse relations, the results show that the semantic orientation of the entire relation is more often that of its nucleus. This is to be expected because, according to the RST framework, the nucleus is the most important EDU of the relation, so it makes sense that the semantic orientation of the nucleus contributes the most to the semantic orientation of the relation as a whole. However, this is not the case for all types of relations. For example, in the EVALUATION group, the semantic orientation of the satellite fits in more occasions. This is because of the characteristics of

this type of relation. In the EVALUATION group, the nucleus presents a situation and the satellite makes an evaluative comment about the situation.

Comparison of the semantic orientation of the text and discourse relation's distance from the central unit

Another aspect we wanted to investigate is the match of semantic orientation of the text and the semantic orientation of a discourse relation, according to the distance of the discourse relation from the central unit in the RST tree. The question is, given the semantic orientation of the entire text, does the semantic orientation of relations in the text change as those relations are farther away from the central unit? Figure 2 shows that the match of semantic orientation between opinion texts and their discourse relations is higher when the discourse relation is closer to the central unit. Distances −2, −1, 0, 1 and 2 present the highest match between all the distances between from −6 to +6[5].

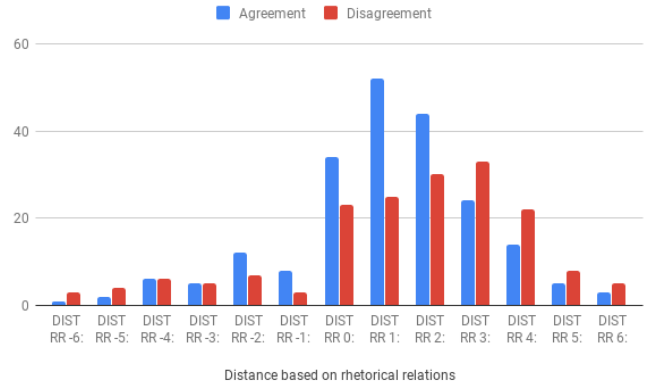


Figure 2. Semantic orientation of the text and discourse relations according to distance from the central unit.

If the results are analyzed for each discourse relation, there are some differences. In the CAUSE discourse relation, the instances that have the same semantic orientation prevail only in distances 0 and 1. In contrast with the overall tendency, the CONCESSION/ANTITHESIS group shows a discontinuous pattern regarding the same semantic orientation of the text. Relations with the same semantic orientation of the text are more frequent in two different non contiguous distance portions. Instances with the same semantic orientation happen more in distances −2, +1, +2, while a different semantic orientation is more frequent in distances 0 and +3.

Our interpretation about the semantic orientation between discourse relations and texts based on position is that when a discourse relation is closer to the central unit, there is typically a match between the semantic orientation of that discourse relation and the semantic orientation of the text overall. In our view, the reason for this is that the topic and the effect of discourse relations that are closer to the central unit are more closely aligned with the global topic and intended effect of the text. As a consequence, the semantic orientations will also be closely aligned.

[5] Distances with the signal – are situated to the left of the central unit. Distances with a positive sign '+' are situated to the right of the central unit.

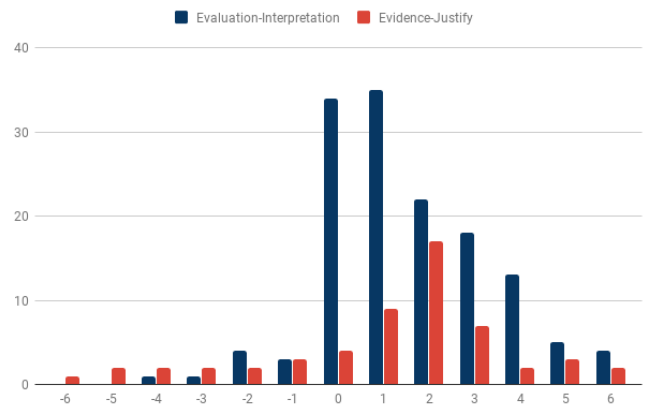


Figure 3. The most usual position of EVALUATION-INTERPRETATION and EVIDENCE-JUSTIFY discourse relations in opinion texts.

Discourse relations in opinion texts

An interesting aspect that deserves attention is the most usual position of different discourse relations in opinion texts. We have analyzed the distribution of instances of discourse relations with respect to the central unit. Figure 3 presents the INTERPRETATION-EVALUATION and EVIDENCE-JUSTIFY discourse relations, showing that for INTERPRETATION-EVALUATION, there are few instances before the central unit and, after it there is a high number of instances in distances 0 and 1. Finally, the number of instances decreases when moving away from the central unit. The CONTRAST multinuclear relation also shows similar characteristics, in fact, there are many instances of this discourse relation in distances -1 and $+3$ but between those distances, there is almost no instance. The situation is different in the case of EVIDENCE-JUSTIFY where, in contrast to the previous relations, it shows a greater degree between distances -1 to $+3$ but a regular distribution in the other distances because its instances are more uniform before and after the central unit.

Additionally, Figure 4 offers the results of the ANTITHESIS-CONCESSION and CAUSE-RESULT-PURPOSE discourse relations. These discourse relations share some similarities regarding their distance from the central unit. As it happened with the EVIDENCE-JUSTIFY group, these groups are distributed in all the distances in greater or lesser degree. But in contrast to other discourse groups, they present two peaks. In both relations, one peak is situated before the central unit and another one after it. In the case of ANTITHESIS-CONCESSION, the peaks are situated in distances -2 and $+3$. In a similar way, the peaks appear in distances -4 and $+1$ when it comes to the CAUSE-RESULTS-PURPOSE group.

Finally, Figure 5 shows the position of types of discourse relations based on our corpus. The results show big differences between instances of types of discourse relations. From distance -1, EVALUATION-INTERPRETATION is the most usual relation. It presents more than 40 instances in distances $+1$ and $+2$.

Although EVALUATION-INTERPRETATION is by far the most usual relation from distance -1, the

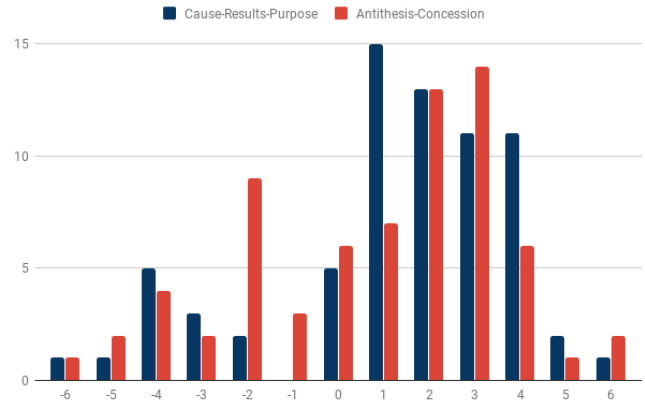


Figure 4. The most usual position of ANTITHESIS-CONCESSION and CAUSE-RESULT-PURPOSE discourse relations in opinion texts.

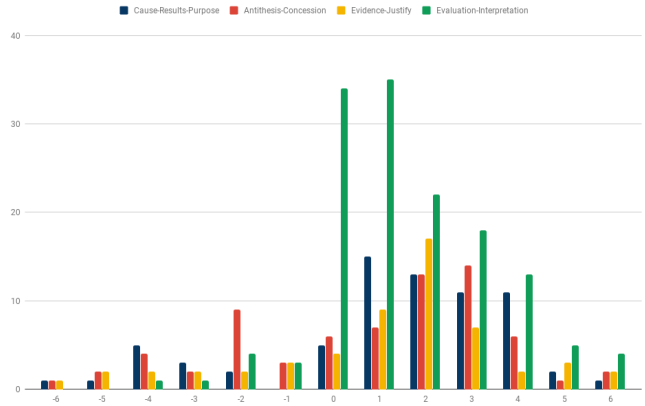


Figure 5. Instances of discourse relations based on distance from central unit.

CONCESSION relation in distance 0 and the CAUSE relation in distance +1 appear with high frequency. In distance +2, CAUSE and JUSTIFY show the same quantity of instances and in distance +3, the CONCESSION relation prevails over other discourse relations. Finally, in distance +4, the RESULT and PURPOSE relations are the most usual.

As far as the position of all the discourse relations of the corpus in texts is concerned, our interpretation is that there is a pattern even when the number of instances in modest to give a clear evidence. Taking into account the results of Figure 5, it seems that the order of discourse relations in texts is the following:

PURPOSE (distance −4), CONCESSION (−3 and −2), and EVALUATION (other distances). Without the EVALUATION relation, the order would be the following in other distances: CONCESSION (0), CAUSE (+1), CAUSE/JUSTIFY (+2), CONCESSION (+3), and RESULTS/PURPOSE (+4).

Conclusion and Future Work

This work presents an empirical study on semantic orientation and discourse relations. Our goal was to explore the interaction of semantic orientation (in the form of positive/negative or neutral sentiment) and discourse relations, in particular certain aspects of discourse relations such as nuclearity, position in the text or type of relation.

We started out by creating an annotated corpus of Basque opinion texts. We relied on a corpus that had already been annotated with discourse relations and assigned semantic orientation to all the discourse relations in the texts and their spans. In order to ensure the quality of the annotations, we performed an inter-annotator agreement study, which showed moderate agreement.

Using this corpus, we calculated different measures of the interplay between sentiment and discourse structure. We examined the semantic orientation of nuclei and satellites, and its match (or not) with the semantic orientation of the entire relation. We also considered any differences across relations.

Secondly, we calculated the distance of each discourse relation from the central unit, and whether the distance entails any difference in semantic orientation matching between the semantic orientation of the entire text and that of the relation under consideration.

The results of all those measures lead to the following conclusions:

1- Discourse relations coincide more in semantic orientation with their nuclei than with their satellites.

2- Discourse relations coincide more in semantic orientation with their last spans (right) than with their first spans (left).

3- The semantic orientation of the relation coincides much more when the satellite appears in the last span (left).

4- When it comes to distance from the central unit, the semantic orientation of relations which are at distances situated between −2 and +2 coincides in more occasions.

This work has had some limitations. Although the number of instances is considerable (in total, 384 instances), their distribution is not uniform across different types of relations. For example, CONDITIONAL (18 instances) and ENABLEMENT/MOTIVATION (6 instances) relations have few instances, making broad generalizations problematic. On the other hand, the inter-annotator agreement as for the semantic orientation of discourse relations and its spans is κ 0.58 which is considered 'moderate agreement' according to Landis and Koch (1977).

In the near future, we would like to pursue further study. First of all, we want to increase the number of discourse relations in the corpus, in order to achieve more precise results. We also plan to rewrite the guidelines to annotate the semantic orientation of discourse relations to achieve a higher inter-annotator agreement. Extending the analysis to others discourse relations is the other objective. Finally, after meeting the previous objectives, we would like to refine the discourse structure of opinion texts proposed in this study. For the future work we plan to automate these findings and add different pipelines to analyze the semantic orientation taking into account these discourse phenomena, following the subsequent tasks:

i) determine the EDUs and the central unit of the text with an existing central unit detector (Atutxa, Bengoetxea, Diaz de Ilarraza, & Iruskieta, 2019), *ii*) after that, add discourse relations with EusDisParser, the Basque RST parser developed by Iruskieta and Braud (2019) and *iii*) finally add the semantic orientation with Sentitegi (Alkorta et al., 2018).

Acknowledgements

This research is funded by the Basque Government PRE_2018_2_0033 grant. We would like to add a special thank also to the Simon Fraser University for given infrastructure to develop this study.

References

Alkorta, J., Gojenola, K., & Iruskieta, M. (2016). Creating and evaluating a polarity-balanced corpus for basque sentiment analysis. In *Iwoda16 fourth international workshop on discourse analysis. santiago de compostela, september* (Vol. 29).

Alkorta, J., Gojenola, K., & Iruskieta, M. (2018). Sentitegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, *22*(4).

Alkorta, J., Gojenola, K., Iruskieta, M., & Pérez, A. (2015). Using relational discourse structure information in basque sentiment analysis. In *Sepln 5th workshop rst and discourse studies.*

Alkorta, J., Gojenola, K., Iruskieta, M., & Taboada, M. (2017). Using lexical level information in discourse structures for basque sentiment analysis. In *Proceedings of the 6th workshop on recent advances in rst and related formalisms* (pp. 39–47).

Asher, N., Benamara, F., & Mathieu, Y. Y. (2008, August). Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters* (pp. 7–10). Manchester, UK: Coling 2008 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C08-2002

Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., & Iruskieta, M. (2019, 09). Towards a top-down approach for an automatic discourse analysis for basque: Segmentation and central unit detection tool. *PLOS ONE*, *14*(9), 1-25. Retrieved from https://doi.org/10.1371/journal.pone.0221639 doi: 10.1371/journal.pone.0221639

Benamara, F., Taboada, M., & Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, *43*(1), 201-264.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015, September). Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2212–2218). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D15-1263 doi: 10.18653/v1/D15-1263

Chardon, B., Benamara, F., Mathieu, Y., Popescu, V., & Asher, N. (2013). Measuring the effect of discourse structure on sentiment analysis. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 25–37). Berlin, Heidelberg: Springer Berlin Heidelberg.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). Polarity analysis of texts using discourse structure. In *Proceedings of the 20th acm international conference on information and knowledge management* (pp. 1061–1070). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2063576.2063730 doi: 10.1145/2063576.2063730

Iruskieta, M. (2014). Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea*.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., & de Lacalle, O. L. (2013). The rst basque treebank: an online search interface to check rhetorical relations. In *4th workshop rst and discourse studies* (pp. 40–49).

Iruskieta, M., & Braud, C. (2019). Eusdisparser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the workshop on discourse relation parsing and treebanking 2019* (pp. 62–71).

Iruskieta, M., da Cunha, I., & Taboada, M. (2015, June). A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, *49*(2), 263–309. Retrieved from http://dx.doi.org/10.1007/s10579-014-9271-6 doi: 10.1007/s10579-014-9271-6

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Mann, W. C., & 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.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: MIT Press.

Mukherjee, S., & Bhattacharyya, P. (2012, December). Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of COLING 2012* (pp. 1847–1864). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C12-1113

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications* (pp. 1–10). Springer.

San Vicente, I., Saralegi, X., & Agerri, R. (2015, June). EliXa: A modular and flexible ABSA platform. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 748–752). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S15-2127 doi: 10.18653/v1/S15-2127

Somasundaran, S., Namata, G., Wiebe, J., & Getoor, L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1* (pp. 170–179). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1699510.1699533

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, *2*, 325-347.

Taboada, M., Voll, K., & Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Trnavac, R., Das, D., & Taboada, M. (2016). Discourse relations and evaluation. *Corpora*, *11*(2), 169–190.

Trnavac, R., & Taboada, M. (2012). The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, *34*(3), 301-318.

Zhou, L., Li, B., Gao, W., Wei, Z., & Wong, K.-F. (2011, July). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 162–171). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D11-1015

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011, November). Fine-grained sentiment analysis with structural features. In *Proceedings of 5th international joint conference on natural language processing* (pp. 336–344). Chiang Mai, Thailand: Asian Federation of Natural Language Processing. Retrieved from https://www.aclweb.org/anthology/I11-1038

CONCLUSION

5

Contributions, conclusions, and future work

As we mentioned at the beginning of this thesis work, nowadays there are numerous opinions on the Internet and extracting subjective information from these texts and, specifically, knowing if these opinion texts have a positive or negative opinion is a challenge. In this direction, we have developed tools and resources for extracting subjective information from Basque opinion texts. Moreover, we have studied different linguistic phenomena that change the sentiment valence of words, known as contextual valence shifters.

In this section, we will present the contributions and conclusions drawn from the sentiment analysis regarding Basque, as well as the challenges facing the future.

5.1 Contributions

This thesis is one of the first works in sentiment analysis of written texts in Basque from an applied linguistics perspective. After building a corpus of opinion texts and developing a sentiment lexicon, the phenomena affecting sentiment valence and semantic orientation of the words have been studied. In the future, these aspects will need to be considered when developing a document-level sentiment classification based on sentiment lexicon. These resources and tools we have developed as well as the study of linguistic phenomena are the result of the methodology we have followed.

5.1.1 Tools and resources related to sentiment analysis

Concerning sentiment analysis, we have developed an annotated corpus of opinion texts in Basque, a sentiment lexicon in Basque called *Sentitegi* and a document-level sentiment classifier in Basque.

- Basque Opinion Corpus (Alkorta et al., 2016). This corpus includes 240 opinion texts from six domains. It has annotated from discourse structure and subjectivity. The *RST* approach (Mann and Thompson, 1988) has been used for annotating discourse structure and, in total, 70 opinion texts have been annotated. 192 central units of opinion texts have also been identified. From sentiment analysis, we have determined whether the text is positive or negative. Furthermore, in 28 opinion texts, for certain types of discourse relations and their constituents (nucleus and satellite), we have assigned the semantic orientation (positive or negative) as well as sentiment value (numerical sentiment value from -5 to $+5$).

- *Sentitegi*, the sentiment lexicon in Basque (Alkorta et al., 2018). This sentiment lexicon includes 1,237 words of 4 grammar categories. It is based on the Spanish lexicon of the SO-CAL tool (Brooke et al., 2009) and it is also enriched with the English lexicon (Taboada et al., 2011). This lexicon has been adapted to six domains: weather, sport, politics, music, film and literature and the Basque Opinion Corpus has been used for that purpose.

- Document-level sentiment classification tool in Basque. Another contribution of this thesis has been the sentiment classifier based on sentiment lexicon. We have based our work on the SO-CAL tool (Taboada et al., 2011) to develop the Basque version. However, because of the typological differences between English and Basque, there are some differences between the tools of both languages.

  In the English sentiment classifier, there are sentiment lexicons, morphology-related rules, and general rules. In contrast, in the Basque version, we have removed the rules related to morphology and we have integrated the *Eustagger* lemmatizer (Alegria et al., 2002). The Basque classifier, firstly, lemmatizes the text and, then, looks up whether the lemmatized word is contained in the lexicon and in this case the classifier assigns a sentiment valence to the word.

5.1.2 Analysis of valence shifters in Basque

To work on sentiment analysis, we have analyzed contextual valence shifters of different grammatical levels that affect to words and phrases. We have focused on (Polanyi and Zaenen, 2006):

- We have studied morphological valence shifters. As we have seen, the phonological phenomenon of expressive palatalization, based on the instances of the corpus, reinforces the sentiment valence of the word. In morphology, we have found prefixes and suffixes that can strengthen or weaken the valence of the word. Besides, we have studied several affixes related to negation.

- We have also worked on the syntactic level and specifically on the negation markers and their scope. Our study shows that negation markers can strengthen or weaken the sentiment valence of words or phrases but also that in some cases they do not change the valence or semantic orientation.

  It is also worth mentioning that the negation marker ("ezin") has a double behavior depending on the grammatical category of the word around it. The results have also shown that some negation markers appear more frequently with other words. We call these lexicalized structures. In terms of CG rules to identify elements related to negation, we have found that it is more difficult to identify scope than negation markers and lexicalized structures because of their irregular structure and variable length.

- Finally, we have also studied the discourse level and, for that purpose, we followed the *RST* approach (Mann and Thompson, 1988). In this work, we have focused on discourse relations and central unit. In discourse relations, we have seen that the nuclei or the final text span of the discourse relations strengthen the agreement of the semantic orientation between these constituents and the discourse relations. Besides, when a discourse relation is closer to the central unit, the semantic agreement of the relation and the whole text is strengthened. Consequently, according to the results, the nucleus and the last text span in discourse relations and central unit in the texts are strengthening valence shifters.

Regarding the central unit, we have also studied the words with sentiment valence. According to our results, in the central unit, 12.40% of the words have sentiment valence and their most usual grammatical categories are adjectives, nouns, and verbs. In the central unit, we have also studied words with sentiment valence from the domain and the results show differences between the domains. For example, in sports, the most common grammatical category with sentiment valence is the verb. In contrast, in weather, the most common grammatical category with sentiment valence is the adjective.

5.2 Conclusions

In this work, we established some goals before starting the thesis and have fulfilled those goals. After completing the thesis, we come to the following conclusions.

- Translation to create a sentiment lexicon in Basque can be a good option for sentiment analysis. In fact, according to Pearson correlation results, the quality of the vocabulary of the Basque language created following our methodology is good. The correlation coefficient is 0.76 for the words annotated by the lexicon and the gold standard. But in some cases, the gold standard, unlike the lexicon we have developed, assigns a certain value of sentiment to the words, and as a result, the coefficient drops to 0.54. That means that the quality of the lexicon is high but there is a possibility to expand the lexicon incorporating more words with sentiment valence.

- To develop a document-level sentiment classifier based on lexicon, it is necessary to take into account contextual valence shifters. As the work has shown, at different levels of grammar, from phonology to discourse, there are valence shifters and they can have three effects on the sentiment valence or semantic orientation of the words and phrases: strengthening, weakening, or no change.

- The rich morphology of Basque also influences sentiment analysis. The results of the thesis work show that Basque prefixes and suffixes can affect the sentiment valence of words, strengthening or weakening its

value and, in some cases, changing the sign of the valence. From semantics, the morphemes in Basque are diverse and they also influence valences. The thesis analysis also shows that expressive palatalization strengths the sentiment valence of words.

- In syntax, we have found that negation markers strengthen, weaken, or do not change the sentiment valence of words or phrases. Besides, we have identified lexicalized structures. They are negation markers that appear with specific words with great frequency and they also affect the sentiment valence of phrases.

- Discourse structure also plays an important role in sentiment classification based on the lexicon. Nucleus and the last text span in discourse relations and central units in texts affect to agreement of semantic orientation between the mentioned elements, increasing the agreement.

Taking into account the above points, from the linguistic point of view, it can be concluded that we have taken the first steps in the sentiment analysis in Basque. However, it is still necessary to begin to work on other aspects or to deepen them, in particular, the contextual valence shifters that are unexplored.

5.3 Future work

In the study of sentiment analysis and its computational processing, we predict the following lines of work for the future:

- To have an even more diverse Basque Opinion Corpus. Nowadays, the corpus includes opinion texts from specialized newspapers and websites and it is composed of 6 domains. The aim for the future is to compile opinion texts. As mentioned in the development of the methodology,we have found few opinion texts with good quality to study discourse. The aim is to compile texts of this type to study other phenomena that have not yet been explored. For example, when ordinary people write texts it is possible to stretch words by repeating a letter as well as use emoticons, which can also affect semantic orientation.

- To improve and expand the annotation of the discourse structure of the corpus. Currently, 70 of the 240 texts in the corpus are annotated and

we want to increase that number. However, before further annotation, it is necessary to reach a better inter-annotator agreement.

- To extend the sentiment lexicon and to deal with different events of the methodology. As we have seen, the quality of the vocabulary we have created is good but it is not yet possible to identify all the words and/or expressions with subjectivity. As a result, we want to increase its size and make it useful for more domains. Finally, we also want to solve the situation of polysemous words, that is to say, decide what to do with words that have two opposing semantic orientations in different contexts. Words like "ikaragarri" (frightening/enormous) have the opposite semantic orientation when it comes to a movie or an accident.

- To continue identifying and studying further contextual valence shifters in the Basque language. In this thesis, we have studied one type of valence shifter in each grammatical category, but there are still some valence shifters that have not yet been studied. We can mention, for example, conditionals, interrogative sentences or, even more complex, the irony. Some of the valence shifters may have specificities in Basque and others may be similar in different languages.

- To deeper study the discourse level. In this work, we focus on discourse relations and central unity. In the future, however, we want to go further and explore other components of discourse structure. In other words, we want to work on other elements of discourse trees, such as the central subconstituent. We also want to make a deeper study on types of discourse relations and find differences that may exist between them.

- To further develop the document level sentiment classifier. The tool is currently made up of Basque lexicon, lemmatizer, and general rules, but we want to integrate more modules into it. For example, we would like to set up a module related to the contextual valence shifters used in this thesis and check if the quality of the tool improves.

Bibliography

Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., and Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6.

Alkorta, J., Gojenola, K., and Iruskieta, M. (2016). Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela, September*, volume 29.

Alkorta, J., Gojenola, K., and Iruskieta, M. (2018). SentiTegi: Semimanually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas*, 22(4).

Altuna, B., Aranzabe, M. J., and de Ilarraza, A. D. (2017). Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria eta Patxi Salaberri (ed.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea, 127-134, Udako Euskal Unibertsitatea (UEU), Bilbo.*

Altuna, P., Salaburu, P., Goenaga, P., Lasarte, M. P., Akesolo, L., Azkarate, M., Charriton, P., Eguskitza, A., Haritschelhar, J., King, A., Larrarte, J. M., Mujika, J. A., Oyharçabal, B. n., and Rotaetxe, K. (1985). Eu-

skal Gramatika Lehen urratsak (EGLU) II. *Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo*.

Barnes, J., Lambert, P., and Badia, T. (2018). MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. *arXiv preprint arXiv:1803.08614*.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From English to Spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Das, D. and Taboada, M. (2018). RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184.

Euskara Institutua, EHU (2011). Sareko euskal gramatika, www.ehu.eus/seg.

Iruskieta, M., Da Cunha, I., and Taboada, M. (2015). A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Karlsson, F., Voutilainen, A., Heikkilae, J., and Anttila, A. (2011). *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Mann, W. C. and 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.

Oñederra, L. (1990). *Euskal fonologia: palatalizazioa: asimilazioa eta hots sinbolismoa*. Servicio Editorial de la Universidad del País Vasco = Euskal Herriko Unibertsitatea.

O'Donnell, M. (2000). RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 253–256. Association for Computational Linguistics.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., and Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

San Vicente, I. and Saralegi, X. (2016). Polarity Lexicon Building: to what Extent Is the Manual Effort Worth? In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

San Vicente, I., Saralegi, X., and Agerri, R. (2017). Elixa: A modular and flexible absa platform. *arXiv preprint arXiv:1702.01944*.

Saralegi, X., San Vicente, I., and Ugarteburu, I. (2013). Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-resourced Languages. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 96–108, Berlin, Heidelberg. Springer-Verlag.

Sarasola, I. (2005). *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

Stone, P. J. and Hunt, E. B. (1963). A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Urdangarin, L. M. M. (1982). Morfología de la composición lexical euskérica. *Fontes linguae vasconum: Studia et documenta*, 14(39):233–272.

Zerbitzuak, E. H. (2013). Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. Usurbil: Elhuyar.

Terminology and abbreviations

Sentimenduen analisi (*Sentiment analysis*)
Orientazio semantiko (*Semantic orientation*)
Sentimendu-balentzi (*Sentiment valence*)
(Testuinguruko) balentzia-aldatzaile (*(Contextual) valence shifter*)
Sentimenduen sailkatzaile (*Sentiment classifier*)
Sentimenduen lexikoi (*Sentiment lexicon*)
Murriztapen Gramatika, MG (*Constraint Grammar, CG*)
Ezeztapen-marka (*Negation mark*)
(Ezeztapenaren) irismena (*Scope*)
Egitura lexikalizatu (*Lexicalized structure*)
Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory, RST*)
Unitate zentrala, UZ (*Central Unit, CU*)
Erlaziozko diskurtso-egitura (*Relational discourse structure*)

The writing of this thesis
was ended on 15th October, 2019.

EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)

Hizkuntzaren eta Literaturaren Didaktika Saila

Doktorego-tesia

Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa

Jon Alkorta Agirrezabala

Donostia, 2019

EUSKAL HERRIKO UNIBERTSITATEA (UPV/EHU)

Hizkuntzaren eta Literaturaren Didaktika

Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa

Jon Alkorta Agirrezabalak Koldo Gojenola Galletebeitia eta Mikel Iruskieta Quintianen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Aitari, Amari eta Josuri

Eskerrak

Lehenik eta behin, nire zuzendariak eskertu nahiko nituzke: Koldo eta Mikel, eman didaten laguntzagatik. Eskerrik asko beti nire atzean egoteagatik eta baita tesian zehar sortu diren zalantzetan laguntza eskaintzeagatik ere. Azkenik, eskerrik asko egin dizkidazuen iruzkinengatik (asko lagundu baitidate) eta ikerkuntzaren mundua nolakoa den erakusteagatik ere. Halaber, eskerrak Maxuxi eta Rodriri tesi-txostena hobetzen laguntzeagatik.

Eskerrik asko tesian lagundu didaten beste pertsonei ere. Eskerrik asko Arantxari *Eustagger*rekin izan ditudan une zailetan laguntzeagatik eta Intxari *include*k Latexen duen garrantzia erakusteagatik. Eskerrak Estherri Kanadatik Ixarako konexioa ahalbidetzeagatik, Kikeri emandako laguntzagatik eta Amaiari administrazioaren munduan laguntzeagatik. Eskerrik asko, halaber, Itziar Gonzalez-Diosi *Murriztapen Gramatika*rekin eta tesi-txostena vs. Latex borrokan laguntzeagatik. Eskerrik asko zerrenda honetan aipatu gabe gelditu diren baina lagundu didaten beste guztiei ere.

Eskertzekoa da tesia hasi eta amaitu arte inguruan mugitu denari ere: 318 bulegoari giro ona sortzeagatik, tupper txokoari askotariko gaiak (batzuetan, frikiak) aipatzeagatik, pintxo-poteetan eta egindako beste planetan atera zareten guztioi eta, azkenik, estatistika klaseetan, Oierri izandako pazientziagatik eta *nukleo duro*ri estatistika ikasteko laguntza morala emateagatik.

También tengo que agradecer a Maite Taboada por su ayuda en la estancia de Vancouver y en los trabajos relacionados con el discurso y análisis de sentimientos.

Familia ere ezin da aipatu gabe utzi. Eskerrik asko aitari, amari eta Josuri

beti hor egoteagatik, eta laguntza behar izan dudanean laguntzeagatik.

Eskerrik asko kuadrillari ere, asteburuetan, musa eta *Comunio*rekin une biziki ziragarriak emateagatik. Eskerrak baita Urkizu eta Gazteluko bazkariengatik eta Kanpezura eta Jakatik mendira egindako planengatik ere.

Bukatzeko eskerrik asko Amets eta Lierni pisukideei eta aipatzea ahaztu zaizkidan beste guztiei.

Mila-mila esker denoi!

Laburpena

Tesi-lan honetan, hizkuntzalaritza aplikatuaren ikuspegitik, euskarazko sentimenduen analisian lehen urratsak egin dira. Bi helburu nagusi egon dira tesi-lanean. Alde batetik, sentimenduen analisia egiteko oinarrizko baliabideak sortu ditugu euskararentzat. Zehatz esanda, Euskarazko Iritzi Corpusa, *Sentitegi* izeneko euskarazko sentimenduen lexikoia eta dokumentu mailako sentimenduen sailkatzailea garatu ditugu. Corpusak sei domeinutako 240 iritzi-testu biltzen ditu. Egitura Erretorikoaren Teoriaz (RST) baliatuta, corpuseko diskurtso-informazioa etiketatuta dago. Gainera, iritzi-testuen orientazio semantikoa ere etiketatuta dago. Sentimenduen lexikoiari dagokionez, 1.237 hitzez osatuta dago eta bertako sarrerek -5 eta +5 arteko sentimendu-balentzia dute. Sentimenduen lexikoia sortzeko itzulpen-metodologia zehatz bat jarraitu dugu. Azkenik, dokumentu mailako sentimenduen sailkatzailea ere garatu dugu. Tresnaren oinarrian aurretik aipatu dugun sentimenduen lexikoia dago.

Beste aldetik, sentimenduen analisiaren lanketa teoriko bat ere egin dugu. Sentimenduen sailkapena lexikoian oinarrituz egin nahi bada, hitzen sentimenduen balentzia jakitearekin ez da nahikoa, izan ere, testuetan badaude zenbait fenomeno hitz horien sentimenduen balentzia eragiten dutenak. Horiei testuinguruko balentzia-aldatzaileak deritze eta horiek euskaran nola agertzen diren landu dugu. Hizkuntza maila bakoitzeko balentzia-aldatzaile mota bat landu dugu: fonologian, bustidura adierazkorra; morfologian, hizkiak; sintaxian, ezeztapen-markak eta, azkenik, diskurtsoan, erlaziozko diskurtso-egiturak eta unitate zentrala. Emaitzek erakusten dutenez, balentzia-aldatzaileek hitzen edo sintagmen sentimenduen balentzia indartu

edo ahuldu egiten dute. Ahultze horren intentsitatearen arabera, sentimendu balentziaren zeinuan aldaketa gerta liteke, positiboa dena negatibo bilakatuz edo alderantziz. Azkenik, kasu batzuetan, balentzia-aldatzaileak ez du eraginik sortzen.

Aurkibidea

METODOLOGIA

ONDORIOAK

Eranskina

Irudien zerrenda

Taulen zerrenda

SARRERA

1. KAPITULUA

Proiektuaren nondik norakoak

1.1 Motibazioa

Iritziak giza jardueraren oinarrian daude eta gure portaeran eragin handia dute (Liu, 2012). Errealitatearen gure uste eta pertzepzioak nahiz egiten ditugun aukeraketak hein handi batean munduak ikusi eta ebaluatzen duenaz baldintzatuta daude. Egoera horren ondorioz, erabaki bat hartu behar dugunean, besteen iritziak bilatu ohi ditugu. Norbanakoetan eta erakundeetan gertatzen den egoera da.

Iritziaren inguruan ere badira zenbait kontzeptu subjektibitatearekin lotura dutenak. Besteen artean, sentimendua, ebaluazioa, jarrerak eta emozioak aipa daitezke eta horiek guztiak, iritziaz gain, sentimenduen analisia edo iritzi-meatzaritza deritzon arloaren ikergaiak dira.

Liuk (2012) gogorarazten duenez, arlo horren hasiera eta hazkundea erabat loturik daude webetako sare sozialen jaiotzarekin. Webetako sare sozialak era askotakoak dira: iruzkinak, foroetako eztabaidak, blogak, mikroblogak eta Twitter. Aplikazio horiek guztiek gizakiaren historian lehen aldiz digitalki jasotako iritzi bolumen handi bat edukitzea ahalbidetu dute. Hori horrela izanik, 2000tik aurrera hizkuntza naturalaren prozesamenduan hazkunderik handiena eta aktiboena bilakatu den arloa da. Gainera, datu-meatzaritzan, web-meatzaritzan eta testu-meatzaritzan ere erabiltzen da.

Oro har, enpresan eta gizartean garrantzizkoa bilakatu da sentimendu-analisia, eta horrek arloa konputazio-zientzietatik kudeaketa-zientzietara eta gizarte-

zientzietara zabaldu du. Bere inguruan, industria garatu da eta enpresek zerbitzuak ere sortu dituzte arloa lantzeko. Beraz, sentimendu-analisiak enpresa edo gizarte eremu askotan aplikazioa aurkitu duela esan daiteke.

Ikusi dugun moduan, nazioartean sentimenduen analisiak garapen handia izan du. Egun, munduko *lingua franca* ingelesa da eta sentimendu-analisiko lan gehienak hizkuntza horretan egin dira. Gurera etorrita, ordea, gutxi dira euskara hartuta egin diren lanak.

Sentimendu-analisia oso arlo zabala da. Ataza batzuek lotura handiagoa dute informatikarekin eta beste batzuek, berriz, hizkuntzalaritzarekin. Gainera, ataza batzuetan bi hurbilpen egoten dira: hizkuntzaren ezagutzan oinarritutakoa eta metodo estatistikoetan oinarritutakoa.

Hizkuntzalaritzaren ikuspegitik, lehen urratsa da zein hitzek duen iritzia edo, modu zabalago batean, informazio subjektiboa den jakitea. Hurrengo urratsean, ordea, hitz horien informazio subjektiboan eragiten duten hizkuntzari lotutako fenomenoak bilatu nahi dira. Azter ditzagun hurrengo adibideak.

(1) Pelikula hori [gustatu$_+$]$_+$ zait.

(2) Pelikula hori [asko gustatu$_+$]$_{++}$ zait.

(3) Pelikula hori ez zait [gustatu$_+$]$_-$.

(4) Pelikula hori [gustatuko$_+$ litzaidake]$_0$ kontakizuna biziagoa izango balitz.

Goian, pelikula bati buruzko zenbait iritzi agertzen dira. Denak antzekoak dira, baina ez berdinak. (1) adibidean, esaterako, pelikula gustukoa izan duela esaten da; beraz, esaldiaren balorazioa positiboa da. (2) adibidean ere esaldiaren balorazioa positiboa da, baina *asko* intentsifikatzailea[1] agertzen da eta, horregatik, esaldi honek aurreko esaldiak baino iritzi positiboagoa du. (3) adibidean, ordea, esaldian iritzia adierazten duen hitz bat, positiboa gainera, egon arren, esaldiaren balorazioa negatiboa da. Azkenik, (4) adibidean ere, iritzia duen hitz bakarrak balorazio positiboa egiten du, baina esaldiak

[1]Intentsifikatzaileak adberbioak edo adjektiboak izan ohi dira. Berezko eduki semantiko gutxi dute, baina aldatzen duen hitzaren edo esaldiaren esanahia intentsifikatzen du.

ez du iritzirik adierazten, erreala ez den egoera batean dagoelako iritzia[2].

Adibide horietan ikusten denez, esaldi guztietan iritzia adierazten duen hitz bat (*gustatu*) egon arren, esaldi osoek adierazten duten iritzia ezberdina da. Tartean, hitzen arteko interakzioak egon dira eta fenomeno linguistiko batzuek iritzia adierazten duen hitzaren balorazioa aldatu egin dute.

Hain zuzen ere, aipatu ditugun bi alderdi horiek dira gure lanaren motibazioak. Alde batetik, euskararentzat sentimendu-analisiaren tresnak eta baliabideak sortu nahi ditugu, gaur egun, euskarazko testuen eduki subjektiboa erauzi ahal izateko. Beste aldetik, berriz, iritzia edo informazio subjektiboa duten hitzetan eragiten duten fenomeno linguistikoak bilatu nahi ditugu, fonologian, sintaxian eta diskurtsoan.

Gure lehen helburua sentimendu-analisiaren arloan ikerketa egiteko oinarrizko tresna eta baliabideak sortzea da. Zehazki, euskarazko iritzi-testuen corpusa eta euskarazko sentimendu lexikoia sortu nahi ditugu.

Izan ere, gutxi dira euskaraz dauden iritzi-testuen corpusak eta sentimendu lexikoiak, eta daudenak, berriz, ez dira baliagarriak gure helburuak betetzeko; izan ere, ez dituzte hizkuntzaren ezaugarriak aintzat hartzen eta hori ez dator bat gure ikusmoldearekin. Iritzi-testuen kasuan, txioak eta zerbitzuak (produktuen salerosketa webguneek eta hotelen web konparatzailea, adibidez) eskaintzen dituzten webguneetako iritziak biltzen dituzten corpusak edota datu-baseak badaude baina guri ez zaizkigu baliagarri testuak motzak direlako eta ondorioz, diskurtsorako zenbait elementu aztertzeko aukera zailduko ligukeelako. Lexikoiei dagokienez, berriz, hitzen orientazio semantikoa (positiboa edo negatiboa) adierazi arren, ez dute intentsitatea adierazten (hau da, *ondo* eta *bikain* positiboak dira, baina intentsitatean duten ezberdintasuna ez da adierazten) eta horregatik, ez zaizkigu baliagarriak. (Stone eta Hunt, 1963) sentimendu lexikoia, esaterako, tankera horretakoa da.

Gure bigarren helburua aurreko erronkan sortutako sentimendu lexikoiko hitzen balentzian eragiten duten hizkuntza-fenomenoak bilatzea eta beren eragina neurtzea da.

[2]Adibidean ematen den iritzia egoera errealean ez dagoela diogu, iritzia emateko erabiltzen den aditza ez dagoelako ez iraganean, ezta orainaldian ere.

(1), (2), (3) eta (4) adibideetan ikusi dugun moduan, lexikoian oinarritzen den sentimenduen sailkatzaile batean, lexikoia bere baitan bakarrik ez da baliagarria. Informazio linguistiko gehiago behar da sentimenduen sailkapen on bat egiteko. Horregatik, Polanyi eta Zaenenen (2006) lanean oinarrituz, euskaran dauden hizkuntza maila ezberdinetako testuinguruko balentzia-aldatzaileak[3] identifikatu nahi ditugu.

Gure hirugarren eta azken helburua dokumentu mailako euskarazko sentimenduen sailkatzaile bat sortzea da. Euskaran lan gutxi daude sentimenduen sailkapena egiten dutenak. Horietako bat EliXa (San Vicente *et al.*, 2015) da eta aspektu mailako sentimenduen sailkapena egiten du[4]. Gure erronka euskarari hizkuntza-prozesamendurako beste baliabide bat ematea da eta, horretarako, lexikoian oinarritutako dokumentu mailako sentimenduen sailkatzailea sortu nahi dugu. Esan behar da erronka hau betetzeko beharrezkoa dela aurretik aipatutako erronkak betetzea: sentimenduen lexikoi bat sortu behar da sentimenduen sailkatzailea inplementatzeko eta balentzia-aldatzaileak landu behar dira lexikoiko hitzetan eragiten duten aldaketa neurtzeko.

Laburbilduz, tesi honek euskara eta sentimenduen analisia uztartzen ditu. Euskara hizkuntzatzat hartzen duten lan gutxi egin dira sentimenduen analisian eta gehienak sentimendu lexikoia sortzera mugatzen dira. Guk euskarari baliabide eta tresnak eman nahi dizkiogu, baita euskarak arlo honetan dituen berezitasunak ikertu ere.

1.2 Hipotesi orokorrak

Aurreko atalean, sentimenduen analisia arloaren motibazioa aipatu dugu, baita berak gizartean duen erabilgarritasuna ere. Atal honetan, tesiaren irismena zehaztuko dugu. Lau hipotesi orokor ditu lan honek.

[3]Testuinguruko balentzia-aldatzaileak (*contextual valence shifters*, ingelesez) hitzen edota esaldien sentimendu balentzian aldaketak eragiten dituzten fenomenoak dira. Aldaketa hori balentziaren indartze bat edo ahultze bat izan daiteke.

[4]Aspektu mailako sentimenduen analisiak aztergai den entitateari loturiko aspektuak identifikatu (entitatea Londres bada, aspektuak ekonomia, turismoa etab. dira) eta horiei buruzko iritzia positiboa edo negatiboa den zehaztu nahi du.

1. ikerketa-hipotesia: *itzulpen bidez lortutako euskarazko sentimenduen lexikoiaren kalitatea corpusetik edo datu-base lexikaletatik sor daitezkeenak bezain baliagarria da.*
Sentimenduen lexikoiek testuetako hitzei beren orientazio semantikoa[5] eta sentimenduen balentzia esleitzen diete eta hori testu baten orientazio semantikoa kalkulatzeko lehen urratsa da. Lexikoiak sortzeko, ohiko bi hurbilpen daude: i) corpus bidez eta ii) datu-base lexikalen bidez.

Guk hautatutako bidea aipaturiko hurbilpenen nahasketa da. Oinarri bezala bi datu-base lexikal, hots, gaztelaniazko eta ingelesezko bi sentimenduen lexikoi erabiliko ditugu eta horiek itzuli egingo ditugu Euskarazko Iritzi Corpusak ematen duen informazioa aberastuz.

Gure ustez, jada sorturik dagoen sentimenduen lexikoia euskaratzea aurreko bi hurbilpenak (corpus bidezkoa eta datu-base lexikal bidezkoa) bezain baliagarria da eta kalitatean ez legoke ezberdintasun handirik. Izan ere, itzulpenaren metodologia aproposa bada, bidean ez da informaziorik galduko edo eraldatuko eta hasierako kalitatea mantenduko luke. Gainera, iritzi-testuen corpusa lagun izanda, sentimendua duten hitzen orientazio semantikoan nahiz sentimenduen balentzian okerreko esleipenak egitea eragotziko litzateke.

2. ikerketa-hipotesia: *hizkuntza maila guztietan (fonologia, sintaxia eta diskurtsoa) zenbait fenomeno linguistiko daude orientazio semantikoa duten hitzetan (eta esaldietan) eragiten dutenak.*
Eragin hori beren sentimenduen balentzia indartzea edo ahultzea izan daiteke. Batzuetan fenomeno linguistikoak ez du balentzian aldaketarik eragiten.

Gure ustez, fenomeno linguistiko horiek gramatika maila ezberdinetan ager daitezke: esateko moduan (hau da, fonologian), hizkien bidez, sintaxiko fenomeno ezberdinen bidez edota baita diskurtsoaren bidez ere. Lexikoian oinarritzen den dokumentu mailako sentimenduen sailkatzaile bati mota honetako informazioa gehitzea beharrezkoa dela uste dugu, bestela sailkatzailearen emaitzak ez baitira nahi bezain zehatzak izango.

[5]Orientazio semantikoa hitz batek duen informazio subjektiboa da eta positiboa edo negatiboa izan daiteke.

3. ikerketa-hipotesia: *diskurtso-egituran badaude osagaiak EDUen*[6]
Gure ustez, EDU edo erlaziozko diskurtso-egiturak orientazio semantiko positiboa edo negatiboa edukitzea ez da ausazko gertaera bat. Hau da, elementu horiek RST-zuhaitzean[7] dituzten guneak horretan eragina dutela uste dugu. Zehazki esanda, diskurtso-egiturako zenbait faktorek edo osagaik EDU eta erlaziozko diskurtso-egiturako orientazio semantikoa baldintzatzen dutela da gure hipotesia.

4. ikerketa-hipotesia: *testu baten domeinuak unitate zentralak ezaugarri jakin batzuk izatea eragiten du, bai bertan agertzen diren hitzen gramatika-kategoriari dagokionez, bai hitz horiek orientazio semantikoa edukitzea edo ez edukitzearekin ere.*
Nahiz eta unitate zentrala testu guztietan dagoen, bere ezaugarriak domeinuaren arabera ezberdinak direla uste dugu. Eta horrek unitate zentraletan orientazio semantikodun hitzak agertzeko moduan eragiten duela ere pentsatzen dugu. Testuen sentimenduen analisia egiterakoan, beharrezkoa domeinua zein den jakitea, horrek arrasto batzuk ematen baititu sentimendu balentziadun hitzak non agertzen diren jakiteko.

1.3 Helburuak

Aurreko ataletan ikerketa-lan honen testuingurua eta gure ikerketa gidatzeko hipotesi orokorrak eztabaidatu ditugu. Atal honetan, berriz, gure helburu zehatzak adieraziko ditugu gure hipotesiak sostengatzen dituzten egiaztatzeko.

- **1. helburua:** Sentimenduen analisia lantzeko oinarrizko tresnak eta baliabideak sortzea.

[6]Oinarrizko Diskurtso Unitatea (*Elementary Discourse Unit*, EDU) *diskurtso-egiturako unitaterik txikiena da.* eta erlaziozko diskurtso-egituren orientazio semantikoa duten bezalakoa izatea eragiten dutenak.

[7]Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory*, RST) testu-sorkuntza automatikorako sortutako teoria da. RSTn testuaren antolakuntza testu-zatien artean daude erlaziozko diskurtso-egitura ezberdinen bidez garatzen da eta koherentzia testuaren egiturak duen hierarkiaren bidez gauzatzen da.

  - ***1.1 helburua:*** euskarazko iritzi-testuak biltzen dituen corpusa sortzea, testuen orientazio semantikoari dagokionez orekatua eta sintaxiari dagokionez aberatsa.
  - ***1.2 helburua:*** euskarazko sentimenduen lexikoia sortzea, hitzen orientazio semantikoa adierazten duena eta horretarako hitzei zenbakizko balio bat esleitzen diena.
  - ***1.3 helburua:*** dokumentu mailako euskarazko sentimenduen sailkatzaile bat sortzea, oinarritzat sentimenduen lexikoia izango duena.

- **2. helburua:** hitzen sentimenduen balentzian eragiten duten euskarazko testuinguruko balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.1 helburua:*** fonologiako balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.2 helburua:*** morfologiako balentzia-aldatzaileak identifikatzea eta beren eragina neurtzea.
  - ***2.3 helburua:*** sintaxian, ezeztapen-markek hitzaren edo esaldiaren sentimendu balentzian nola eragiten duten neurtzea.
  - ***2.4 helburua:*** diskurtsoan, erlaziozko diskurtso-egituretan nukleartasunaren eta iritzi-testuetan unitate zentralaren eragina neurtzea orientazio semantikoari dagokionez.
  - ***2.5 helburua:*** iritzi-testuetan, erlaziozko diskurtso-egiturek testuaren posizio jakin batean agertzeko joera duten aztertzea.
  - ***2.6 helburua:*** iritzi-testuetako unitate zentralen azterketa: hitzen gramatika-kategoriaren azterketa eta orientazio semantikoa duten hitzen banaketaren lanketa.

1.4 Argitalpenak

Tesi-lan honetan zehar garatutako zenbait lan aldizkari eta kongresu ezberdinetan aurkeztu dira. Atal honetan, argitalpen horiek zerrendatu ditugu tesi-laneko gaiaren arabera sailkatuta.

- **Corpusa**
  - Alkorta J., Gojenola K. eta Iruskieta M. Creating and evaluating a polarity-balanced corpus for basque sentiment analysis. In *Fourth International Workshop on Discourse Analysis*, 54-58, Santiago de Compostela, Spain, 2016a. ISBN 978-84-608-9305-9.
- **Sentimendu-hitzak eta sentimendu-lexikoia**
  - Alkorta J., Gojenola K. eta Iruskieta M. SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas* 22(4): 1295-1306, 2018b, ISSN 2007-9737, DOI 10.13053/CyS-22-4-3075
  - Alkorta J., Gojenola K. eta Iruskieta M. Sentimenduak deskribatzeko hurbilpen teoriko konputazionala euskaraz. *UPV/EHUko I. Doktorego Jardunaldiak. Elkarrekin ikertuz*: 245-246, 2017a, Euskal Herriko Unibertsitateko Argitalpen Zerbitzua, ISBN 978-84-9082-619-5.
  - Alkorta J., Gojenola K., Iruskieta M. eta Taboada M. Using lexical level information in discourse structures for Basque sentiment analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, 39-47, Santiago de Compostela, Spain, 2017b. ISBN 978-1-945626-78-4.
- **Ezeztapena**
  - Alkorta J., Gojenola K. eta Iruskieta M. Saying no but meaning yes: negation and sentiment analysis in Basque. In Balahur A., Mohammad S. M., Hoste V., eta Klinger, R., editoreak, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 85-90, Brussels, Belgium, 2018a. The Association for Computational Linguistic. ISBN 978-1-948087-80-3.
- **Diskurtsoa**

  - Alkorta J., Gojenola K., Iruskieta M. eta Perez A. El uso de la información de la estructura retórica en el análisis de sentimiento. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., eta Romá Ferri M.T., editoreak, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015b. Sociedad Española para el Procesamiento del Lenguaje Natural. ISBN 978-84-608-1989-9.
  - Alkorta J., Gojenola K. eta Iruskieta M. Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena. *Gogoa: Euskal Herriko Unibersitateko hizkuntza, ezagutza, komunikazio eta ekintzari buruzko aldizkaria (Xabier Arrazola Gogoan (1962-2015))*:14, 131-152, 2016b, ISSN 1577-9424
  - Alkorta J. El uso de información del discurso en el análisis de sentimientos en euskera. In Almela A., Alcaraz-Mármol G., Valencia R., Fernández G. T. *Proceedings of Doctoral Symposium of the 33rd Conference of the Spanish Society for Natural Language Processing (SEPLN 2017)*, Murcia, Spain, 2017c. ISSN 1613-0073
  - Alkorta J., Gojenola K. eta Iruskieta M. 2019b. Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, 144-152, Association for Computational Linguistics. ISBN 978-1-948087-98-8

- **Testuinguruko balentzia-aldatzaileak oro har**

  - Alkorta J., Gojenola K. eta Iruskieta M. 2019a. Sentimenduen tratamendu konputazionalerantz: gramatika maila ezberdineta-ko sentimendu balentzia-aldatzaileen bila. Olatz Arbelaitz, Urtzi Etxeberria, Ainhoa Latatu, Miren Josu Ormaetxebarria (arg.), *III. Ikergazte. Nazioarteko Ikerketa Euskaraz, Giza Zientziak eta Artea* (1. liburukia), 39-46. Udako Euskal Unibertsitatea (UEU). Bilbo. ISBN: 978-84-8438-682-7

- **Bestelakoak**

  - Alkorta J. Euskararako HPSG gramatikaren lehen proposamena. Alegria I., Latatu A., eta Omaetxebarria M.J., editoreak, *I. Ikergazte, Nazioarteko ikerketa euskaraz*, 24-31, Bilbao, Spain, 2015a. Udako Euskal Unibertsitatea. ISBN 978-84-8438-540-0.

- **Bidalitakoak**

  - Alkorta J., Taboada M., Gojenola K. eta Iruskieta M. Bidalita. The role of hierarchical structure and coherence relations in sentiment analysis: a study in Basque. *Journal of Corpora and Discourse Studies*

1.5 Tesi-lanaren antolakuntza

Tesi hau zazpi kapituluz osatuta dago. Hurrengo lerroetan, kapitulu bakoitzaren edukia laburbilduta azalduko dugu.

1. kapitulua: proiektuaren nondik norakoak.
Lehen kapituluan, lanaren motibazioa, hipotesi orokorrak, helburuak eta tesiarekin lotutako argitalpenak azaldu ditugu.

2. kapitulua: sentimenduen analisiko baliabideak eta testuinguruko balentzia-aldatzaileak.
Bigarren kapitulua bost atalez osatuta dago. Lehenik eta behin, sentimenduen analisiaren arloan dauden atazak aipatu ditugu. Gero, ataza horiek gauzatzeko behar diren baliabideetako bi (iritzi-testuen corpusak eta sentimenduen lexikoiak) azaldu ditugu. Ondoren, testuinguruko balentzia-aldatzaileak aztergai dituzten lanen inguruan jardungo dugu. Laugarren atalean, sentimenduen analisia eta euskara uztartzen dituzten lanak aipatu ditugu. Azkenik, kapitulu guztia laburbildu dugu.

3. kapitulua: metodologiaren diseinua.
Kapitulu honetan, lehenik eta behin, euskarazko iritzi-testuen corpusa eta euskarazko sentimenduen lexikoia sortzeko urratsak azaldu ditugu. Ondoren, berriz, euskarazko balentzia-aldatzaileak identifikatzeko metodologia zehaztu dugu. Fonologian, morfologian, sintaxian eta diskurtsoan egin dena bananbanan aipatu dugu. Gero, dokumentu mailako sentimenduen sailkatzailea

den SO-CAL tresnaren ingelesezko bertsioa euskaratzeko prozesua nolakoa izan den azaldu dugu. Azkenik, kapituluaren laburpena egin dugu.

4. kapitulua: sentimenduen analisirako baliabideak.
Laugarren kapituluan, tesi-lan honetan zehaztutako metodologia jarraitzearen ondorioz garatutako baliabideen eta beren ebaluazioaren berri eman dugu. Aipatu ditugun baliabideen artean, Euskarazko Iritzi Corpusa, *Sentitegi* izeneko sentimenduen lexikoia eta dokumentu mailako euskarazko sentimenduen sailkatzailea daude. Bukatzeko, kapituluaren laburpena egin dugu.

5. kapitulua: balentzia-aldatzaileak.
Bosgarren kapitulu honetan, lorturiko emaitzen berri eman dugu. Baina aurreko kapitulukoan ez bezala, hemengo emaitzak teorikoak dira. Zehazki, maila fonologiko eta morfologikoan, sintaktikoan eta diskurtsokoan aurkitu ditugun testuinguruko balentzia-aldatzaileak azaldu ditugu eta hitzen edota sintagmen orientazio semantikoan edo sentimenduen balentzian zer-nolako eragina duten ere aipatu dugu. Kapituluarekin amaitzeko, laburpena egin dugu.

6. kapitulua: ekarpenak, mugak eta etorkizuneko lanak.
Kapitulu honek tesi-lanaren ondorioak aipatuko ditu lehenik. Hasierako helburuak bete diren aipatu dugu eta ikerketa-galderei erantzuna emango diegu. Lanaren ekarpenak ere aipatu ditugu. Azkenik, etorkizuneko lanen inguruan aritu gara.

Terminologia eta laburdurak.
Eranskin honetan, tesi-txosten honetan aipatutako terminologia eta laburdurak zerrendatu ditugu.

AURREKARIAK

2. KAPITULUA

Sentimenduen analisiko baliabideak eta testuinguruko balentzia-aldatzaileak

Aurrekarien kapitulu hau bost zatitan banatuta dago. Lehenik eta behin, 2.1 atalean, sentimenduen analisiaren ataza nagusiak zeintzuk diren esplikatzen da. Ondoren, 2.2 atalean, iritzi-testuen corpusak eta sentimenduen lexikoiak nola sortu diren azaltzen da. 2.3 atalean, berriz, testuinguruko balentzia-aldatzaileekin egin diren lanak aipatzen dira. Euskara eta sentimenduen analisia uztartzen dituzten lanak 2.4 atalean zehazten dira. Azkenik, 2.5 atalean, kapitulu osoa laburtuta dago.

Kapitulu honetan, 2.1 eta 2.2 atalak Liuren (2012) lanean eta bertako sailkapenean oinarrituz egin dira. Sentimenduen analisia eta iritzien erauzketa arloen barnean, hainbat ataza eta hurbilpen daude eta egile bakoitzak sailkapen ezberdin bat egiten du. Guk Liuren (2012) sailkapenaren alde egin dugu; izan ere, arloaren ikuspegi orokorrago bat emateaz gain, garrantzi berezia ematen dio lexikoian oinarritutako edo hizkuntzalaritzaren ikuspegitik gertuago dauden hurbilpenei. Sentimenduen analisiaren inguruan egin diren beste sailkapen batzuk honako lanetan aurki daitezke: Pang eta Lee (2008), Vinodhini eta Chandrasekaran (2012), Cambria *et al.* (2017), Mohammad (2016) edo Westerski (2007), besteak beste.

2.1 Sentimenduen analisiaren atazak

Sentimenduen analisian, jarraian azaltzen diren hiru ataza nagusi bereizi ohi dira eta ataza horietako bakoitza hizkuntza maila batekin (hitz, esaldi eta diskurtso mailekin) lotuta dago (Liu, 2012).

- Entitate eta aspektu mailako ataza: ezaugarri maila (*feature level*, ingelesez) lantzen da. Hau da, iritzia positiboa edo negatiboa den jakiteaz gain, iritziaren jomuga (entitatea eta bere aspektuak) identifikatu nahi da. Beraz, ataza hau hitz mailakoa da. Lan aipagarrien artean Wiebe *et al.*ena (1999) dago.

- Esaldi mailako ataza: subjektibitatearen sailkapena (*subjectivity classification*, ingelesez). Ataza honetan, esaldiak informazio subjektiboa edo objektiboa (hots, informazio faktuala) duen bereiztea da helburua. Ataza hau ere lantzen duen lanetako bat Wiebe *et al.*ena (1999) da.

- Dokumentu mailako ataza: dokumentu mailako sentimenduen sailkapena (*document-level sentiment classification*, ingelesez). Ataza honen helburua dokumentu baten subjektibitateak balorazio positiboa edo negatiboa duen identifikatzea da. Ataza honetan aipagarriak dira Pang *et al.*ren (2002), Turneyren (2002) eta Pang eta Leeren (2008) lanak. Sentimenduen analisiko hiru ataza nagusi hauen artean, gure lanak dokumentu mailakoarekin du lotura; izan ere, dokumentu mailako sentimenduen sailkatzailea garatzea dugu helburu.

- Sentimenduen analisiko beste atazak. Ataza hauek sentimenduen analisiaren eta HPko beste atazen arteko nahasketa dira. Atazak mota askotarikoak dira: iritzi-testuen laburpena, konparaziozko iritzien analisia, iritzi-bilaketa eta -berreskuraketa, iritzi *spam*en detekzioa eta iritzien kalitatearen neurketa.

Jarraian, lau atazak banan-banan esplikatuko ditugu.

2.1.1. Entitate eta aspektu mailako sentimenduen analisia

Sentimenduen analisiko atazetako bat entitate eta aspektu mailako sentimenduen analisia da. Ataza honetan, iritzia positiboa edo negatiboa den jakiteaz gain, iritziaren hartzailea zein den ere jakin nahi da.

Entitate bati buruzko iritzi bat ematean, entitate horren eta aspektu guztien iritziak beti ez datoz bat. Hori dela eta, beharrezkoa da entitatearen aspektu bakoitzak nolako iritzia duen jakitea. Adibidez, film bat entitate bat da eta bere aspektuak izango lirateke pertsonaiak, atrezzoa, soinu-banda eta argumentua, besteak beste. Aspektu horietako batzuk positiboak izan daitezke eta beste batzuk, berriz, negatiboak. Entitate horren eta aspektu horien orientazio semantikoa identifikatzea da ataza honen motibazioa.

Aspektuak hitz edo sintagma mailan agertzen dira. Horregatik, aspektu bakoitzaren orientazio semantikoa lortzeko, hitz edo sintagmaren orientazioa kalkulatzen da. Horretarako, beharrezkoa da iritziak erreferentzia egiten dion entitatea bere aspektuetan deskonposatzea.

Aspektuan oinarritutako sentimenduen analisian sei urrats daude. Urratsak esplikatzeko Liuren (2012) (5) adibidea erabiliko dugu.

(5) Posted by: bigJohn Date: Sept. 15, 2011. [1] I bought a Samsung camera and my friends brought a Canon camera yesterday. [2] In the past week, we both used the cameras a lot. [3] The photos from my Samy are not that great, and the battery life is short too. [4] My friend was very happy with his camera and loves its picture quality. [5] I want a camera that can take good photos. [6] I am going to return it tomorrow.

1- Entitateen erauzketa eta kategorizazioa. Entitate guztiak erauzi eta ezaugarrietan oinarrituta multzokatu egiten dira. (5) adibidean, *Samsung*, *Samy* eta *Canon* entitateak erauzten dira eta *Samsung* eta *Samy* entitateak multzokatzen dira entitate bera errepresentatzen dutelako.

2- Aspektuen erauzketa eta kategorizazioa. Aspektu guztiak erauzi eta

multzokatu egiten dira. (5) adibidean, *picture*, *photo* eta *battery life* aspektuak erauzten dira eta *picture* eta *photo* aspektuak multzokatzen dira automatikoki, kamerari dagokionez sinonimoak direlako.

3- Iritziaren jabearen erauzketa eta kategorizazioa. Iritzi-hartzaile guztiak erauzi eta beren ezaugarrietan oinarrituta, kategorizatu egiten dira. (5) adibidean, iritziaren jabea [3] esaldian *bigJohn* (blogaren egilea) da eta [4] esaldian, aldiz, iritziaren jabea *bigJohn*en laguna da.

4- Denboraren erauzketa eta kategorizazioa. Denbora-adierazpenak erauzi, formatu bakar batean jarri eta sailkatu egin behar dira. Aztertzen ari garen adibidean, erauzitako denbora-adierazpena Sept-15-2011 da.

5- Aspektuan oinarritutako sentimenduen analisia. Urrats honetan egiten da sentimenduen analisia, hots, iritziaren aspektuak orientazio positiboa edo negatiboa duen zehazten da eta iritziaren hartzaileari kalifikazio bat esleitzen zaio.

   [3] esaldian, Samsung kameraren irudi-kalitateari eta bateria-bizitzari iritzi negatiboa ematen zaio. [4] esaldian, iritzi positiboa ematen zaio bai Canon kamerari, bai haren irudi-kalitateari. Gainera, esaldi honetan, *its* eta *this camera* hitzek nori egiten dioten erreferentzia ere jakin behar da. Azkenik, [5] esaldian, iritziak positiboa dela ematen du, baina ez da horrela. Esaldi horretan, hiztunak bere desio edo nahi bat adierazten baitu eta ez objektu baten (kasu honetan, kamera baten) iritzia edo ebaluazioa.

6- Aspektu bakoitzari lotutako iritziaren sorkuntza. Azken urratsean, aspektuak, aspektuari loturiko entitatea, aspektuari buruzko iritzia, iritziaren jabea eta iritzia eman den denbora zehazten dira.

   (Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
   (Samsung, battery_life, negative, bigJohn, Sept-15-2011)
   (Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
   (Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

Hu eta Liuren (2004) lana erreferentziazkoa ataza honetan. Izan ere, bezero askoren iritzi-testuak laburtzeaz gain, laburpen horietatik balorazioren

batzuk dauden zatiak bakarrik hartzen dituzte. Horrela, iritzi datu-base bat hartu eta, lehenik eta behin, POS etiketatzea, ezaugarrien maiztasunaren identifikazioa eta ezaugarrien kimatzea burutzen dute. Ondoren, ohiko ezaugarrietatik iritzidun hitzak lortzen dituzte eta azken honetatik, ezohiko ezaugarriak ere lortzen dituzte. Gero, iritzidun hitzak eta ezohiko ezaugarriak baliatuz, iritzidun esaldiaren orientazio identifikazioa egiten dute eta azkenik, berriz, iritzi-testuaren laburpen bat lortzen dute.

2.1.2. Esaldi mailako sentimenduen analisia

Sentimenduen analisiko ataza honek esaldi mailako sentimenduen sailkapena egiten du. Ataza honek bi urrats ditu: i) subjektibitatearen sailkapena eta ii) sentimenduen sailkapena.

Ataza lantzeko azter dezagun Liuren (2012) adibidea hau:

(6) "[1] I bought a Motorola phone two weeks ago. [2] Everything was good initially. [3] The voice was clear and the battery life was long, although it is a bit bulky. [4] Then, it stopped working yesterday."

Subjektibitatearen sailkapena izeneko atazean, esaldiak iritzia adierazten duen edo ez zehazten da. Esaldiak iritzia ez du adieraziko baldin eta bertan azaltzen den informazioa faktuala edo objektiboa baldin bada (hau da, informazioa egiazkoa bada). Esaldi batean iritzia, ebaluazioa, emozioa, usteak, espekulazioa, epaia, salaketa edota jarrera, besteak beste, agertzen bada, aldiz, esaldia subjektiboa da. Hala ere, batzuetan posible da esaldi objektibo batek ere iritzia adieraztea[1].

(6) adibidea aztertzen badugu, [1] esaldiak informazio faktuala du; beraz, esaldia objektiboa da. [2] eta [3] esaldiek, berriz, iritziak adierazten dituzte *good*, *clear*, *long* eta *bulky* adjektiboak agertzen baitira. Azkenik, [4] esaldia iritzia adierazten duen esaldi objektiboa da; izan ere, bateriak funtzionatzeari uztea informazio faktuala da baina zerbait negatiboa da.

[1]Esaterako, *mugikor honek bateria azkar xahutzen du* esaldia objektiboa da baina iritzi negatiboa ere badu.

Lehen urratsean, subjektibitatearen sailkapena egiteko hainbat teknika aipatu ditugu. Teknika horiek 2.1 taulan laburbilduta eta, ondoren, azalduta daude.

Lana	**Teknika**
(Wiebe *et al.*, 1999)	Ikasketa gainbegiratua
(Wiebe, 2000)	Ikasketa gainbegiratu gabea
(Yu eta Hatzivassiloglou, 2003)	Esaldien arteko antzekotasuna eta Naïve Bayes
(Riloff *et al.*, 2006)	Unigrama, n-grama eta patroi lexiko-sintaktikoen arteko harremanak

2.1 taula: Esaldien subjektibitatearen sailkapena egiteko teknika ezberdinak.

Wiebe *et al.*ek (1999) subjektibitatearen sailkapena egiteko urre-patroia nola sortzen duten azaltzen dute. Etiketatzean desadostasunak agertu zaizkienez, aurreiritziak zuzentzeko etiketak sortzen dituzte bi helbururekin: i) eskuz etiketatzeko gidalerroen bertsio hobea sortzeko eta ii) sailkatzaile automatikoa sortzeko.

Urtebete geroago, Wiebek (2000) subjektibitatearen sailkapena egiteko ikasketa gainbegiratu gabea[2] egiten du. Kasu honetan, hitz-multzoen distribuzio semantikoa aintzat hartuz, subjektibitateari loturiko aztarnak lortzen ditu.

Yu eta Hatzivassiloglouk (2003), berriz, iritzidun esaldiak bilatzeko hiru teknika ezberdin erabiltzen dituzte: i) esaldien arteko antzekotasuna, ii) Naïve Bayes sailkatzailea[3] eta iii) Naïve Bayes anizkuna.

Bukatzeko, Riloff *et al.*ek (2006) ezaugarrien errepresentazioa egiteko sub-

[2]Ikasketa gainbegiratu gabea ikasketa automatikoko metodo bat da, behaketan oinarritzen den eredua da. Bertan, modelatze prozesu guztia sistemako sarreran dauden adibide multzo baten bidez gauzatzen da. Ez du adibide horien kategorien inguruko informaziorik. Hori horrela izanik, sistemak gauza izan behar du patroiak identifikatu eta, modu horretan, sarrera berriei etiketak esleitzeko.

[3]Naïve Bayes sailkatzaile probabilistiko bat da eta Bayes-en teoreman eta aldagaien arteko independentziaren hipotesian oinarritzen da. Sailkatzaile honen arabera, klase aldagai bat emanda, ezaugarriak elkarren artean independienteak dira eta, beraz, ez dago korrelaziorik.

suntzio hierarkia (*subsumption hierarchy*) erabiltzen dute, bertan, ezaugarri lexikal ezberdinak eta beren arteko harremanak finkatzeko. Modu horretan, esaldien subjektibitatearen sailkapena lortzen dute.

Bigarren urratsean, berriz, sentimenduen sailkapena egiten da eta esaldiak iritzia adierazten badu, iritzia positiboa edo negatiboa den zehatzen da.

(6) adibidearen kasuan, [2] esaldiak iritzi positiboa du, *good* ("ona") hitza dagoelako, baina [3] eta [4] esaldiek iritzi negatiboa dute: [3] esaldiaren kasuan, *bulky* ("tamaina handikoa") hitza ageri da eta [4] esaldian, azkenik, mugikorrak funtzionatzeari utzi diola aipatzen da.

Sentimenduen sailkapena egiteko ere hainbat teknika daude eta horiek 2.2 Taulan laburbilduta eta, ondoren, azalduta daude.

Lana	**Teknika**
(Yu eta Hatzivassiloglou, 2003)	Egiantz-arrazoiaren logaritmoa (*log-likelihood ratio*, LLR)
(Hu eta Liu, 2004)	Lexikoian oinarritutako algoritmoa
(Gamon, 2004)	Ikasketarako algoritmo erdi-gainbegiratua
(McDonald *et al.*, 2007)	Sekuentzia hierarkikoen ikasketa automatikoa

2.2 taula: Esaldi mailako sentimenduen sailkapena egiteko zenbait teknika.

Sentimenduen sailkapenaren urratsa gauzatzeko, Yu eta Hatzivassiloglouek (2003) hitzen orientazio semantikoa kalkulatzen dute eta horretarako Turneyk (2002) proposaturiko teknika erabiltzen dute. Hurrengo urratsean, esaldi osoaren orientazio semantikoa kalkulatzeko, berriz, egiantz-arrazoiaren logaritmoa[4] erabiltzen dute.

[4]Egiantz-arrazoiaren logaritmoak (*log-likelihood ratio*, ingelesez) bi eredu estatistikoen egokitzeko egokitasuna ebaluatzen du. Bi ereduok beren probabilitateen proportzioaren arabera lehiatzen dira; zehazki, bata parametroen espazio osoaren maximizazioarekin aurkitzen da eta bestea murriztapen batzuk ezarri ondoren aurkitzen da. Behatutako datuek murriztapena babesten badute (hau da, hipotesi nulua) bi probabilitateek ez dute laginketa-errore bat baino gehiagotan diferitu behar. Beraz, egiantz-arrazoiak erlazio hau bata bestearengandik oso ezberdina den edo bere logaritmo naturala zerotik guztiz ezberdina den ziurtatzen du.

Hu eta Liuek (2004), aldiz, hiru urrats egiten dituzte sentimenduen sailkapena gauzatzeko: i) kontsumitzaileen iruzkinetan agertu diren produktuaren ezaugarriak erauzi, ii) iruzkin bakoitzean iritzidun esaldia identifikatu eta positiboa edo negatiboa den adierazi eta iii) emaitzak laburbildu.

Bestalde, Gamonek (2004) ikasketa erdi-gainbegiratuen teknika erabiltzen du. SMV algoritmoa ezaugarri askoko bektoreekin erabiltzen du, zeintzuek bektore murrizketa jasan duten. Halaber, azterketa linguistiko sakonaren ezaugarriak hitzen n-grama funtzioei ere aplikatzen dizkie.

Azkenik, McDonald *et al.*ek (2007) testuaren granularitate-maila ezberdinetan testuen sailkapen semantikoa egiteko eredu egituratuak erabiltzen dituzte. Ereduaren inferentzia Viterbiren sekuentzia estandarren sailkapen tekniketan oinarritzen dira.

Ikusten den moduan, teknika ezberdinak daude esaldi mailako sentimenduen analisia egiteko, hots, esaldien iritzi positiboa edo negatiboa duten jakiteko.

2.1.3. Dokumentu mailako sentimenduen sailkapena

Aurretik esan bezala, ataza honetan dokumentu baten eduki subjektiboa aztertzen da, duen balorazioa positiboa edo negatiboa den identifikatuz. Ataza honetarako hainbat hurbilpen daude (Liu, 2012): i) sentimenduen sailkapena ikasketa gainbegiratua erabiliz (2.1.3.1 atala), ii) sentimenduen sailkapena ikasketa gainbegiratu gabea erabiliz (2.1.3.2 atala), iii) sentimenduen estimazio iragarpena (2.1.3.3 atala), iv) domeinu arteko sentimenduen sailkapena eta (2.1.3.4 atala) v) hizkuntza arteko sentimenduen sailkapena (2.1.3.5 atala).

2.1.3.1. Sentimenduen sailkapen gainbegiratua

Sentimenduen sailkapena ikasketa gainbegiratua egitean, dokumentu baten sentimendu sailkapen bitarra egiten da, hots, haren iritzia positiboa edo negatiboa den zehazten da. Zehaztapen hori kalifikazio bat emanez egiten da. Ikasketa gainbegiratuan, entrenamenduko eta testeko datuak iritzi-testuak dira eta bertan, iritzia 1 eta 5 bitarteko eskala erabiliz adierazten da non 1 eta 2 kalifikazioak negatiboak diren eta 4 eta 5 kalifikazioak, aldiz, positiboak.

Ikertzaile gehienek ez dute iritzi neutroa aintzat hartzen, modu horretan, sailkapena egitea erraza delako, baina beharko balitz, 3 kalifikazioa izango litzateke iritzi neutroa.

Sentimenduen sailkapena bereziki dokumentu mailako arazo bat da. Testu-sailkapen tradizionalean, hots, testu bat domeinu batean sailkatzeko arazoan, domeinu ezberdinetako dokumentuak sailkatu izan dira: politika, zientzia eta kirola, besteak beste, eta mota horretako sailkapenetan, domeinuari loturiko hitz gakoak garrantzitsuak dira, modu horretan, testua gai nagusiari edo bigarren mailako gai bati buruzko iritzia ematen ari den jakin daitekeelako. Baina sentimenduen sailkapenean, iritzi positiboa edo negatiboa adierazten duten sentimendu eta iritzidun hitzak garrantzitsuagoak dira (esaterako, *ederra*, *bikaina*, *txarra*), Liuren (2012) arabera.

Dokumentu mailako sentimenduen sailkapena testu-sailkapenaren ataza bat denez, ikasketa gainbegiratuan dauden teknika ezberdinak erabil daitezke. Dokumentu mailako sentimenduen sailkapena egiteko ikasketa automatiko gainbegiratuko aplikazioek zenbait ezaugarri erabili izan dituzte:

1- Hitzak eta horien maiztasunak. Banakako hitzek (unigramek) eta beren n-gramak[5] maiztasun-kontaketarekin osatzen dute ezaugarri hau. Batzuetan, posizioa ere kontutan hartzen da. Ezaugarri honek eraginkortasun handia erakusten du sentimenduen analisian.

2- *Part-Of-Speech* (POS). Hitz bakoitzaren gramatika-kategoria jakitea ere garrantzitsua da; izan ere, sentimenduen analisian gramatika-kategoria bakoitzak garrantzi ezberdina du. Hala, adjektiboak iritziaren adierazle garrantzitsuak dira. Adjektiboek halako garrantzia dutenez, batzuetan ezaugarri berezi moduan tratatzen dira. Beste batzuetan, aldiz, n-gramak eta POS batera erabiltzen dira ezaugarri berean.

3- Sentimendu-hitzak eta -sintagmak. Sentimendu hitzak iritzi positiboak eta negatiboak adierazteko erabiltzen dira. Adibidez, *on*, *zorioneko* eta

[5]N-grama sekuentzia baten parte den azpi-sekuentzia bat da. Azpi-sekuentzia hori hainbat n elementuz osatuta dago eta elementu horiek hainbat motatakoak izan daitezke; letrak edo hitzak, esaterako.

eder hitz positiboak dira. *txar*, *zoritxarreko* eta *itsusi*, aldiz, hitz negatiboak dira. Sentimendu-hitz gehienak adjektiboak eta adberbioak dira baina izenak (*maitasuna*) edota aditzak (*gorrotatu*) ere izan daitezke. Sentimendu-hitzez gain, sentimendu-sintagmak ere badaude; *leher eginda* lokuzioa oso nekatuta zaudela adierazteko bezalakoak.

4- Iritzi-erregelak. Sentimendu-hitz eta -sintagmez gain, badaude iritziak adierazteko zenbait adierazmolde iritzia edo sentimendua adieraz edo inplika dezaketenak. Esaterako, *bateria-bizitza luzea du* eta *bateria-bizitza laburra du* esaldietan, aurkako bi orientazio semantiko daude; baina halaber, *bateria-bizitza* hitzak berak orientazio semantiko positiboa inplikatzen du. Hau da, mugikorrak *bateria-bizitza* bera edukitzea positiboa da baina horren baitan, *bateria-bizitza luzea* positiboa den bezala *bateria-bizitza laburra* negatiboa da.

5- Sentimendu-aldatzaileak. Badaude zenbait adierazmolde orientazio semantikoa aldatzeko erabiltzen direnak. Orientazio semantiko positiboa negatibo bilaka dezakete edota alderantziz. Sentimenduen aldatzaileetako bat ezeztapena da. *Kamera hau ez zait gustatzen* esaldian, esaterako, *gustatu* hitzaren orientazio semantiko positiboa negatibo bihurtzen du.

6- Dependentzia sintaktikoak. Dependentzia-zuhaitzetatik sortutako hitzen dependentzia-ezaugarriak ere erabiltzen dira dokumentu mailako sentimendu-sailkapenean.

Dokumentu mailako sentimenduen sailkapena egiteko erabilitako teknikak 2.3 Taulan daude laburbilduta.

Aurretik aipatutako ezaugarriak erabiliz, zenbait teknika erabili izan dira. Pang *et al.*ek (2002) filmeen iruzkinak testu positibo eta negatibo bezala sailkatzen dituzte eta horretarako n-gramak (hitz-poltsak) erabiltzen dituzte sailkapeneko ezaugarri moduan Naïve Bayes nahiz SVM sailkatzaileak erabiliz.

Lana	Teknika
(Pang *et al.*, 2002)	1) Naïve Bayes sailkatzailea 2) Euskarri bektoredun makina (*Support Vector Machines*, SVM)
(Mullen eta Collier, 2004)	SVM + Erlazio sintaktikoak eta ohiko ezaugarriak
(Nakagawa *et al.*, 2010)	Dependentzia-zuhaitzean oinarritutako sailkatzailea + baldintzazko ausazko eremuak (*Conditional Random Fields*), CRF
(Kouloumpis *et al.*, 2011)	Ezaugarri linguistikoak + mikroblogetako hizkuntza kreatiboa eta ez-formala identifikatzeko ezaugarria

2.3 taula: Dokumentu mailako sentimenduen sailkapena egiteko erabilitako zenbait teknika.

Mullen eta Collierek (2004) SVM sailkatzailea[6] hainbat ezaugarrirekin konbinatzen dute, hala nola PMI[7] bidez lorturiko orientazio semantikoak, Wordneten oinarrituz egindako desberdintze semantikoa eta gaiarekiko hurbiltasuna eta erlazio sintaktikoen ezaugarriak.

Nakagawa *et al.*ek (2010), berriz, ingeleseko eta japonierako datu subjektiboak sailkatzeko dependentzia-zuhaitzak eta baldintzazko ausazko eremua[8]

[6]Euskarri bektoredun makina (ingelesez, *Support Vector Machine*, SVM) sailkapen nahiz erregresiorako erabiltzen den algoritmo multzoa da. Bere oinarrian bektore-espazioak daude eta sarreran, bere n dimentsio ereduan adieraz daitezkeen hainbat datu jasotzen ditu, bi kategorietatik batean sailkatuta daudenak. Ondoren, teknika honek bere ereduan dauden puntuak bereizteko hiperplanoa bilatzen du. Teknika honen kasuan gertueneko puntuetarako distantziarik handiena duena aukeratzen du eta hori da emaitza.

[7]Elkarrekiko Informazio Puntuala (ingelesez, *Pointwise mutual information*) estatistikan eta informazioaren teorian erabiltzen den asoziazio neurria da. Hitz batek (x) beste hitz batekin (y) duen probabilitatea hitzaren beraren (x) probabilitatearekin zatituz lortzen da. Esaterako, *puerto* hitza 1.938 aldiz agertzen da eta *rico* 1.311 aldiz. Biak batera, berriz, 1.159 aldiz agertzen dira. Horrenbestez, PMI neurria 10,03 da.

[8]Baldintzazko ausazko eremuak (ingelesez, *Conditional Random Fields*, CRF) dokumentuetatik datuak erauzteko edo datu-sekuentziak segmentatu eta etiketatzeko erabiltzen den eredu estokastikoa da. Eredu honek datu-sekuentzia bat emanda elementuetako bakoitzari (*Oi*) etiketa bat (*Si*) ematen dio. Eredu sortzailea denez, behaketen eta eti-

erabiltzen dituzte. Lanean, esaldi bakoitzaren azpi-zuhaitz baten orientazio semantikoa aldagai bat da eta esaldi osoaren orientazio semantikoa aldagai horien arteko interakzioa kalkulatuz lortzen da.

Kouloumpis *et al.*ek (2011), berriz, txioen sentimenduen sailkapena egiteko, hainbat ezaugarri erabiltzen dituzte: n-gramen ezaugarriak, lexikoiko ezaugarriak (hitzak positiboak, negatiboak edo neutralak diren ere barne hartuz), gramatika-kategoria ezaugarriak eta, bukatzeko, mikroblogetako testuen ezaugarriak.

2.1.3.2. Sentimenduen sailkapen gainbegiratu gabea

Dokumentu mailako sentimenduen sailkapena ikasketa gainbegiratu gabea erabiliz egiten duten lanak ere badaude. Sentimendu-hitzak oinarrizko faktoreak dira sentimenduen sailkapenean eta horregatik, horiek modu gainbegiratu gabean erabil daitezke. Bi hurbilpen daude ikasketa gainbegiratu gabean: i) erregela sintaktikoetan oinarritutakoa eta ii) lexikoian oinarritutakoa.

Bi hurbilpenak argitzeko Liuren (2012) *this piano produces beautiful sounds* esaldia erabiliko dugu.

	Lehen hitza	**Bigarren hitza**	**Hirugarren hitza (ez erauzia)**
1	JJ	NN edo NNS	edozer
2	RB, RBR edo RBS	JJ	ez NN edo NNS
3	JJ	JJ	ez NN edo NNS
4	NN edo NNS	JJ	ez NN edo NNS
5	RB, RBR edo RBS	VB, VBD, VBN edo VBG	edozer

2.4 taula: Dokumentu mailako sentimenduen sailkapen gainbegiratu gaberako erregela sintaktikoak (Turney, 2002).

Turneyk (2002) erregela sintaktikoen hurbilpena jarraitzen du. Erregela sin-

keten probabilitatearen banaketa aldi berean modelatzen ditu eta baldintzazko ausazko eremuek behaketek baldintzatutako etiketen sekuentzia zuzenen probabilitatea modelatzen dute (*P(S—O)*). Beraz, eredu diskriminatzaile bat da.

taktikoak gramatika-kategorien POS etiketetan (Liu, 2012)[9] oinarritzen dira eta guztira hiru urrats daude hurbilpen honetan:

1. urratsa. Jarraian dauden bi hitz erauzi, baldin eta bi hitz horien POS etiketek 2.4 taulako patroi bat osatzen badute. Esaterako, bigarren patroiak adierazten du lehen hitzak adberbioa, bigarrenak adjektiboa eta hirugarrenak (erauzten ez denak) izena izan behar duela. *This piano produces beautiful sounds* esaldiaren kasuan, *beautiful sounds* dugu, lehen hitza adjektiboa da eta bigarrena, berriz, izena. Beraz, 2.4 taulako lehen patroiarekin bat egiten du eta bi hitzak erauzi egiten dira. JJ, RB, RBR eta RBS POS etiketak dituzten hitzek askotan iritzia adieraztea da patroiak erabiltzearen arrazoia eta berekin agertzen diren izen eta aditzek beren testuingurua dira. Izan ere, adjektiboek esaterako (*ikaragarri* kasu) testuinguruaren arabera sentimendu positiboa edo negatiboa izan dezakete.

2. urratsa. Aurretik erauzitako hitz-multzoaren orientazio semantikoa kalkulatzen da elkarrekiko informazio puntuala (PMI) neurria (2.1 irudia) erabiliz.

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right)$$

2.1 irudia: *Pointwise mutual information*, PMI.

2.1 irudiko PMI neurriak bi hitzen arteko dependentzia estatistikoa neurtzen du. Pr(term1∧term2) *term1* gertatzeko ko-okurrentzia probabilitatea eta Pr(*term1*)Pr(*term2*) bi hitzen ko-okurrentzia probabilitatea da, baldin eta bi hitzak estatistikoki independienteak badira.

(7) SO(phrase) = PMI(phrase, “beautiful”) - PMI(phrase, “ugly”).

[9]Hauexek dira erregeletan azaltzen diren laburpenen esanahiak: JJ: adjektiboa, RB: adberbioa, RBR: konparaziozko adberbioa, RBS: superlatibozko adberbioa, NN: izen singularra, NNS: izen plurala, VB: aditza erro moduan, VBD: aditza iraganaldian, VBN: aditza iragan partizipioan eta VBG: aditza orainaldian eta orainaldiko partizipioan.

Liuren (2012) *this piano produces beautiful sounds* sintagmaren orientazio semantikoa (SO) kalkulatzeko, (7) adibidean ikusten den moduan, erreferentzia positiboko hitza den *beautiful* eta negatibokoa den *ugly* hitzaren arteko asoziazioa neurtzen da.

Probabilitatea kalkulatzeko, bilaketa-motor[10] bati kontsulta eskatzen zaio eta bisita kopurua biltzen du hark. Bilaketa bakoitzeko kontsulta batean, bilaketa-motor batek kontsultari lotutako dokumentu kopuru aipagarriak eskaintzen ditu eta hori emaitza kopurua da. Beraz, bi hitz batera eta bereiz bilatuta 2.1 irudiko probabilitatea kalkula daiteke.

3. urratsa. Iritzi-testu bat emanda, algoritmoak iritzi-testuan dauden sintagma guztien orientazio semantikoa neurtzen du eta iritzi-testua positibotzat (orientazio semantiko guztien batezbestekoa positiboa bada) edo negatibotzat (orientazio semantiko guztien batezbestekoa negatiboa bada) sailkatzen du.

Taboada *et al.*ek (2011) ere ikasketa gainbegiratu gabea erabiltzen du. Bertan, dokumentu mailako sentimenduen sailkapena egiteko SO-CAL tresna aurkezten du[11]. Tresna honek hitz, esaldien eta testuen orientazio semantikoa kalkulatzen du.

Tresna honen abiapuntua −5 eta +5 arteko orientazio semantikoa duten hitzen lexikoia da. Guztira, lexikoiak lau gramatika-kategorietako hitzak biltzen ditu: izenak, adjektiboak, adberbioak eta aditzak. SO-CAL tresnak hitz horiek iritzi-testu batean bilatzean erauzi egiten ditu. Hurrengo urratsean, tresnak hizkuntzaren beste ezaugarri batzuk erauzten ditu, hala nola intentsifikatzaileak, ezeztapena edota puntuazio-zeinuak. Erauzketa egin ostean, lehen urratsean erauzitako hitzen sentimendu-balentziak aldaketa batzuk jasatzen ditu, aipatu berri ditugun hizkuntza fenomenoen ondorio. Aldaketa hori sentimendu-balentzia indartzea edo ahultzea izan daiteke.

[10]Bilaketa-motorra informazioa bilatzeko helburuarekin sortutako informazioa eskuratzeko sistema da. Baldintza batzuk zehazten dira motorrean eta horren araberako emaitzak lortzen dira zerrenda batean garrantzi handienetik txikienera.

[11]Sentimenduen sailkapena egiten duen SO-CAL tresnari buruzko informazio gehiago: atal hauetan dago: 4.3.1 atalean non tresnari buruzko informazio gehiago ematen den eta 4.3.2 atalean non funtzionamendu bera duen euskarazko bertsioa deskribatzen den.

Esaterako, intentsifikatzaileek hitzaren orientazio semantikoa indartu edo ahuldu dezake. Bestalde, sintagman edo esaldian ezeztapen-marka agertuz gero, horren orientazio semantikoa ahuldu egiten du tresnak, ezeztapen-markaren ±4 balentzia aplikatuz. Hau da, sintagma edo esaldiaren orientazio semantikoa negatiboa bada eta ezeztapena badago, ezeztapen-markaren +4 aplikatuz, haren orientazio semantikoa indartu egingo du. Sintagma edo esaldiaren orientazio semantikoa positiboa bada eta ezeztapena badago, aldiz, ezeztapen-markaren −4 balentziak, haren orientazio semantikoa ahulduko du.

Bukatzeko, SO-CAL tresnak puntuazio-zeinuak ere kontuan hartzen ditu eta, esaterako, komen artean dauden hitzei ez die orientazio semantikoa esleitzen, izenburuak edota aipuak bezalakoak direlako.

Liuren (2012) esaldia hurbilpen honetan aplikatzen badugu, lehenik eta behin, SO-CALen lexioiak esaldiko *beautiful* (+4) hitza erauzten du. Hurrengo urratsean, inguruan intentsifikatzailerik eta ezeztapenik ez dagoela antzematen du, eta horrenbestez, hitzaren orientazio semantikoari ez dizkio hizkuntza-fenomeno hauen eraginak aplikatzen. Bukatzeko, tresnak hitza komen artean ez dagoela egiaztatzen du eta horrenbestez, hitzaren orientazio semantikoa bere horretan mantentzen du. Esaldian, orientazio semantikodun beste hitzik ez dagoenez, *beautiful* hitzaren orientazio semantikoa da (+4) esaldiaren orientazio semantikoa.

Gure tesi-lanak sentimenduen analisi gainbegiratu gabearekin du zerikusia. Izan ere, garatu nahi dugun sentimendu-sailkatzailea lexikoian oinarritutakoa izango da.

2.1.3.3. Sentimenduen estimazio-iragarpena

Dokumentu mailako sentimenduen sailkapeneko beste arloetako bat sentimenduen estimazio-iragarpena (*Sentiment Rating Prediction*) da. Arlo honetan, dokumentu batek iritzi positiboa edo negatiboa duen sailkatzetik urrunago joan nahi da eta zenbakizko kalifikazio bat (1 eta 5 artean) esleitu nahi zaio. Kasu honetan, ataza erregresio[12] moduan trata daiteke; izan ere, kalifikazio-puntuazioak ordinalak dira.

[12]Erregresioa aldagai dependente baten eta independente zenbaiten arteko erlazioa zehaztu eta aztertzen duten teknika estatistikoa da.

Sentimenduen estimazio iragarpena-egiteko aipatuko ditugun teknika ezberdinak 2.5 taulan daude laburbilduta.

Teknika	**Lana**
SVM erregresioa	Pang eta Lee (2005)
Hitz-poltsen errepresentazioa	Qu *et al.* (2010)
Aspektuaren sentimenduen estimazio-iragarpena: Erregresio estandarra	Snyder eta Barzilay (2007)
Aspektuaren sentimenduen estimazio-iragarpena: Sare Bayesiarrean oinarritutako sailkatzailea	Long *et al.* (2010)

2.5 taula: Sentimenduen estimazio-iragarpena egiteko teknika ezberdinak.

Pang eta Leek (2005) SVM erregresioa erabiltzen du. Zehazki, klase anitzeko SVM sailkatzailea[13] erabiltzen du.

Bestalde, Qu *et al.*ek (2010) dokumentuko hitz-poltsen errepresentazioa darabilte iritzia duten n-gramen indarra neurtzeko. Berau ezberdina da ohiko hitz-poltsen errepresentaziotik. Teknika honetan iritzietako bakoitza hirukoitza da: sentimendu hitza, aldatzailea eta ezeztapena baititu. Esaterako, *ez oso ona* adibidean, *ez* ezeztapena da, *oso* aldatzailea eta *ona*, berriz, sentimendu-hitza. Modu honetan, sentimendu-hitzetik haratago dauden baino berengan eragiten duten fenomenoak identifika ditzakete.

Snyder eta Barzilayk (2007) beste ikuspegi batekin lantzen dute ataza. Iritzi bakoitzari zenbakizko kalifikazio bat esleitu beharrean, aspektuetako bakoitzari esleitzen diote kalifikazioa eta, horretarako, erregresio estandarra edo sailkatze-teknikak erabiltzen dituzte.

Azkenik, Long *et al.*ek (2010) iritzietako aspektu bakoitzari kalifikazioa esleitzeko sare Bayesiarra darabilen sailkatzailea erabiltzen dute.

[13]Klase anitzeko SVM teknikan, datuak bi kategoria baino gehiagotan sailkatzeko bi hurbilpen daude: i) kategoria bakoitza beste batzuetan zatitzen da eta denak konbinatzen dira eta ii) k(k-1)/2 ereduak sortzen dira non k kategoria kopurua den.

2.1.3.4. Domeinuen arteko sentimenduen analisia

Dokumentu mailako beste arlo bat domeinu arteko sentimenduen sailkapena da. Ikasketan erabilitako domeinuek eragin handia dute sentimenduen sailkatzaileetan. Horregatik, domeinu jakin bateko iritzi-testuekin ondo dabilen sentimenduen sailkatzaile bat, beste domeinu bateko iritzi-testuekin ez da ondo ibiltzen (Liu, 2012).

Hitz batek aurkako bi orientazio semantiko eduki ditzake bi domeinu ezberdinetan eta hori da arazo honen jatorria. Adibidez, *luzea* adjektiboaren kasuan, telefono mugikorren domeinuan bateria-bizitzari zuzenduta dagoenean orientazio positiboa du. Adjektibo berak, aldiz, zinemaren domeinuan orientazio negatiboa du: *filma luzea egin zitzaidan.*

Arlo honetan bi hurbilpen daude. Horietako bat iturri-domeinua (*source domain*) da. Gamon *et al.* (2005) da lan aipagarrietako bat hurbilpen honetan, etiketatutako datu ikasiak domeinu berri batean erabiltzen dira. Beste hurbilpena helburu-domeinua (*target domain*) da. Blitzer *et al.* (2007); Tan *et al.* (2005) dira lanik nabarmenenak eta kasu honetan, domeinu berri baterako ez dute etiketatutako daturik erabiltzen.

2.1.3.5. Hizkuntza arteko sentimenduen analisia

Orain arte deskribatu eta aipatu ditugun hurbilpenak eta lanak hizkuntza bati bideratuta egon dira. Hala ere, badira hizkuntza arteko sentimenduen analisia egiten duten hurbilpen eta teknikak ere, bai esaldi mailan, bai dokumentu mailan.

Esaldi mailako hizkuntza arteko sentimenduen analisian ere bi ataza daude: i) hizkuntza arteko subjektibitatea eta ii) sentimenduen sailkapena.

Arlo honetan, hizkuntza askotan baliabide falta dagoenez, ingelesetik beste hizkuntzetara sentimenduen analisirako baliabide bat automatikoki itzultzea da helburua. Liu (2012) lanaren arabera, hiru estrategia nagusi daude baliabideak itzultzerakoan:

1- Helburu-hizkuntzako test-multzoko esaldiak iturri-hizkuntzara itzuli eta bertan, iturri-hizkuntza iturriko hizkuntzan dagoen baliabidea edo sis-

tema erabiliz test-multzoko esaldiek informazio objektiboa edo subjektiboa duten sailkatu.

2- Iturri-hizkuntzako ikasketarako corpusa[14] helburu-hizkuntzara itzuli eta corpusean oinarritutako sailkatzailea sortu helburu-hizkuntzan.

3- Iturri-hizkuntzako sentimenduen lexikoia helburu-hizkuntzara itzuli eta lexikoian oinarritutako sailkatzailea sortu helburu-hizkuntzan.

Hurbilpen honetan aipaturiko lanak 2.6 taulan laburbilduta daude.

Teknika	**Lana**
Helburu-hizkuntzako datuak jatorri-hizkuntzara itzuli eta sentimenduen sailkapena egin	Kim eta Hovy (2006) Banea *et al.* (2008)
Sentimenduen sailkapena egiteko lexikoia helburu-hizkuntzara itzuli eta, gero, sentimenduen sailkapena egin	Kim eta Hovy (2006) Banea *et al.* (2008) Bautin *et al.* (2008)
Sentimenduen sailkapena jatorri-hizkuntzan eta, ondoren, itzulpena egin	Banea *et al.* (2008)
Eleaniztasunaren konparagarritasuna	Kim *et al.* (2010)

2.6 taula: Esaldi mailako hizkuntza arteko sentimenduen analisirako zenbait teknika.

Kim eta Hovyek (2006) bi esperimentu gauzatzen dituzte: i) alemanerazko emailak ingelesera itzuli eta itzulpenei ingelesezko sentimendu-hitzak aplikatu, horien orientazio semantikoa ikusteko eta ii) ingelesezko sentimendu-hitzak alemanerara itzuli eta itzulitakoa alemanerazko emailetan aplikatu.

Banea *et al.*ek (2008) hiru esperimentu egin dituzte: i) ingelesezko datu etiketatuak (iturri-hizkuntza) automatikoki errumanierara itzuli (helburu-hizkuntza), ii) jatorri-hizkuntza automatikoki subjektibitatez etiketatu eta ondoren helburu-hizkuntzara itzuli eta iii) helburu-hizkuntza jatorri-hizkuntzara itzuli eta bertan subjektibitatearen sailkapen tresna aplikatu.

[14]Ikasketa automatikoan, corpusa hiru zatitan banatu ohi da: garapena, ikasketa eta testa. Ikasketa zatia, hitzak berak adierazten duen moduan, ikasketa automatikoko teknikek esperientzia trebetasun edo jakintza bihurtzeko baliatzen dute.

Bautin *et al.*ek (2008), berriz, aspektuan oinarritutako hizkuntza arteko sentimenduen analisia egiten dute. Helburu-hizkuntza iturri-hizkuntzara itzultzen dute (ingelesera) eta ondoren, entitatea duen esaldi bakoitzari orientazio semantikoa esleitzen diote ingelesezko lexikoian oinarritzen den metodoa erabiliz.

Bukatzeko, Kim *et al.*ek (2010) eleaniztasunarem konpagarritasuna (ingelesez, *multilanguage-comparability*) metodoa aplikatzen dute. Hots, testu eleanitz bateko hitz-bikoteek subjektibitate esanahi bera dutela baliatuz, sailkapen emaitzen hitz-bikoteek duten adostasun-maila neurtzen dute.

Dokumentu mailako hizkuntza arteko sentimenduen analisia ere badago. Hizkuntza arteko sentimenduen analisia egiteko bi motibazio nagusi daude:

i) Ikertzaile batek sentimenduen analisirako sistema bere hizkuntzan sortu nahi du eta, bere hizkuntzan horrelakorik sortzeko baliabiderik ez dagoenez eta sistema gehienak ingelesez daudenez, ingeleseko sisteman oinarrituz bere hizkuntzan sistema sortzeko erabiltzen du.

ii) Aplikazio edo produktuetan, enpresek pertsonen iritziak jaso nahi dituzte eta, hori lortzeko, ingelesez duten sistema beste hizkuntzetara itzultzen dute.

Dokumentu mailan, aipatuko ditugun hizkuntza arteko sentimenduen analisirako teknika ezberdinak 2.7 taulan laburbilduta daude.

Teknika	**Lana**
Hainbat itzultzaile automatiko	Wan (2008)
Ko-ikasketa metodoa (SVM barne)	Wan (2009)
Transferentzia-ikaskuntza metodoa	Wei eta Pal (2010)
Gai-eredua	Guo *et al.* (2010)

2.7 taula: Dokumentu mailan, hizkuntza arteko sentimenduen analisia egiteko zenbait teknika.

Wanek (2008) ingeleseko baliabideak ustiatzen ditu txinerazko iritzi-testuen sentimenduen sailkapena egiteko. Bere algoritmoan, txinerazko iritzi-testuak

hainbat itzultzaile erabiliz ingelesera itzultzen ditu eta itzulitako testuei ingelesezko sentimenduen lexikoia aplikatzen die.

Wanek (2009), ordea, ko-ikasketa metodoa[15] erabiltzen du. Zehazki esanda, txinerazko iritzi-testuen sailkapena egiteko etiketatuta dagoen ingelesezko corpusa erabiltzen du modu gainbegiratu batean (SVM teknika erabiliz).

Wei eta Palek (2010), berriz, transferentzia-ikaskuntza metodoa[16] proposatzen dute hizkuntza arteko sentimenduen analisia egiteko. Itzulpen-prozesuan bertan, itzultzaile automatikoak sortzen duen zarata txikiagotzen du.

Azkenik, Guo *et al.*ek (2010) gai-ereduan oinarritzen den metodoa[17] proposatzen dute hizkuntza askotako iritzi-adierazpenak multzokatu eta, modu horretan, estatu askotako iritzien aspektuan oinarritutako sentimenduen sailkapena egiteko.

2.1.4. Sentimenduen analisiko beste atazak

Liuk (2012) adierazten duenez, dokumentu maila, esaldi maila eta entitate eta aspektu maila sentimenduen analisiko ataza nagusiak dira; izan ere, maila horietako bakoitza objektiboa edo subjektiboa den zehaztu, horien orientazio semantikoa zehaztu eta azkenik, iritziaren hartzailea eta entitate baten aspektu askoren orientazio semantikoa zehazten baita.

Hala ere, sentimenduen analisia zabalagoa da eta badaude bestelako atazak ere non sentimenduen analisia hizkuntza naturalaren prozesamenduko beste

[15]Ikasketa automatikoan, ko-ikasketa metodoa (ingelesez, *co-training*) etiketatutako datu-multzo txiki bat eta etiketatu gabeko datu-multzo handi bat erabiltzean datza.

[16]Transferentzia-ikaskuntza (ingelesez, *transfer learning*) ikasketa automatikoko ikerketa arazo bat da eta arazo bat ebatzi ostean lorturiko jakintza gorde egiten du; ondoren, arazo ezberdin baina antzeko batean aplikatzeko. Adibidez, autoak sailkatzeko erabilitako jakintza kamioiak sailkatzeko arazoan erabil daiteke. Psikologian dagoen transferentzia-ikaskuntzarekin nolabaiteko harremana du.

[17]Ikasketa automatikoan eta lengoaia naturalaren prozesamenduan gai-eredua (ingelesez, *topic model*) testu multzo batean ageri diren "gai" abstraktuak ezagutzeko eredu estatistiko mota bat da. Testuen baitan dauden ezkutuko egitura semantikoak ezagutzeko maiz erabiltzen den testu erauzketarako tresna da. Adibidez, eguraldia gaian den testuetan *eguzki* eta *tenperatura* hitzak maiztasun handiz agertzea espero da eta kirolean, aldiz, *irabazi* eta *talde*. Bestalde, *da* hitza bietan antzeko maiztasun batez agertzea espero da. Teknika honek sortzen dituen "gaiak" antzekotasuna duten hitz-multzoak dira.

arloekin uztartzen den. Ataza horiek gure tesi-lanaren helburuetatik urrutiago daudenez, azaletik esplikatuko ditugu.

2.1.4.1. Iritzi-testuen laburpena

Sentimenduen analisiko ataza ezberdinetan, pertsona askoren iritzia jakitea beharrezkoa da; bakoitzak bere ikuspegia duelako. Pertsona askoren iritzia bildu behar izateak iritzi-testuen laburpena egin beharra eragin du. Bestalde, aspektuan oinarritutako sentimenduen sailkapenak (2.1.1 atalean erakutsi bezala) iritzi-testuen laburpen bat egiteko adina informazio eskatzen du. Egoera honetan, iritzi-testuen laburpena eta aspektu mailako sentimenduen sailkapena batzearen ondorioz, aspektuan oinarritutako iritzi-testuen laburpena sortu da. Aspektuan oinarritutako iritzi-laburpenaren emaitza bi modutakoa izan daiteke: i) testu egituratua (2.2 irudian bezala) edota ii) testu desegituratua.

Digital Camera 1:
Aspect: **GENERAL**
Positive: 105 <individual review sentences>
Negative: 12 <individual review sentences>
Aspect: **Picture quality**
Positive: 95 <individual review sentences>
Negative: 10 <individual review sentences>
Aspect: **Battery life**
Positive: 50 <individual review sentences>
Negative: 9 <individual review sentences>
...

2.2 irudia: Aspektuan oinarritutako iritzi-testuen laburpenaren emaitza egituratua (Cambria *et al.*, 2017).

2.2 irudian, iritzi-testu baten laburpenaren emaitza egituratua ikus daiteke. Bertan argazki-kamera bati buruzko iritzia ematen da eta beraren zenbait aspektu azaltzen dira: orokorra, argazkien kalitatea eta bateriaren bizitza. Aspektu horietako bakoitzaren eskuinetara, zenbat iritzi-testuk positiboki edo negatiboki ebaluatu duten zehazten da. Adibidez, argazkiaren kalitatea 95 iritzi-testuk positiboki baloratu dute eta 10 iritzi-testuk negatiboki.

Aspektuan oinarritutako iritzi-laburpena dagoen moduan, badago dokumen-

tu askotako testu-laburpena ere. Carenini *et al.* (2006) eta Hu *et al.* (2017) eremu horretako lanak dira. Hala ere, iritzi-testuen laburpena ezberdina da dokumentu baten edo askoren laburpenetik. Izan ere, iritzi-testuen laburpena aspektu eta entitateetan eta aspektu eta entitate horietako bakoitzaren orientazio semantikoan oinarritzen da. Gainera, alderdi kuantitatiboa ere badago, hau da, horietako bakoitza positiboa edo negatiboa den zehaztu behar da. Testuen laburpen tradizionalean, aldiz, esaldirik garrantzitsuenak erauzten dira eta dokumentu askotako testu-laburpenean, berriz, errepikatzen ez den informazioa bilatu eta errepikatzen den informazioa ezabatzen da.

Aspektuan oinarritutako iritzi-laburpenaren lanik aipagarrienak Hu eta Liurenak (2004) eta Liu *et al.*renak (2005) dira. Dokumentu askotako iritzi-laburpenean, aldiz, oinarrizko lana Das eta Martinsena (2007) da.

2.1.4.2. Konparaziozko iritzien analisia

Entitate bati edo bere aspektuei buruzko iritzi positiboa edo negatiboa zuzenean adierazteko aukeraz gain, antzeko bi entitateen arteko konparazioa eginez ere iritzia adieraz daiteke. Horrelako iritziei konparaziozko iritziak deritze.

Konparaziozko esaldiek beti ez dute zuzenean iritzia adierazten. Hau da, objektu bati buruzko iritzia zuzenean eman beharrean, beste objektu batetiko konparazioa da iritzia. Gainera, konparaziozko iritzien barruan, superlatibozko iritziak ere topa ditzakegu.

2.8 taulan, konparaziozko iritzietan aipatuko ditugun lanak laburbilduta daude.

Teknika	**Lana**
Gako-hitzak + SVM	Jindal eta Liu (2006)
Konparaziozko esaldiak identifikatzeko erregelak	Liu *et al.* (2010)
Lexikoian oinarritutako hurbilpena + aspektuan oinarritutako sentimenduen sailkapena	Ding *et al.* (2009)

2.8 taula: Konparaziozko iritziak lantzeko zenbait teknika.

Konparaziozko iritziak identifikatzeko hainbat teknika ezberdin erabili izan

dira. Jindal eta Liuk (2006) erakusten dutenez, konparaziozko iritziek gako-hitzak eduki ohi dituzte, mota horretako hitzak antzematen laguntzen dutenak. Adibidez, konparaziozko adjektiboak (*gehiago*, *gutxiago*) eta superlatibozko adjektiboak (*gehien*, *gutxien*) gako-hitzak dira.

Bestalde, Liu *et al.*ek (2010) konparaziozko galderak eta konparatzen diren entitateak ikertzen dituzte. Konparaziozko esaldiak identifikatzeko erregelak erabiltzen dituzte.

Azkenik, Ding *et al.*ek (2009) konparaziozko iritzi batean entitate gustukoena identifikatzea dute helburu eta bi urrats burutzen dituzte: i) entitatea ezagutzea eta ii) entitatea esleitzea.

2.1.4.3. Iritzi-bilaketa eta -berreskuraketa

Iritzi-bilaketa web-bilaketaren atazaz eta sentimenduen analisiko atazaz osatuta dago:

1- Web-bilaketako ataza: i) bilaketarekiko garrantzitsuak diren dokumentuak eta esaldiak bilatu behar dira eta, ondoren, ii) esaldi edo dokumentuak garrantziaren arabera sailkatu behar dira.

2- Sentimenduen analisiko ataza: i) bilatutako dokumentu edo esaldiak bilaketa-gaiari (entitatea edo aspektua) buruzko iritzia baduen zehaztu behar da eta, hala bada, ii) iritzia positiboa edo negatiboa den zehaztu behar da.

Iritzi-bilaketan aipatu ditugun teknika ezberdinak 2.9 taulan esplikatuta daude.

Iritzi-bilaketa egiteko zenbait teknika proposatu izan dira. Zhang *et al.*ek (2007) lehenik eta behin, bilaketarekiko garrantzitsuak diren dokumentuak bilatzen dituzte. Horretarako, kontzeptuen desanbiguazioa, kontzeptu horien sinonimoen bilaketa eta dokumentuen antzekotasuna neurtzen dituzte eta, ondoren, SVM erabiltzen dute sentimenduen sailkapena egiteko.

Eguchi eta Lavrenkok (2006), berriz, sorkuntzazko hizkuntza modelatzen[18]

[18]Modelatu esatean, hizkuntzaren eredua sortu nahi dela esan nahi da.

Teknika	Lana
Desanbiguazioa, sinonimoen bilaketa, dokumentuen antzekotasuna + SVM	Zhang *et al.* (2007)
Sorkuntzazko hizkuntza-eredua	Eguchi eta Lavrenko (2006)
Lexikoian oinarritutako iritzi-bilaketa	Na *et al.* (2009)
Lexikoen eta sentimenduen ezaugarriak + algoritmo ezberdinak	Liu *et al.* (2009)

2.9 taula: Iritzi-bilaketa egiteko aipatutako teknikak.

dute iritzi-bilaketa egiteko. Horretarako, bilaketako hitz multzo adierazgarri bat, intereseko sentimendu-balentziadun hitzak erabiltzen dituzte.

Na *et al.*ek (2009), ordea, ikuspegi ezberdina lantzen dute; izan ere, lexikoian oinarrituta egiten baitute iritzi-bilaketa. Domeinuari loturiko lexikoia sortzen dute horretarako, eta lexikoiaren oinarrian *feedback* bidezko ikasketa dago. Ikasketa egiteko, bilaketa bateko dokumentuak erabiltzen dituzte.

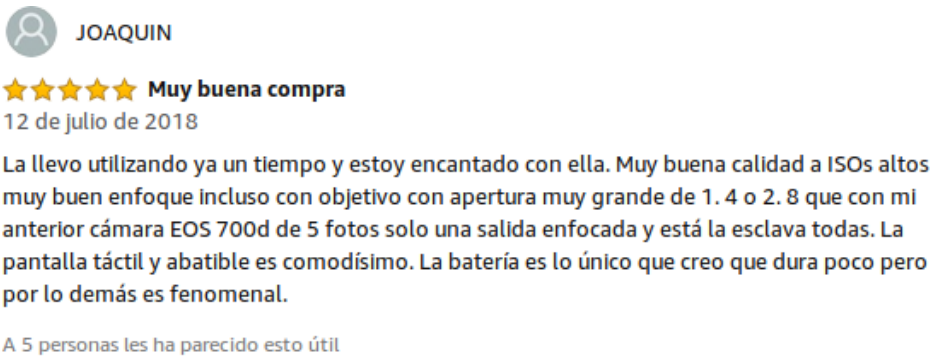
Azkenik, Liu *et al.*ek (2009) iritziak dauden blog batetik algoritmo ezberdinak eta lexikoen eta sentimenduen ezaugarriak ikasten dituzte. Modu horretan, sentimenduen analisia eta osagaien berreskuraketa uztartzen dituen estrategia garatzen dute.

2.1.4.4. *Spam*ak diren iritzien detekzioa

Norbanakoek eta erakundeek sare sozialetako iritziak gero eta gehiago hartzen dituzte kontuan. Iritzi horiek produktu baten erosketan, hauteskundeetan, marketinean edota produktu baten diseinuan eragin dezakete. Testuinguru honetan, iritzi positiboak arrakasta eta irabaziekin lotzen dira eta horregatik, norbanakoek gezurrezko iritziak eta iruzkinak idazten dituzte; produktu, zerbitzu, pertsona edo ideiei ospea emateko edo ospea kentzeko. Horrelako iritzi-motak antzemateko *spam*ak diren iritzien detekzioa garatu da. Li *et al.* (2011), Lim *et al.* (2010) eta Wang *et al.* (2012) dira lanik aipagarrienak iritzi *spam*ak diren iritzien detekzioan.

2.1.4.5. Iritzien kalitatearen neurketa

Sentimenduen analisiaren beste ataza bat iritzien kalitatearen neurketa da. Ataza honek iritzi-kritika baten kalitatea, lagungarritasuna, erabilgarritasuna edo baliagarritasuna neurtu nahi du. Lehenik eta behin, iritzi-testuak erauzi egiten dira eta, ondoren, iritzi-testuak sailkatu. Sailkapena egiteko irizpideak bere kalitatea eta bere erabilgarritasuna dira.


JOAQUIN

★★★★★ **Muy buena compra**

12 de julio de 2018

La llevo utilizando ya un tiempo y estoy encantado con ella. Muy buena calidad a ISOs altos muy buen enfoque incluso con objetivo con apertura muy grande de 1. 4 o 2. 8 que con mi anterior cámara EOS 700d de 5 fotos solo una salida enfocada y está la esclava todas. La pantalla táctil y abatible es comodísimo. La batería es lo único que creo que dura poco pero por lo demás es fenomenal.

A 5 personas les ha parecido esto útil

2.3 irudia: Iritziaren kalitatearen neurketa Amazon webgunean.

Gaur egun, helburu horrekin, zenbait webgunek (Amazon barne) kritikaren kalitatea edo erabilgarritasunari buruzko kalifikazioa lortzen dute irakurleei horien inguruan galdetuz. 2.3 irudian esaterako, kritika hori bost pertsonei baliagarria suertatu zaie. Hala ere, iritziaren kalitatea neurtzeak luze jo dezake eta tresna automatikoen beharra dago.

Iritzien kalitatea neurtzeko aipatuko ditugun teknikak 2.10 taulan laburbilduta daude.

Teknika	Lana
SVM erregresioa	Kim *et al.* (2006)
Ospe ezaugarriak + eduki ezaugarriak + ezaugarri sozialak + sentimendu ezaugarriak	Lu *et al.* (2010)
Hurbilpen gainbegiratu gabea	Tsur eta Rappoport (2009)

2.10 taula: Iritzien kalitatea neurtzeko zenbait teknika.

Kim *et al.*ek (2006) iritzien kalitatearen neurketa SVM erregresioa baliatuz egin dute. Horretarako, ezaugarri hauek erabili dituzte: egitura-ezaugarriak

(kritikaren luzera, esaldi-kopurua, galdera- eta harridura-ikurrak, etab.), lexiko ezaugarriak (unigramak eta bigramak), ezaugarri sintaktikoak (hitzen gramatika-kategoria eta horien agerpen-kopurua), ezaugarri semantikoak (produktuaren aspektuak eta sentimendudun hitzak) eta meta-datu ezaugarriak (kritikak dituen izar kopurua).

Lu *et al.*ek (2010), ordea, sailkatze hurbilpenean, kritika lagungarriak eta ez-lagungarriak sailkatzen dituzte. Erabiltzen dituzten ezaugarriak ospe-ezaugarriak, eduki-ezaugarriak, ezaugarri sozialak eta sentimendu-ezaugarriak dira.

Azkenik, Tsur eta Rappoportek (2009) liburu-kritiken erabilgarritasuna hurbilpen gainbegiratu gabea erabiliz neurtzen dute. Lehenik, kritiketako hitzik garrantzitsuenak identifikatzen dituzte eta horiekin bektore bat osatzen dute. Ondoren, kritika bakoitza bektore bilakatzen dute aurretik zehaztutako hitzik garrantzitsuen agerpenean oinarrituz. Azkenik, kritikak sailkatu egiten dituzte.

2.2 Sentimenduen analisirako baliabideen sorkuntza

Atal honetan, sentimenduen analisiko ataza ezberdinetan erabili ohi diren bi baliabideen inguruan jardungo dugu. Alde batetik, iritzi-testuen corpusen inguruan arituko gara (2.2.1 atala). Hauek aniztasun handia erakusten dute. Batzuk elebakarrak dira, besteak eleanitzak. Badaude anotatu gabekoak eta anotatutakoak ere. Eta batzuek testu-genero asko biltzen dituzten bitartean, besteek testu-genero bakarra lantzen dute. Bestetik, berriz, sentimenduen lexikoien inguruan arituko gara (2.2.2 atala). Baliabide hau sortzerakoan, bi hurbilpen nagusi daude: hiztegian oinarritutakoa eta corpusean oinarritutakoa.

2.2.1. Iritzi-testuen corpusak

Hizkuntza naturalaren prozesamenduko sentimenduen analisi-atazarako mota askotako eta helburu ezberdinetarako corpusak sortu izan dira. Leturia *et al.*ek (2014) corpus-mota hauek bereizten ditu.

- **Eremuaren arabera**
  - Orokorrak. Denetariko gaiez osatutakoak dira. Hizkuntzaren erabilera-eremu guztietarako baliagarriak edo "adierazgarriak" izatea dute helburu.
  - Espezializatuak. Hizkuntzaren erabilera-eremu espezializatu bateko testuak biltzen dituzte. Terminologiarako egokiak dira.
- **Erregistroaren arabera**
  - Denetarikoak. Genero edo erregistro guztiak biltzen dituzten corpusak dira.
  - Erregistro jakin batekoak. Testu guztiak genero edo erregistro jakin batekoak dira. Esaterako, kazetaritza eta administrazioko testuez osatutako corpusak badaude.
- **Adierazgarritasunaren arabera**

  - Orekatuak. Unibertsotik ausaz hartutako testuez osatutako corpusak. Unibertso horren ezaugarriak eta banaketa badituzte. Corpusik objektiboenak dira.
  - Ereduzkoak. Corpusaren egileak eredugarritzat jotzen dituen testuez osatuta dago. Helburu preskriptiboa dute horrelako corpusek.
  - Oportunistak. Irizpide bakartzat testuak erraz biltzea duen corpusmota da.

- **Hizkuntza-kopuruaren arabera**
  - Elebakarrak. Hizkuntza bakarrez osatutako corpusak dira. Lexikografian erabiltzen dira, batez ere.
  - Eleaniztunak. Hizkuntza bat baino gehiagoz osatutako corpusak dira. Itzulpen automatiko estatistikoko sistemak entrenatzeko egokiak dira.

- **Corpus eleaniztun motak lerrokatze mailaren arabera**
  - Paraleloak. Hizkuntza bakoitzeko azpicorpuseko testuak elkarren itzulpenak diren corpusei esaten zaie.
  - Konparagarriak. Azpicorpusek ezaugarri amankomunen bat dutenean esaten da. Testuak ez dira elkarren itzulpenak eta ez daude lerrokatuta. Amankomunean dutena generoa edo garaia izan daiteke.

Irizpide horietaz gain, badago beste bat corpusak sailkatzeko erabil daitekeena. Sailkapen hori aberastasunaren araberakoa da, hots, ea corpusa etiketatuta dagoen edo ez.

Jarraian, sentimenduen analisirako sortu diren zenbait corpus deskribatu ditugu. Horretarako, aurretik aipaturiko ezaugarrietan oinarritu gara.

Sentimenduen analisirako sortu diren eta aipatu ditugun corpusak 2.11 taulan laburbilduta daude.

Corpusa	Tamaina	Eremua	Erregistroa	Adierazgarritasuna	Hizkuntzakopurua	Lerrokatzemaila	Aberastasuna
Taboada (2008)	300 testu	Orokorra (6 gai)	Bakarra (iritziak)	Orekatua	Ingelesa		Etiketatu gabea
Refaee eta Rieser (2014)	8.868 txio	-	Bakarra (txioak)	Oportunista	Arabiera		Hizkuntza maila ezberdinetan etiketatuta
Clematide *et al.* (2012)	270 esaldi	Orokorra	Denetarikoak	Orekatua	Alemana		Hizkuntza maila ezberdinetan etiketatuta + sentimendua
Shin *et al.* (2012)	8.050 esaldi		Bakarra (albisteak)		Koreera		MPQA hurbilpenaz etiketatuta
Li *et al.* (2012)	108 dokumentu 714 esaldi	Espezializatua (politika)	Bakarra (albisteak)		Alemana		Hizkuntza-fenomenoak + Sentimendua
Fernández *et al.* (2011)	300 testu 975 esaldi ~90.000 hitz	Orokorra	Bakarra (Web 2.0ko testuak)		Gaztelania Italiera Ingelesa	Konparagarria	Subjektibitate-adierazpenak (sentimendua)
Schulz *et al.* (2010)	750 iruzkin	Espezializatua	Bakarra (kritikak)	Orekatua	Ingelesa Gaztelania Alemana	Konparagarria	
Karoui *et al.* (2017)	38.262 txio		Bakarra (txioak)	Orekatua (ironiko eta ez-ironiko)	Frantsesa Ingelesa Italiera	Konparagarria	Sentimendua etiketatuta
Bond *et al.* (2016)	3.824 esaldi 74.732 hitz	Espezializatua	Bakarra (ipuinak)	Orekatua	Ingelesa Txinera Japoniera	Paraleloa	Kontzeptuak sentimenduarekin etiketatuta

2.11 taula: Sentimenduen analisirako sortu diren zenbait corpus.

Taboadaren (2008) *The SFU Review Corpus* izenekoa lanetako bat da. Corpus honek *Epinions* webguneko iritzi-testuak biltzen ditu, sei gai edo kategorien ingurukoak: liburuak, autoak, ordenagailuak, sukaldeko tresneria, hotelak, filmak, musika eta telefonoa. Kategoria bakoitzak 25 iritzi positibo eta 25 iritzi negatibo ditu. Beraz, guztira 300 iritzi-testuko corpusa da. Bere ezaugarriei dagokienez, corpus orokorra dela esan daiteke, baita erregistro bakarrekoa dela ere, eta adierazgarritasunari dagokionez, orekatua dela. Bestalde, corpusa ingelesez dago eta hainbat hizkuntza mailatan etiketatuta dago: diskurtsoa RST hurbilpena erabiliz, subjektibitatearen ebaluazioa *Appraisal Theory* erabiliz eta, azkenik, ezeztapena eta espekulazioa.

Bestalde, Refaee eta Rieserek (2014) Twitter erabiltzen dute eta arabierazko 8.868 txio biltzen dituzte. Arabieraren dialekto ezberdinak biltzen dituzte ausaz egindako bilaketa batean. Beraz, corpus oportunista da adierazgarritasunaren aldetik eta txioak bakarrik biltzen dituelako erregistro bakarrekoa da. Gainera, corpusa ezaugarri ezberdinez etiketatuta dago, hala nola, morfologia, sintaxia, semantika, estilistika[19] eta seinale sozialak[20].

Clematide *et al.*ek (2012), berriz, corpusa alemanerako sortzen du. Corpusa guztira 270 esaldiz osatuta dago eta *DeWaC Corpus* izeneko corpusean oinarritzen da. Corpus honek webguneetako hainbat generotako testuak biltzen ditu. Corpusa geruza ezberdinez osatuta dago eta horietako bakoitza etiketatuta. Lehena esaldi mailako geruza da eta esaldi bakoitzaren objektibitatea, subjektibitatea eta polaritatea zehazten da. Bigarren geruza hitz eta sintagma mailakoa da, eta informazio subjektiboa eta faktuala dago etiketatuta. Azken geruza adierazpen maila da. Hizketa-gertaerak objektiboak edo zuzenekoak diren zehazten du. Beste modu batera esanda, egoera pribatuko ingurunea zehazten da. Beraz, corpus hau eremuari dagokionez orokorra dela esan daiteke, eta erregistroari dagokionez orekatua eta etiketatua.

[19]Hizkuntzalaritzan, estilistika adierazkortasuna eta literaturan estiloa eragiten duten hizkuntza erabilerak aztertzen dituen arloa da. Erretorikazko figurak eta sintaxizko egiturak lantzen dituzte.

[20]Seinale sozialak gizarte gertakariei buruzko informazioa ematen duten seinale komunikatiboak edo informatiboak dira. Hainbat motariko informazioa eman dezakete: gizarte elkarrekintzak, gizarte-emozioak, gizarte-ebaluazioak, gizarte-jarrerak edota gizarte harremanak, besteak beste.

Li *et al.*ek (2012) albisteez osaturiko sentimenduen analisirako beste corpus bat aurkezten dute. Kasu honetan, alemanezko albiste politikoak bildu eta etiketatu dituzte. 108 dokumentuz eta 714 esaldiz osatuta dago corpusa. Guztira jarrerarekin lotura duten lau elementu anotatu dituzte. Horietako bat testu-aingura da eta jarrera adierazten duten testu-zatiak dira hitz, sintagma edota lokuzioak. Bigarren elementua helburua da, hots, jarreraren helburua edo jomuga. Anotatu beharreko hirugarren elementua jarreraren iturria da, hau da, iritzia adierazten duen pertsona. Azken elementua, berriz, laguntzailea deiturikoa da; sentimenduan eragin dezakeen hitza: ezeztapena, intentsifikatzailea edota ahultzailea.

Shin *et al.*ek (2012) koreerazko sentimenduen corpusa aurkezten dute. Corpusak albiste-artikuluak biltzen ditu eta guztira 8.050 esaldiek osatzen dute. Corpusa persperktiba anitzeko galde-erantzunen (*Multiperspective Question Answering*, MPQA) hurbilpenean oinarrituz etiketatuta dago. MPQAk esaldiaren orientazioa positiboa edo negatiboa den zehazteko aukera ematen du.

Orain arte, sentimenduen analisi-atazarako sortu diren corpus ezberdinak deskribatu ditugu. Ikusten denez, corpusek ezaugarri ezberdinak dituzte, bai bildutako testu edo esaldiei dagokienez, bai etiketatzeko moldeari dagokionez. Baina denek dute ezaugarri bat amankomunean eta hori corpusa hizkuntza batez osatuta dagoela da. Hala ere, badira zenbait corpus hizkuntza askoz osatuta daudenak, jarraian datozenak bezalakoak.

Fernández *et al.*ek (2011) aurrekoekiko zertxobait ezberdina den corpusa aurkezten dute. *EmotiBlog* izeneko corpusak Web 2.0 web-zerbitzuaren bidez sortutako testu-genero berriak biltzen ditu. Bertako adierazpen subjektiboak etiketatuta ditu corpusak. Corpusak 975 esaldi ditu. Horietatik 519 esaldi objektiboak dira eta gainerakoak, 456 esaldi subjektiboak. Hauen artean, 260 esaldi positiboak dira eta 188 esaldi, aldiz, negatiboak. Hiru gai lantzen dituzte testuek: Kioto protokoloa, Zimbabweko Mugaberen gobernua eta AEBetako 2008ko hauteskunde presidentzialak. Gai bakoitzeko 100 testu bildu dituzte. Bestalde, corpusa eleanitza da, gaztelaniazko, italierazko eta ingelesezko testuak baitaude. Hizkuntza bakoitzak 30.000 hitz inguru ditu.

Schulz *et al.*ek (2010) hizkuntza anitzeko corpus bat aurkezten dute. Corpus honek ingeleseko, gaztelaniako eta alemaneko testuak biltzen ditu eta

eskuz etiketatuta daude. Hu eta Liuen (2004) corpusean, ingelesez dagoena, oinarritzen dira corpusa sortzeko eta produktu berberen alemanezko eta gaztelaniazko beste 500 iruzkin gehitzen dituzte. Iritzien ezaugarri hauek etiketatu dira: i) produktuen ezaugarriak, ii) horien orientazio semantikoa eta iii) horien indarra (0-3 artekoa). Kasu honetan, corpus eleanitza lerrokatze mailari dagokionez konparagarria da; izan ere, azpicorpusek amankomunean dutena produktu berberen iritziak izatea da.

Karoui *et al.*ek (2017) ere hiru hizkuntzaz osaturiko corpusa sortu dute baina testu-mota ez da bera, txioz osaturiko corpus hirueleduna sortu baitute. Gainera, corpusa helburu batekin sortu dute eta helburua ironia detektatzeko anotazio-irizpide bat finkatzea da. Corpuseko hiru hizkuntzak frantsesa, ingelesa eta italiera dira. Frantseseko azpicorpusean, 2.073 txio ironiko eta 16.179 txio ez-ironiko daude. Ingeleseko azpicorpusean, 5.173 txio ironiko eta 6.116 txio ez-ironiko daude. Azkenik, italierakoan, 3.079 txio ironiko eta 5.642 txio ez-ironiko daude. Ikusten den moduan; beraz, corpusa ez da paraleloa, txioak ez direlako itzulpenak; baina konparagarria bada, ezaugarri amankomuna ironia edukitzea edo ez edukitzea delako. Geruza askotariko anotazio-irizpidea baliatuta etiketatu dute corpusa: ironia (esplizitua edo inplizitua), ironia-kategoriak eta ironia-markatzaileak.

Azkenik, Bond *et al.*ek (2016) ere beste corpus eleanitz bat osatu dute. Corpus hau ingelesezko bi istorio laburrez osatuta dago. Aurretik aipatutako beste bi corpusekin corpus honek duen ezberdintasuna horixe da; corpus eleanitza itzulpen bidez lortu denez, corpusa paraleloa dela. Anotazioa bi mailatan egin dute. Kontzeptu mailan, sentimendu positiboa edo negatiboa duten hitzak (kontzeptuak) etiketatu dituzte. *Chunk* mailan, berriz, sentimendu bera duten segidako hitz-multzoak etiketatu dituzte. Ingeleseko testuak originalak dira eta japonierara eta txinerara itzuli dituzte.

Gure iritzi-corpusak lotura gehien Shin *et al.*ekin (2012), Li *et al.*ekin (2012) eta Taboadarekin (2008) du. Taboadak (2008) bezala, gure corpusa orekatua da eta Shin *et al.* (2012) eta Li *et al.* (2012) lanek bezala, albisteak edo kazetaritzako testuak biltzen ditu gure corpusak.

2.2.2. Sentimenduen lexikoia

Sentimendua duten hitzei hainbat izen hartu izan dituzte: sentimendu-hitzak (*sentiment words*), iritzi-hitzak (*opinion words*) edota polaritate-hitzak (*polar words*), besteak beste. Liuk (2012) dioenez, sentimendu positiboko hitzek egoera edo kualitate desiratuak adierazten dituzte: *ederra*, *bikaina* edota *txundigarria*. Sentimendu negatiboko hitzek, aldiz, egoera edo kualitate ezdesiratuak adierazten dituzte: *txarra*, *beldurgarria* edota *pobrea*. Hitz horiek guztiek osatzen duten multzoari sentimenduen lexikoia esango diogu.

Sentimendu-hitzak bi motarikoak izan daitezke: oinarrizkoak (*base type*) eta konparaziozkoak (*comparative type*). Azken honetan, konparatiboak edo superlatiboak diren hitzak biltzen dira: *hobea*, *okerragoa*, *onena* eta *okerrena*, esaterako.

Liuren (2012) lanean sentimendu-hitzak biltzeko bi modu agertzen dira:

1- Hiztegian oinarritutako hurbilpena.

2- Corpusean oinarritutako hurbilpena.

Jarraian, bi hurbilpen horiek deskribatuko ditugu, tesi-lan honetan sortu dugun *Sentitegi*, euskarazko sentimenduen lexikoia, bi hurbilpen horietaz elikatu delako.

2.2.2.1. Hiztegian oinarritutako sentimenduen lexikoia

Sentimenduen lexikoia sortzeko hiztegia erabiltzea (Millerren (1995) *Wordnet* izeneko datu-base lexikala kasu) nahiko hurbilpen ohikoa izan da; izan ere, hiztegiek sinonimoak eta antonimoak biltzen dituzte eta horiek baliagarriak dira lexikoia sortzeko. Beraz, teknika sinpleena honako hau da: sentimendua duten hitz multzo txiki bat hartu eta, hiztegiaren sinonimo eta antonimo egituran oinarrituta hiztegiko egituran barrena mugitzea *boostrapping* teknika[21] erabiliz sentimendua duten hitz gehiago bilatzeko.

[21]Estatistikan, *boostrapping* berlanginketa egiteko metodo bat da eta laginen banaketa egiteko erabiltzen da. Laginen datu multzo bat baliatuz, populazio baten gainean inferentzia egitean; ondoren, inferentzia hori laginen datuen berlaginketa baten bidez modelatzean eta, azkenik, berlangiketako datuekin inferentzia berri bat egitean datza.

Wordnet erabilita lexikoi bat Hu eta Liuren (2004) arabera modu honetan sor daiteke:

1- Lehenik eta behin, eskuz, hazi-hitz positibo edota negatiboak bildu behar dira. Hazi-hitzak datu-base lexikaletik hartutako hasierako zenbait hitz dira. Ondoren, hitz horietan oinarrituta bilaketa gehiago egiten dira datu-basean hitz gehiago aurkitzeko. Beigman Klebanov *et al.*ek (2012), esaterako, *anxiety*, *conflict* eta *improve* hitzak hazi-hitz moduan hartzen dituzte.

2- Ondoren, algoritmo batek hitz horien sinonimo eta antonimoak bilatzen ditu bai *Wordnet*en, bai sareko beste hiztegietan. Aurkitutako hitzak zerrendara gehituko dira. Beigman Klebanov *et al.*en (2012) lanean, algoritmoaren emaitzak dira bilatutako *anxiousness* (<*anxiety*), *battle* (<*conflict*) eta *amend* (<*improve*) hitzak.

3- Azkenik, iterazioa behin eta berriz hasten da hitza bilatzen. Ezin direnean hitz gehiago bilatu, iterazioa amaitzen da.

Hiztegian oinarrituta lexikoiak sortzea erraza eta azkarra da. Modu honetan sortutako lexikoiek akats batzuk eduki ditzakete, baina gero, eskuz konpon litezke. Modu honen desabantaila da sortutako lexikoiak oso loturik daudela domeinuari nahiz testuinguruari. Ondorioz, ez dago domeinu askotarako balio duen sentimenduen lexikoirik.

Lan hauek guk sortu dugun sentimendu lexikoiarekin bi ikuspegitatik dute lotura. Alde batetik, denak dira hiztegian oinarrituta sortuak. Beste aldetik, berriz, hitzen sentimenduen balentzia kalkulatu da hitzek elkarren artean duten lotura semantikoa aintzat hartuta.

Sentimenduen lexikoia sortzeko hiztegia erabiltzen duten eta aipatuko ditugun lanak 2.12 taulan laburbilduta daude.

Blair-Goldensohn *et al.*ek (2008) *bootstrapping* teknika konplexuagoa erabiltzen dute. Izan ere, lehenik eta behin hazi-hitz positibo, negatibo eta neutralak biltzen ditu. Ondoren, hitz horiei pisua esleitzen diete grafo semantikoa

Teknika	Lana
Wordnet + Bootstrapping	Hu eta Liu (2004) Blair-Goldensohn *et al.* (2008)
Grafoan oinarritutako ikasketa gainbegiratu gabea	Rao eta Ravichandran (2009)
Markoven ausazko ibilaldia (*Markov random walk*)	Hassan *et al.* (2014)
Elkarrekiko informazio puntuala (*Pointwise Mutual Information*, PMI)	Turney eta Littman (2003) Taboada *et al.* (2011)

2.12 taula: Hiztegian oinarrituz sentimendu lexikoia sortzeko zenbait teknika.

erabiliz. Grafo[22] semantiko horretan, ondoko adabegiak (*neighboring nodes*) *Wordnet*eko hitzen sinonimoak eta antonimoak dira eta ez daude hazi-hitz neutralen multzoan. Hazi-hitz neutralak sentimendu orientazioa hitz neutraletara ez zabaltzeko erabiltzen dira.

Rao eta Ravichandran (2009) lanak, berriz, hitzen polaritatea antzematea du helburu eta, horretarako, grafoan oinarritutako ikasketa erdi-gainbegiratua proposatzen dute. Beren ekarpena da sinonimia eta hiperonimia moduko erlazioak erabiltzea etiketen zabalkundearen emaitzak hobetzeko. Lanaren oinarrian Wordnet datu-base lexikala dago.

Hassan *et al.*en (2014) lanak, hitzen orientazio semantikoa kalkulatzeko asmoz, Markoven ausazko ibilaldia (*Markov random walk*) eredua erabiltzen du. Beren metodoak *Wordnet* datu-basean hedapena orientazio semantiko bera duten hitzen bidez egiten du lehenik eta, ondoren, hedapena orientazio semantiko ezberdineko hitzetara igarotzen da.

Bestalde, Turney eta Littmanen (2003) lanak elkarrekiko informazio puntuala (*Pointwise Mutual Information*, PMI) neurrian oinarritutako metodoa erabiltzen du. Hitz baten orientazio semantikoa kalkulatzeko, hitz batek hazi-hitz positiboekiko (*ona*, *bikaina*, *zuzena*) duen asoziazioaren indarra ken hazi-hitz negatiboekiko (*txarra*, *pobrea*, *okerra*) duen asoziazio semantikoa-

[22]Grafoa datu-egitura mota bat eta, bertan, erlazioak edo loturak adierazten dira.

ren indarra[23] kalkulatuz zenbatzen du. Elkarrekiko informazioa puntuala erabiliz neurtzen du hitz batek duen asoziazioaren indarra. Taboada *et al.*en (2011) SO-CAL tresnako lexikoia ere modu berean sortu da.

2.2.2.2. Corpusean oinarritutako sentimenduen lexikoia

Liuk (2012) azaltzen duenez, corpusa erabiliz sortutako sentimenduen lexikoiak bi arrazoi jakinengatik sortzen dira.

1- Ezagunak diren sentimenduen hitzak edukita, domeinu bakarreko corpusean sentimendua duten hitzak eta horien orientazio semantikoa ezagutzeko. Hau da, aldez aurretik lortutako sentimendu-hitzak domeinu bakar bateko corpusean aplikatzen dira eta corpusetik sentimendu-hitzak lortzen dira.

2- Helburu orokor baterako edo domeinu askotarako sortutako sentimenduen lexikoia beste domeinu berri batera moldatzeko, domeinu bakarreko corpusean oinarrituz. Kasu honetan, aurretik badagoen lexikoi bat domeinu batera murrizten da corpus baliatuz.

Bi arrazoiek antzekoak diruditen arren, badaude ezberdintasun batzuk. Lehen kasuan, sentimendu-hitzak corpusetik lortu dira. Bigarrengoan, aldiz, aldez aurretik daude sentimendu-hitzak, baina corpusaren iragazkia pasatu dute.

Corpusean oinarritutakoetan helburua lexikoia domeinu batera mugatzea bakarrik dela dirudien arren, hori egitea ez da uste bezain erraza. Izan ere, domeinuaren arabera, hitz batek bi aurkako esanahi izan ditzake eta hori tratatzea zaila da.

Corpusean oinarritutako lexikoiak erabili izan dira helburu orokorreko sentimenduen lexikoiak sortzeko, beti ere corpusa oso handia eta anitza bada. Hala ere, helburu horretarako hiztegian oinarritutako lexikoiak egokiagoak dira.

[23] Asoziazioa estatistikan erabiltzen den kontzeptua da eta bi aldagaien arteko erlazioa edo lotura adierazten du. Sentimenduen analisira mugatuz, bi hitzek elkarren artean duten lotura semantikoa adierazten du.

Corpusean oinarrituta sentimenduen lexikoiak sortzen dituzten lanak 2.13 taulan laburbilduta daude.

Teknika	**Lana**
Esaldi barnea: juntagailuak	Hatzivassiloglou eta McKeown (1997)
Esaldi barnea: juntagailuak Esaldi artean: aurkaritzako juntagailuak	Kanayama eta Nasukawa (2006)
Aspektua (iritziaren testuingurua)	Ding *et al.* (2008)
i) Elkarrekiko errefortzua ii) Domeinuen arteko zati amankomuna	Du *et al.* (2010)
Antzekotasun distribuzionala	Wiebe eta Mihalcea (2006)

2.13 taula: Corpusean oinarrituta sortutako zenbait sentimenduen lexikoi.

Corpusean oinarritutako hurbilpenak guk sortu dugun sentimenduen lexikoiarekin duen lotura da biak direla hein batean corpus batetik sortuak edo corpus batean oinarrituak. Gure kasuan, lexikoi bat gaztelaniatik euskaratu ostean, lexikoia corpuseko domeinuetara egokitu dugu. Hau da, 2.2.2.2 atalaren hasieran aipatu dugun moduan, helburu orokor baterako edo domeinu askotarako sortutako sentimenduen lexikoia beste domeinu berri batera moldatu dugu, domeinu bakarreko corpusean oinarrituz.

Positiboa	**Negatiboa**
adequate, central, clever, famous, intelligent, remarkable, reputed, sensitive, slender, thriving	contagious, drunken, ignorant, lanky, listless, primitive, strident, troublesome, unresolved, unsuspecting

2.14 taula: Hatzivassiloglou eta McKeownen (1997) laneko hazi-hitzak.

Hatzivassiloglou eta McKeownena (1997) da hurbilpen honetan lehenengoetariko lana. Bertan, corpusa eta hazi-hitzak (2.14 taula) erabili dira corpusean sentimendua duten adjektibo gehiago bilatzeko. Lehenik juntagailuz lotuak dauden adjektiboak lortzen dituzte zenbait erregela linguistiko erabiliz: *eta, edo, baina, edo... edo...* eta *ez... ez...* izan dira. Ondoren, adjektibo bakoitzaren orientazio semantikoa lortu eta orientazioaren arabe-

Positibo bezala sailkatua	Negatibo bezala sailkatua
bold, decisive, disturbing, generous, good, honest, important, large, mature, patient, peaceful, positive, proud, sound, stimulating, straightforward, strange, talented, vigorous, witty	ambiguous, cautious, cynical, evasive, harmful, hypocritical, inefficient, insecure, irrational, irresponsible, minor, outspoken, pleasant, reckless, risky, selfish, tedious, unsupported, vulnerable, wasteful

2.15 taula: Hatzivassiloglou eta McKeownen (1997) sentimenduen lexikoia zati bat.

ra multzokatzen dituzte. Azkenik, multzo bakoitzaren maiztasunak alderatu eta maiztasun handienekoak positibo moduan etiketatzen dira (2.15 taula).

(8) Koordinaziozkoa ("eta" esanahia gutxi gorabehera): *-te, -shi, -weni, -dakedenaku, - nominarazu.*

(9) Kausazkoa ("-(e)lako" esanahia gutxi gorabehera): *-tame, -kara, -node.*

(10) Aurkaritzakoa ("baina" esanahia gutxi gorabehera): *-ga, -kedo, -keredo, -monomo, -nodaga.*

(11) *shikashi* ("hala ere"), *demo* ("baina"), *sorenanomi* ("nahiz eta"), *tadashi* ("baldin eta"), *dakedo* ("baina"), *gyakuni* ("bestela"), *tohaie* ("arren"), *keredomo* ("hala ere"), *ippou* ("bestalde").

Lan horretatik abiatuz, Kanayama eta Nasukawaren (2006) lanak zenbait aurrerapen egin dituzte. Aurreko kontzeptua garatu dute. Esaldi barneko (*intra-sentential*) eta esaldi arteko (*inter-sentential*) kontzeptuak sortu dituzte. Lehena Hatzivassiloglou eta McKeown (1997) laneko kontzeptu bera da eta esaldi barnean kokatzen denez, erregela linguistikoak ere (*eta, edo...*) esaldi barnekoak dira ((8), (9) eta (10) adibideak). Esaldi artekoan, aldiz, jarraian dauden bi esaldiek ideia edo kontzeptu bera adierazten dute. Kasu honetan, bestelako erregela linguistikoak sortu behar izan dituzte: *baina, hala ere* ((11) adibidea). Esaldi artekoa izendatzeko sentimendu sendotasuna (*sentiment consistency*) kontzeptua erabili dute.

Ding *et al.*ek (2008) diotenez, goiko metodologia domenuari loturiko hitzak bilatzeko erabilgarria den arren, praktikan hori bakarrik ez da nahikoa. Izan ere, hitz batek aurkako bi orientazio semantiko izan ditzake bi testuinguru ezberdinetan. Esaterako, *bateriaren bizia luzea da* eta *fokuratzeak denbora luzea eskatzen du* esaldietan, *luzea* hitzak lehenengoan orientazio semantiko positiboa du eta bigarrengoan, berriz, negatiboa. Horregatik, domeinuari loturiko sentimenduen hitza eta haren balentzia ez direla nahikoa ohartzen dira eta horren aurrean, aspektua eta sentimendua duen hitzari garrantzi gehiago eman eta iritzi-testuingurua irudikatzeko bikoteak (aspektua, sentimendu_hitza egitura dutenak) proposatzen dituzte. Adibidez, goiko adibidearen bikotea (*bateriaren bizia*, *luzea*) izango litzateke.

Aurreko lanak ezagunak diren hitzetatik abiatuz domeinu bateko corpusean sentimendua duten hitz gehiago bilatu eta horiei orientazio semantikoa esleitzeko hurbilpen eta teknikak dira. Jarraian, berriz, jada sortuta dagoen sentimenduen lexikoi orokor bat domeinu zehatz batera moldatzeko teknikak eta hurbilpenak aipatuko ditugu.

Du *et al.* (2010) lanean domeinu bateko lexikoia beste domeinu batera egokitzen dute. Horretarako, domeinu barruko dokumentu etiketatutak, domeinu barruko sentimendu-hitzak eta domeinuz kanpoko dokumentuak erabiltzen ditu algoritmoak. Bi aspektu lantzen dituzte: dokumentu bat positiboa bada, bertan dauden hitz askok ere orientazio semantiko positiboa dute eta dokumentua negatiboa bada, bertako hitz askok orientazio semantiko negatiboa dute (honek elkarrekiko errefortzuarekin, *mutual reinforcement*ekin, du lotura) eta nahiz eta bi domeinuek distribuzio ezberdina izan, beren arteko zati amankomuna identifikatzea posible da.

Wiebe eta Mihalceak (2006) corpusean oinarrituta hitz adierei subjektibitate etiketak esleitzea proposatzen dute. Bi azterketa egiten dituzte. Lehenengoan, *Wordnet*eko adierei bi pertsonek "subjektiboa", "objektiboa", "biak" etiketak esleitzeko duten adostasuna neurtzen dute. Bigarrengoan, hitzen adierei etiketak automatikoki esleitzen dizkie antzekotasun distribuzionala metodoan[24] oinarrituz.

[24]Antzekotasun distribuzionala (ingelesez, *distributional similarity* teknika bat da. Teknika honen arabera, antzekoak diren objektu linguistikoek antzeko edukia (dokumentuetan

Gure sentimenduen lexikoia aurreko bi hurbilpenen uztarketa da. Alde batetik, hiztegian oinarritutakoa da, SO-CAL tresnaren gaztelaniazko lexikoian (Brooke *et al.*, 2009) oinarrituz egin baita. Beste aldetik, berriz, corpusean oinarritutakoa ere bada, itzulitako lexikoia corpuseko sei domeinuetara egokitu baitugu eta anbiguoak diren hitzetan, corpuseko adiera eta, horrekin batera, sentimenduen balentzia hartu baititugu aintzat.

eta esaldietan, esaterako) eta antzeko testuingurua (hitzen kasuan) duten.

2.3 Balentzia-aldatzaileak

Behin lexikoiak nola lortzen diren aipatuta, testuinguruko balentzia-aldatzaileek horien balioetan nola eragiten duten azalduko dugu. Testuinguruzko balentzia-aldatzaileen lanketan erreferentziazko lana Polanyi eta Zaenenena (2006) da. Lan horretan, aurretik egin diren zenbait lan osatu gabe daudela adierazten dute eta beren analisia proposatzen dute, jarraian azalduko dugun moduan.

Polanyi eta Zaenenek (2006) testuinguruko balentzia-aldatzaileak (ingelesez, *contextual valence shifters*) proposatzen dituzte. Beren arabera, testuak gertaerak eta egitateak deskribatzeaz gain, deskribatzen den gertaerarekiko idazleak edo parte-hartzaileek duten jarrera komunikatzen du. Horretarako, idazleak aukeraketa lexikala egiten du, baina testuaren antolakuntzak berak ere idazlearen jarrera adierazteko informazio kritikoa ematen du. Beste modu batera esanda, nahiz eta idazleek aukeratutako hitzen sentimendu balentzia jakin bat izan, testuaren antolakuntzak eta bertako elementuek hitzen sentimendu balentzia alda dezakete. Atal honetan aipatuko diren lanak 2.16 taulan laburbilduta daude.

2.3.1. Testuingurua kontuan hartzen ez duten lanak

Atal honetan, hitzen testuingurua kontuan hartzen ez dutelako osatugabetzat jo dituzten zenbait lan aipatu eta beren analisia deskribatuko dugu.

Edmonds eta Hirstek (2002) sinonimo hurkoak eta aukeraketa lexikala erabiltzen dituzte. Zehatzago esanda, sinonimo hurkoen esanahiak eta beren arteko ezberdintasunak errepresentatzeko modeloa aurkezten dute, baita egoera jakin batean sinonimo hurko horien artean aukeraketa lexikal egokiena egiteko prozesua ere. Sinonimo hurkoak ia sinonimoak dira, baina ez guztiz. Esanahiaren aldetik oso antzekoak dira, baina ez dira berdinak eta ezin dira elkartrukatu. Denotazioan, konnotazioan, inplikazioan, enfasian edo erregistroan ezberdintzen dira. 2.17 taulan, zenbait sinonimo hurko ageri dira beren arteko ezberdintasuna zer den aipatuz.

Errepresentazioaren granularitatea erabiliz[25] sinonimo hurkoak multzokatu

[25]Errepresentazioaren granularitatean (ingelesez, *representation granularity*), hitz ba-

Hitzen orientazio semantikoa testuingurua kontuan hartu gabe	**Lana**
Sinonimo hurkoak Denotazioa, konnotazioa, inplikazioa, enfasia, erregistroa	(Edmonds eta Hirst, 2002)
Juntagailuz elkartuta dauden adjektiboak	(Hatzivassiloglou eta McKeown, 1997)
Web bilaketa motorraren kontsulta Elkarrekiko informazio puntuala	(Turney eta Littman, 2002)
Maiztasun gutxiko hitzak, kolokazioa, adjektiboak, aditzak	(Wiebe *et al.*, 2004)
Berrikuntza: **testuinguruko balentzia-aldatzaileak**	
Esaldi mailako testuinguruko balentzia-aldatzaileak Diskurtso mailako testuinguruko balentzia-aldatzaileak	(Polanyi eta Zaenen, 2006)
Hitzen orientazio semantikoa testuingurua kontuan hartuta	
General Inquirer lexikoia, web corpusa, Asoziazio kalifikazioa	(Kennedy eta Inkpen, 2006)
SVM Balentzia-aldatzaileen unigramak eta bigramak	(Kennedy eta Inkpen, 2006)
Ezaugarrien konbinaketa Testuinguruko balentzia-aldatzaileak	(Morsy eta Rafea, 2012)
Termino kontaketa Testuinguruko balentzia-aldatzaileak	(Ngoc Phu eta Thi Tuoi, 2014)

2.16 taula: Hitzen orientazio semantikoa esleitzeko testuingurua kontuan hartzen eta hartzen ez duten lanak.

Bariazio mota	**Adibidea**
Abstrakzio dimentsioa	*seep : drip*
Enfasia	*enemy : foe*
Denotaziozkoa, zeharkakoa	*error : mistake*
Denotaziozkoa, lausoa	*woods : forest*
Estilistikoa, formaltasuna	*pissed : drunk : inebriated*
Estilistikoa, indarra	*ruin : annihilate*
Jarrera adierazpena	*skinny : thin : slim, slender*
Emotiboa	*daddy : dad : father*
Kolokaziozkoa	*task : job*
Aukeraketazkoa	*pass away : die*
Azpikategorizazioa	*give : donate*

2.17 taula: Sinonimo hurkoen sailkapena beren ezberdintasunetan oinarrituta (Edmonds eta Hirst, 2002, 5. orr.).

tean esanahia sortzeko; alde batetik, testuinguruz independienteak diren esanahiek dituz-

egiten dituzte. Ondoren, ezagutza lexikalaren eredu multzokatua garatzen dute ohiko eredu ontologikoetan oinarrituz. Multzo bakoitzean azpikontzeptuak daude eta sinonimo hurkoak adierazten dituzten estiloaren, aldeko edo aurkako jarreraren, inplikazioaren eta denotazioaren arabera daude multzokatuta. Prozesua bi mailetan aritzen da aukeraketa lexikala egiteko: multzoetan eta sinonimo hurkoen multzoetan.

Hatzivassiloglou eta McKeownek (1997) adjektiboen orientazio semantikoa iragartzeko lana aurkezten dute. Horretarako juntagailuz elkartuta dauden adjektiboak baliatzen dituzte, (12), (13) eta (14) adibideetan bezalakoak.

(12) The tax proposal was {simple and well-received} by the public.

(13) The tax proposal was {simplistic but well-received} by the public.

(14) The tax proposal was {*simplistic and well-received} by the public.[26]

Erregresio linealean, (14) adibideko murriztapenak bezalakoak baliatuz (*eta* juntagailuarekin batera ezin direla orientazio semantiko negatiboko eta positiboko bi hitz agertu), adjektiboen orientazioa semantikoa iragartzea lortzen dute.

Turney eta Littmanek (2002) adjektiboen, adberbioen, izenen eta aditzen orientazio semantikoa iragartzen dute baina, horretarako, beste hurbilpen bat (metodo gainbegiratua) darabilte. Oinarritzat ehun bilioi hitzeko corpus bat, web-bilaketa motorraren kontsulta eta elkarrekiko informazio puntuala darabiltzate.

Wiebe *et al.*ek (2004) hizkuntza-subjektiboa lantzen dute, hots, iritziak, ebaluazioak edo usteak adierazteko hizkuntzaren aspektuak. Corpus batzuetatik subjektibitatearen arrastoak sortzen eta probatzen dituzte. Arrasto horiek maiztasun gutxiko hitzak, kolokazioak eta antzekotasun distribuzionala baliatuz lortutako adjektiboak eta aditzak dira. Halaber, lan horrek aipaturiko arrastoak dentsitate handi batez edo askotan hitz baten ondoan agertzeak, hitz hori subjektibo bihurtzen duela ere adierazten dute.

ten testuinguruz dependienteak diren konbinazioak eta, beste aldetik, sinonimo hurkoen ezberdintasun esplizituen multzo bat baliatzen dira.

[26]Izartxoak (*) esaldia ez dela zuzena esan nahi du.

Laburbilduz, aipaturiko lanetan, esaldien edo testuen orientazio semantikoa edo sentimendu-balentzia kalkulatzeko, hitz jakin batzuen sentimendu-balentzia erabiltzen dute. Baina, hitz horien inguruan dauden eta hitzaren sentimendu-balentzia aldatu diezaioketen hizkuntza-fenomenoak ez dira kontuan hartzen eta, horregatik, lan osatugabetzat jo izan dira.

2.3.2. Berrikuntza: testuinguruko balentzia-aldatzaileak

Polanyi eta Zaenenen (2006) arabera, aurretik aipaturiko lanek hutsune bat dute eta hutsune hori orientazio semantikoa duten hitzek ondoan duten testuingurua kontuan ez hartzea da. Izan ere, Edmonds eta Hirsten (2002) kasuan, sinonimoen arteko ezberdintasuna hitz horien zenbait ezaugarri zehatzetan oinarrituz egiten da. Hitzetik kanpo dauden ezaugarriak, hots, hitzaren testuinguruari loturikoak, ez dira aintzat hartzen. Hatzivassiloglou eta McKeownen (1997) kasuan ere, corpusetik juntagailuz loturik dauden adjektiboak ateratzen dituzte, baina ez da haien testuingurua kontuan hartzen. Turney eta Littmanen (2002) eta Wiebe *et al.*en (2004) lanekin ere berdin gertatzen da, testuinguruarekin loturarik ez duten ezaugarriak bakarrik erabiltzen dira hitzaren orientazio semantikoa zehazteko.

Gure lanari dagokionez, lexikoian oinarritzen den dokumentu mailako sentimenduen sailkatzailea garatzeko bidean, jakitun gara lexikoiko sarrerek duten sentimenduen balentzia bere horretan aintzat hartuta ez dela nahikoa. Izan ere, Polanyi eta Zaenenek (2006) dioten moduan, hitz baten sentimenduen balentzia jakiteko beharrezkoa da bere testuingurua ere kontuan hartzea. Hori horrela izanik, gure lanaren eta beren lanaren arteko lotura da bietan esaldi eta diskurtso mailako balentzia-aldatzaileak identifikatzea dutela helburu. Hori dela eta, gure lanaren aurrekaria da jarraian deskribatuko dugun hurbilpena.

Osagai lexikalek jarrera positiboa edo negatiboa adierazten dute. Edozein klase irekik adieraz dezake jarrera positiboa edo negatiboa: izenak (*onarpen/katastrofe*), adjektiboak (*erraz/zail*), aditzak (*arindu/huts egin*) eta adberbioak (*azkar/mantso*) daude eta hitz anitzeko unitate lexikalak (*porrot egin*) ere izan daitezke. Horren erakusle dira (15), (16), (17) eta (18) adibideak.

(15) 25 urteko bat bizi zen hiriko partean oinez ibili zen[27].

(16) [Gizon gazte bat]$_{+}$ [bere auzoan zehar]]$_{+}$ [patxadaz paseatu]]$_{+}$ zuen.

(17) [Gazte ar bat]$_{-}$ [bere lurraldean zehar]$_{-}$ [harrokeriaz ibili]$_{-}$ zen.

(18) Filma [interesgarria]$_{+}$ eta [hunkigarria]$_{+}$ izan arren, ez zitzaidan [gustatu]$_{-}$.

(15), (16) eta (17) adibideek sentimendu-hitzen garrantzia erakusten dute. Hirurek gertaera bera azaltzen dute baina egin den aukeraketa lexikalak[28], hau da, aukeratutako sentimenduen balentziek ikuspegi guztiz kontrajarriak adierazten dituzte. Izan ere, (15) adibideak jarrera neutrala du, (16) adibideak positiboa eta (17) adibideak, aldiz, jarrera negatiboa. Gainera, jarrera jakin bateko hitz asko egoteak ez du esan nahi esaldiak eta testuak jarrera hori adierazten dutenik. Esaterako, (18) adibidean, jarrera positiboko hitz gehiago daude, baina esaldiak jarrera negatiboa du. Halaber, kontuan hartu behar da sentimenduen hitzek duten intentsitatea; izan ere, *ona* eta *bikain* ez baitira semantikoki berdinak. Baina Polanyi eta Zaenenek (2006) diotenez, sentimendu-hitzen ezaugarri horiek bakarrik kontuan hartzea ez da nahikoa, testuinguruko balentzia-aldatzaileek ere hiztunek bere jarrera adierazteko momentuan zerbait edo bere aurkako zerbait esatea eragin dezaketelako.

Esaldi mailan hainbat faktore (ezeztapena, aditzen tempusa, etab.) aipatzen dituzte sentimenduen balentzia alda dezaketenak.

(19) Luar [azkarra]$_{+2}$ da. Luar [ez da azkarra]$_{-2}$.

(20) Gertaera [susmagarria]$_{+2}$ da. Gertaera [oso susmagarria]$_{+3}$ da.

(21) Soluzioak [eraginkorra]$_{+2} \rightarrow_{0}$ izan beharko luke.

(22) [Ana pertsona txarra da]$_{-1}$. [Bere lagunak gaizki tratatzen ditu]$_{-1}$. [Ana pertsona txarra izango balitz, bere lagunak gaizki tratatuko lituzke]$_{0}$.

[27](15) adibide hau nahiz beste hauek (16, 17, 20, 24, 25, 26, 28 eta 30) Polanyi eta Zaenenetik (2006) hartuta eta euskarara itzulita daude.

[28]Psikolinguistikan, aukeraketa lexikala (ingelesez, *lexical selection*) hizkuntza ekoizterakoan, mezua identifikatu ondoren, erantzuteko mezua irudikatzeko hiztunak aukeratzen dituen item lexikalak dira eta, prozesu horretan, hitzari buruzko informazio semantikoa eta gramatikala aktibatzen da.

(23) [Azterketa gainditu du.]$_{+}$ [Azterketa nekez gainditu du]$_{-}$.

(24) [Antolakuntza [oso [distiratsuak]$_{+2}$]$_{+3}$ $\rightarrow_{-3}$ [huts egin]$_{-1}$ zuen [arazoa konpontzen]$_{+1}$ $\rightarrow_{0}$]$_{+5}$ $\rightarrow_{-4}$.

(19) adibidean, ezeztapenak *azkarra* adjektiboaren orientazio semantikoa aldarazten du, positibo izatetik negatibo izatera pasatzen baita. (20) adibidean, berriz, *susmagarria* adjektiboaren orientazio semantikoak bera izaten jarraitzen badu ere, balentzian[29] igoera bat egon da *oso* intentsifikatzailearen ondorioz. (46) adibidean, ordea, modalak esaldiaren orientazio semantikoa neutralizatu egiten du. Izan ere, deskribatzen den egoera ez da gertatu eta horrenbestez, *eraginkorra* adjektiboaren balentzia +2 izatetik 0 izatera pasatzen da. Antzeko zerbait gertatzen da (22) adibidean. Bertan, azken esaldian, tempusa alegiazkoa denez, aurreko esaldietan gertatzen den egoera zalantzan jartzen da eta ondorioz, esaldiaren orientazio semantikoa neutralizatzen da. (23) adibidean egoera ezberdina da, izan ere, *nekez* adberbioak, osagai aurresuposizionala[30] denak, orientazio semantiko positiboko esaldia negatibo bilakatzen du. Azkenik, (24) adibidean, ironia dago eta horrek esaldiko hitzen sentimenduen balentzian aldaketak sortarazten ditu. Aldaketen ondorioz, esaldia +5 sentimendu balentzia izatetik, −4 sentimendu balentzia izatera igarotzen da.

Diskurtso mailan ere hainbat testuinguruko balentzia-aldatzaile daudela adierazten dute. Besteak beste, jarraian ageri diren adibideak aipatzen dituzte.

(25) [Boris matematikan [aparta]$_{+2}$ $\rightarrow_{0}$ izan arren, bera oso irakasle [txarra]$_{-2}$ da]$_{0}$ $\rightarrow_{-2}$.

(26) John asko ibiltzen da oinez. Aurreko hilean, astearteetan 25 milia egiten zituen oinez.

(27) Etxea. Kokalekua. Prezioa. Hedadura.

[29]Sentimendu-balentzia hitz batek duen informazio subjektiboa adierazteko zenbakizko eredua da. Zenbakiak eskala batean kokatuta egon ohi dira eta negatiboak edo positiboak izan daitezke hitzak duen orientazio semantikoaren arabera.

[30]Pragmatikan, aurresuposizioa (ingelesez, *presupposition*) diskurtsoan ontzat jotzen den egia duen adierazpen batekin zerikusia duen sinesmenaren edo munduaren gaineko jarrera inplizitu bat da.

(28) Hura pertsona [zerria]$_{-1}$ da. Hark esan du hura pertsona [zerria]$_{-1}$ $\rightarrow_0$ dela.

(29) Bilbo hiriaren egoera. Kokapenari dagokionez, Bilbo leku bikainean kokatzen da. Itsasoa gertu du eta inguruko mendiak ikusgarriak dira. Hiriari dagokionez, Bilbo oso kutsatuta dago eta zikinkeria edonon topa daiteke.

(30) [Filmak [txundigarria]$_+$ $\rightarrow_0$ izan beharko luke.
Aktoreak [lehen mailakoak]$_+$ $\rightarrow_0$ dira. Stallone gizon [zoriontsua]$_+$ $\rightarrow_0$ eta [zoragarria]$_+$ $\rightarrow_0$ da. Bere emaztea [gozoa]$_+$ $\rightarrow_0$ [ederra]$_+$ $\rightarrow_0$ da eta bere gizona [gurtzen]$_+$ $\rightarrow_0$ du. Gizonak [opari]$_+$ $\rightarrow_0$ [txundigarria]$_+$ $\rightarrow_0$ du [bizitza]$_+$ $\rightarrow_0$ [guztiz]$_+$ $\rightarrow_0$ [bizitzeko]$_+$ $\rightarrow_0$.
Argumentuak sekulakoa$_{+1}$ $\rightarrow_{-1}$ dirudi, hala ere, filma [porrot]$_{-1}$ bat da]$_{-2}$.

(25) adibidean, *izan arren* kontzesio menderagailuaren ondorioz, *aparta* adjektiboaren +2 sentimendu balentzia neutralizatu egiten da. (26) adibideak, aldiz, diskurtso-egitura bat islatzen du. Zehatz esanda, bertako bi esaldiak ELABORAZIOA izeneko erlaziozko diskurtso-egituraren bidez lotuta daude. ELABORAZIOA eta LISTA[31] bezalako erlazioek aurretik esandakoari buruzko informazioa gehitzen dute. Ondorioz, (26) adibidean, bigarren esaldiak lehena indartzen du.

(27) adibideak, entitate askoko ebaluazio bat islatzen du. Hau da, esaldiko gai nagusia edo aspektua etxea da eta, ondoren, horri buruzko zenbait entitate daude: kokalekua, prezioa, hedadura. Bereizi egin behar dira entitateetako bakoitzak eta aspektuak berak duen orientazio semantikoa. Kasua ezberdina da (28) adibidean, bertako perpaus osagarria edo konpletiboa dago. Perpaus osagarria denez, jarreraren ebaluazioa ez du esaldia esan duenak egiten, baizik eta beste batek. Ondorioz, *zerria* hitzaren sentimendu-balentzia neutralizatu egiten da. (29) adibidean, berriz, esaldi batean hainbat azpigai daude Bilbori buruzkoak. Kokapen geografikoari dagokionez, iritziak positiboak dira, baina hiriari buruzkoak negatiboak dira. Honek lotura zuzena

[31]ELABORAZIOA erlaziozko diskurtso-egiturak aurretik aipatutako proposizio baten aspektuari buruz informazio xeheagoa ematen duen erlazioa da. LISTA erlazioan, aldiz, lotura duten proposizioak zerrendatu egiten dira.

du aspektu eta entitate mailako sentimendu-analisiarekin, izan ere, hemengo gaia eta azpigaiak hurbilpen horretako entitatea eta aspektuak dira.

(30) adibidea genero murriztapenaren adibidea da eta (27) adibidearen antzekoa da. Bertan, filmari buruzko iruzkina egiten da baina, aldi berean, filmeko pertsonaien ebaluazioa ere egiten da. Genero murriztapena dago testuan entitate bat (filma) eta bere zenbait aspektu (tartean pertsonaiak) ageri direlako. Ondorioz, filmaren iritzia positiboa edo negatiboa zehaztu nahi bada, beharrezkoa da filmeko pertsonaiei dagozkien hitzen sentimenduen balentziak kontutan hartzea.

2.3.3. Balentzia-aldatzaileetan oinarritzen diren lanak

Polanyi eta Zaenenen (2006) laneko hurbilpen teorikoa proposatu ondoren, hainbat lan daude sentimenduen sailkapena egiteko balentzia-aldatzaileak erabiltzen dituztenak.

Kennedy eta Inkpenek (2006), esaterako, filmen iruzkinen sentimenduen sailkapena egiten dute balentzia-aldatzaileak (ezeztapena, intentsifikatzaileak eta ahultzaileak) aintzat hartuz. Lehen metodoan, iruzkinak hitz-kopuru positiboak eta negatiboak zenbatuz egiten dituzte. *General Inquirer*[32] bidez hitz positibo eta negatiboak eta balentzia-aldatzaileak identifikatzen dituzte eta sentimenduen lexikoi bat eta web corpusa ere badarabilte. Hitzen sentimenduen balentzia kalkulatzeko asoziazio kalifikazioa erabiltzen dute. Bigarren metodoan, berriz, euskarri bektoredun makina metodoa erabiltzen dute, balentzia-aldatzaileen unigramak eta bigramak ezaugarriak erabiliz.

Bestalde, Morsy eta Rafeak (2012) ere filmetako eta produktuetako iruzkinen sentimenduak sailkatzeko balentzia-aldatzaileak erabiltzen dituzte. Zehazki, intentsifikatzaileak, ezeztapena eta balentzia zeinua aldatzen duten sentimenduen hitzak erabiltzen dituzte ezaugarri moduan. Esperimentuetan, ezaugarri ezberdinak konbinatzen dituzte eta balentzia-aldatzaileekin eragin positiboa dutela adierazten dute.

[32] *General Inquirer* (Stone *et al.*, 1966) testuko edukia ordenagailu bidez aztertzeko lexikoia da. Lexikoi honek hitzen informazio sintaktiko, semantiko eta pragmatikoa biltzen du.

Azkenik, Ngoc Phu eta Thi Tuoiek (2014) dokumentu mailako sentimenduen sailkapena egiteko ere balentzia-aldatzaileak erabiltzen dituzte. Lehenik eta behin, esaldia hitzetan segmentatzen dute. Ondoren, termino kontaketa egiten dute, hots, sentimendu-balentzia duten hitzak kontatu eta horren araberako eragiketa egiten dute. Azkenik, sentimendu hitzei balentzia-aldatzaileen faktorea gehitzen diete eta eragiketa egiten dute berriz. Aintzat hartzen dituzten balentzia-aldatzaileak ezeztapena, konparaziozkoak eta superlatibozkoak, intentsifikatzaileak eta ahultzaileak dira.

Kasu honetan, gure lanak harreman zuzena du Ngoc Phu eta Thi Tuoiren (2014) lanarekin. Izan ere, bietan dokumentu mailako sentimendu-sailkapena egin nahi da sentimendu lexikoi bat erabiliz. Horrez gain, bi lanetan testuinguruko balentzia-aldatzaileak ere erabili nahi dira.

2.4 Sentimenduen analisia eta euskara

Orain arte ikusi dugun moduan, sentimenduen analisia oso arlo zabala da ataza, teknika eta ikuspegi asko baitaude. Aurrekariek erakusten duten bezala, lan gehienak ingeles hizkuntzan egin dira eta, haren atzetik, beste zenbait hizkuntza handi (tartean, gaztelera, frantsesa eta txinera) daude. Egoera arrunt ezberdina da hizkuntza gutxituetan eta horren arrazoi nagusietako bat baliabideen falta da. Atal honetan, euskara aintzat hartuta sentimenduen analisia arloan egin diren lanak azalduko ditugu. Lan horiek 2.18 taulan laburbilduta daude.

Lana	**Ataza**	**Teknikak**
Chen eta Skiena (2014)	Sentimendu-lexikoia	Hainbat lexikoi + jakintza-grafoa
Cruz *et al.* (2014)	Sentimendu-lexikoia	SentiWordNet 3.0 eredua
Barnes *et al.* (2018b)	Sentimendu-lexikoia	Embedding elebidunak
Saralegi *et al.* (2013)	Sentimendu-lexikoia	i) Corpusean oinarrituta ii) Hizkuntza arteko proiekzioan oinarrituta
Vicente eta Saralegi (2016)	Sentimendu-lexikoia	i) Beste hizkuntza batean oinarrituta ii) Corpusean oinarrituta ii) Jakintza base lexikalean oinarrituta
Barnes *et al.* (2018a)	Corpusaren sorkuntza	Webguneetatik iritzi-testuak lortu Hainbat geruzatan etiketatu
Vilares *et al.* (2017)	Lexikoian oinarritutako sentimenduen sailkapena	i) Sentimendu-lexikoia (SO-CAL eta ML-Senticon) ii) Analizatzaile sintaktikoa
San Vicente *et al.* (2015)	Aspektuan oinarritutako sentimenduen sailkapena	IXA pipe tools + Weka
Rodriguez *et al.* (2017)	Robotika	Buruaren eta besoen mugimenduak Begien mugimenduak Ahots-intonazioa Sentimendua: EliXa (IXA pipe tools + Weka)

2.18 taula: Euskara eta sentimendua uztartzen dituzten zenbait lan.

Baliabideen sorkuntzan, zehazki, sentimenduen lexikoiaren sorkuntzan, Chen eta Skienak (2014) munduko 136 hizkuntza nagusietarako, tartean euskararako, sentimenduen lexikoiak sortzen dituzte. Hainbat baliabide linguistikotan oinarrituz, jakintza-grafo handi bat sortzen dute eta hazi-hitzak zabalduz, grafoak hizkuntza bakoitzerako sentimenduen lexikoia sortzen du.

Bestalde, Cruz *et al.*ek (2014) euskara, ingelesa, gaztelania, galiziera eta katalan hizkuntzetarako sentimenduen lexikoiak sortzen dituzte. Lema mailako sentimenduen lexikoia automatikoki sortzeko metodoa aurkezten da bertan.

Beren lexikoiek hainbat geruza[33] dituzte; modu horretan, sentimendu lexikoi horiek erabiltzen dituzten aplikazioek aukera dute doitasun bat edo beste aukeratzeko eta doitasunaren arabera, lexikoiak eskaintzen dituen hitz gehiago edo gutxiago aukeratzeko.

```
<?xml version="1.0" encoding="UTF-8" ?>
<layers lang="en">
  <layer level="0">
    <positive>
      <lemma pos="a" pol="1.0" std="0.0"> admirable </lemma>
      <lemma pos="a" pol="0.438" std="0.088"> amorous </lemma>
      <lemma pos="n" pol="0.594" std="0.12"> approval </lemma>
      <!-- it continues... -->
    </positive>
    <negative>
      <lemma pos="a" pol="-0.437" std="0.157"> afraid </lemma>
      <lemma pos="n" pol="-0.275" std="0.105"> aggression </lemma>
      <lemma pos="v" pol="-0.25" std="0.0"> alarm </lemma>
      <!-- it continues... -->
    </negative>
  </layer>
  <layer level="1">
    <positive>
      <lemma pos="a" pol="1.0" std="0.0"> adept </lemma>
      <lemma pos="r" pol="0.25" std="0.0"> admirably </lemma>
      <lemma pos="n" pol="0.667" std="0.315"> admiration </lemma>
      <!-- it continues... -->
    </positive>
    <negative>
      <lemma pos="v" pol="-0.875" std="0.0"> abase </lemma>
      <lemma pos="a" pol="-0.25" std="0.0"> abashed </lemma>
      <lemma pos="v" pol="-0.25" std="0.0"> abhor </lemma>
      <!-- it continues... -->
    </negative>
  </layer>
  <layer level="2">
      <!-- it continues... -->
  </layer>
</layers>
```

2.4 irudia: ML-SentiCon lexikoiaren ingelesezko bertsioaren zati bat (Cruz *et al.*, 2014).

2.4 irudian, lexikoia geruzetan nola antolatuta dagoen ikus daiteke. Bertan, hitzak hiru geruza-mailetan (*layer level*) banatuta daude eta, horietako maila bakoitzean, hitzak daude beren sentimendu-balentziarekin. Adibidez, *amorous* ("maitekor") hitza 0 geruza-mailan dago, adjektiboa da eta 0,088

[33]Lexikoiak guztira 8 geruza ditu. Lehen geruza da murriztapenik gehiena eta lema gutxi dituena eta, ondoren, geruzaz igo ahala murriztapenak gutxitu egiten dira eta lema kopurua igo egiten da. Horrela, esaterako, bigarren geruzak lehen geruzako lemak eta beste berri batzuk ditu.

sentimendu balentzia du. 0 geruza-mailan dagoenez, bere doitasuna ez da oso handia.

Lema mailako lexikoia sortu aurretik, ingeleserako *synset* mailako lexikoia sortzen dute, *SentiWordNet 3.0*ren antzekoa dena. Azken hau sentimenduen lexikoirik erabiliena da gaur egun. Euskarazko sentimenduen lexikoiak 26.392 lema ditu (22.879 izen, 57 adjektibo, 3.456 aditz eta 0 adberbio) eta *ML-SENTICON* izeneko baliabidean eskuragarri dago.

Barnes *et al.*ek (2018b), berriz, besteen artean, euskarazko sentimenduen lexikoia sortzen dute; baina aurreko lanekiko hurbilpena ezberdina da. Euskara moduko baliabide gutxiko hizkuntzetan, corpus etiketatu falta handia dago eredu onak sortzeko. Horregatik, sentimendu *embedding* elebidunak (*Bilingual Sentiment Embeddings*, BLSE) sortzen dituzte beste hizkuntza bat eta lexikoia sortu nahi den hizkuntza uztartuz. Horretarako, baliabide hauek erabiltzen dituzte: i) lexikoi elebidun txiki bat; ii) jatorri hizkuntzako corpusa sentimenduz etiketatuta; iii) hizkuntza bakoitzeko hitz *embedding* elebakarrak. Esperimentuak gaztelania, katalanera eta euskaraz konbinatuz egiten dituzte esaldi mailako hizkuntzen arteko sentimenduak sailkatzean.

Saralegi *et al.*ek (2013) ere sentimenduen lexikoia sortzen dute. Euskara du aztergaitzat eta bi hurbilpen ezberdin alderatzen eta aztertzen ditu euskarazko sentimenduen lexikoiak sortzeko. Bi hurbilpenak automatikoak izateaz gain, erraz eskuratzeko baliabideetan oinarritzen dira eta, beraz, baliabide gutxiko hizkuntzetarako egokia da. Hurbilpen horietako bat corpusean oinarritutakoa da eta bestea, berriz, hizkuntzen arteko proiekzioan. Vicente eta Saralegik (2016), berriz, euskarazko sentimendu lexikoia sortzeko hiru metodo alderatzen dituzte: i) beste hizkuntza batetik itzuliz, ii) corpusetik sentimendu kalifikazioa duten hitzak erauziz eta iii) jakintza-base lexikaletik (*Lexical Knowledge Bases*) sentimenduak etiketatuz. Metodo bakoitzak zenbat eskuzko lan eta lortzen diren emaitzekiko eskuzko lan horrek merezi duten ebaluatzen dute lanean.

Bestalde, Barnes *et al.*ek (2018a) sentimenduen analisirako corpus bat sortzen dute. Bertan, aspektu mailako sentimenduen sailkapena egiteko, katalaneraz eta euskaraz dauden hotelei buruzko iritzien corpus anotatua sortzen dute. Katalanerazko 568 eta euskarazko 343 iruzkin biltzen dituzte.

Iruzkinak aurreprozesatu ondoren, OpeNER proiektuko hurbilpenaz etiketatzen dituzte eta etiketatzeko *KafAnnotator* tresna erabiltzen dute. Anotatu beharreko elementuak iritzi-helburua edo jomuga, iritzi-adierazpenak, iritzi-eramailea eta polaritatea dira.

San Vicente *et al.*en (2015) *EliXa* sistema aspektuan oinarritutako sentimenduen analisia egiteko tresna da. Sistema honen abantaila da plataforma modularra dela; horrek aukera ematen du tresnari ezaugarriak kentzeko edo berriak gehitzeko. Sistema *IXA pipe tools* (Agerri *et al.*, 2014) eta Wekan oinarritzen diren hiru modulu independientez osatuta dago. Tresnari ezaugarri berriak gehitu ahal izateak egiten du posible sistema hau euskararako baliagarri izatea. *EliXa*k hizkuntza-baliabide zehatzak erabil ditzake, hala nola polaritate lexikoak eta testua normalizatzeko balio duten baliabideak. Momentuz, lau hizkuntzarako lau baliabide eskaintzen dituzte, tartean, euskararako. Euskararako baliabideak Elhuyarren sentimenduen lexikoia, maiztasun handiko hitzen zerrenda (ingelesez, *stopwords*) eta hiztegitik kanpoko hitzak (ingelesez, *Out-Of-Vocabulary*, OOV) biltzen ditu, besteak beste.

Vilares *et al.*ek (2017) sentimenduen lexikoia eta sintaxia uztartzen dituzte sentimenduen sailkapena egiteko. Iberiar penintsulako hizkuntzetarako polaritatearen sailkapena egitea dute helburu eta, horretarako, sentimenduen lexikoiaz eta erregela sintaktikoez[34] baliatzen dira. Hurbilpen hori analizatzaile sintaktikoan oinarritzen da. Gainera, erregelak dependenteak dira eta egoera eleanitzetan moldaketa bat behar dute. Euskararen kasuan ez ezik, baita katalaneran eta galizieran ere, aurretik dauden hainbat lexikoi bateratu egiten dituzte. Erabiltzen duten lexikoietako bat SO-CAL tresnakoa (Taboada *et al.*, 2011) da. Euskarazko bertsioa itzulpena eginez lortzen dute eta 4.066 sarrerako lexikoia lortzen dute. Erabilitako beste lexikoia *ML-Senticon* da. Lexikoi horren euskarazko bertsioak 1.662 sarrera ditu. Bi lexikoien bateraketa egin ondoren, lortzen duten lexikoiak 5.134 sarrera ditu. Sortzen duten baliabidearen beste atala gramatika-kategorien etiketatzailea eta dependentzia etiketatzaileak dira. Kasu hauetan, aurretik sortuta dauden baliabideak erabiltzen dituzte. Azken urratsean, konposizio-eragiketak egiten dituzte.

[34]Erregela sintaktikoak esaldien egitura finkatzen dituzten erregelen multzoa da.

Rodriguez *et al.*ek (2017), berriz, aurretik aipatu dugun *EliXa* sistema robotikan aplikatzen dute. Gizakiari izaera soziala ematen diona emozioen adierazpena da. Idatzizko eta ahozko moduez gain, pertsona baten egoera mentala adierazten duen moduetako bat da. Robot baten eta gizaki baten artean, hori islatzea beharrezkoa da, interakzioa ahalik eta naturalena izateko. Lan honetan, ahozko testuari eduki emozionala jartzen diote. Buru- eta beso-mugimenduekin, begien mugimenduarekin eta ahots intonazio ezberdinak erabiliz, robota gai da pozezko edo tristurako emozioak adierazteko, bai euskaran, baita ingeles eta gaztelanian ere. Robota, euskararen kasuan, emozio bat 0 eta +10 arteko eskalan sailkatzeko (bi muturrak tristura eta poztasuna dira) *EliXa* sisteman oinarritzen da.

Atal honetan aipatutako lanek badituzte zenbait berdintasun eta ezberdintasun guk garatutako tresna eta baliabideekin. Sentimenduen lexikoiaren sorkuntzan, gure hurbilpena izan da SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskaratzea eta ingelesekoarekin aberastea. Bestelako hurbilpenak erabili dituzte aipatu ditugun lanetan, izan ere, batzuetan *SentiWordnet* erabili dute (Cruz *et al.*, 2014), beste batzuetan, berriz, *embedding* bezalako testu-analisirako teknikak erabili dituzte (Barnes *et al.*, 2018b) eta azkenik, badira gure antzera corpusak erabili dituztenak ere (Saralegi *et al.*, 2013), nahiz eta ikuspegia ezberdina izan.

Corpusaren sorkuntzari dagokionez, Barnes *et al.*ek (2018a) gure teknika bera erabiltzen dute; hots, webguneetatik iritzi-testuak lortzen dituzte. Gero, hala ere, egindako etiketatzea ezberdina da. Azkenik, dokumentu mailako sentimenduen sailkapenean, gure lanak Vilares *et al.*enarekin (2017) antzekotasunak ditu, izan ere, bietan lexikoian oinarritutako hurbilpena eta sintaxia baliatzen dira, nahiz eta guk diskurtso maila ere lantzen dugun.

2.5 Laburpena

Kapitulu honetan, sentimenduen analisi arloan dauden ataza eta aurrekari ezberdinak aztertu ditugu. Ikusi dugunez, sentimenduen arloan hiru ataza nagusi daude: i) dokumentu mailako sentimenduen sailkapena, ii) subjektibitatearen sailkapena esaldi mailan eta iii) iritziaren jabearen erauzketa entitate eta aspektu mailan. Horietaz gain, beste ataza batzuk ere badaudela ikusi dugu eta hizkuntzaren prozesamenduaren beste arloekin lotura zuzena dutela, iritzi-laburpena edota *spam*ak diren iritzien detekzioa, adibidez.

Baliabide-sorkuntzari dagokionez, hainbat motatako iritzi-testuen corpusak daudela ikusi dugu. Testu-genero ezberdinak biltzen dituzte batzuek. Beste batzuk, berriz, corpus eleanitzak dira eta corpus gehienak etiketatuta daudela ere ikusi dugu. Sentimenduen lexikoiari dagokionez, horiek sortzeko biderik ohikoena corpus edo hiztegi bidezkoa da.

Balentzia-aldatzaileen identifikazioan egin diren lanak ere landu ditugu. Balentzia-aldatzailearen kontzeptua testuaren idazleak testuak eskaintzen dituen baliabideen bidez iritzia adierazteko moldeak identifikatzeko sortu dela azaldu dugu eta, ondoren, horren inguruko erreferentziazko lana den Polanyi eta Zaenenek (2006) zer-nolako balentzia-aldatzaileak deskribatzen dituen aztertu dugu. Azkenik, balentzia-aldatzaileak kontuan hartuz sentimenduen analisian egin diren zenbait lan aipatu ditugu.

Azken atalean, euskara eta sentimenduen analisia lantzen dituzten lanak deskribatu ditugu. Lan gehienek euskarazko sentimenduen lexikoia sortzea dute helburutzat, askotan, horretarako SentiWordNet datu-base lexikala erabiliz.

Gure aukerei dagokienez, corpusaren kasuan, gai ezberdinetako iritzi-testuez osatuta egotea nahi dugu, baita orientazio semantikoaren aldetik orekatua eta diskurtso-egitura RST hurbilpenaz etiketatuta egotea ere. Sentimenduen lexikoiari dagokionez, gure aukeraketa hiztegian eta corpusearen oinarritzen da. Guk sortu nahi dugun dokumentu maila sentimenduen sailkatzailea, berriz, lexikoian oinarritutakoa izatea hautatu dugu. Azkenik, balentzia-aldatzaileei dagokienez, Polanyi eta Zaenenek (2006) deskribatzen dituenen artean, sintaxi eta diskurtso mailakoak aztertzea erabaki dugu, fonologia eta morfologia mailekoez gain.

METODOLOGIA

3. KAPITULUA

Metodologiaren diseinua

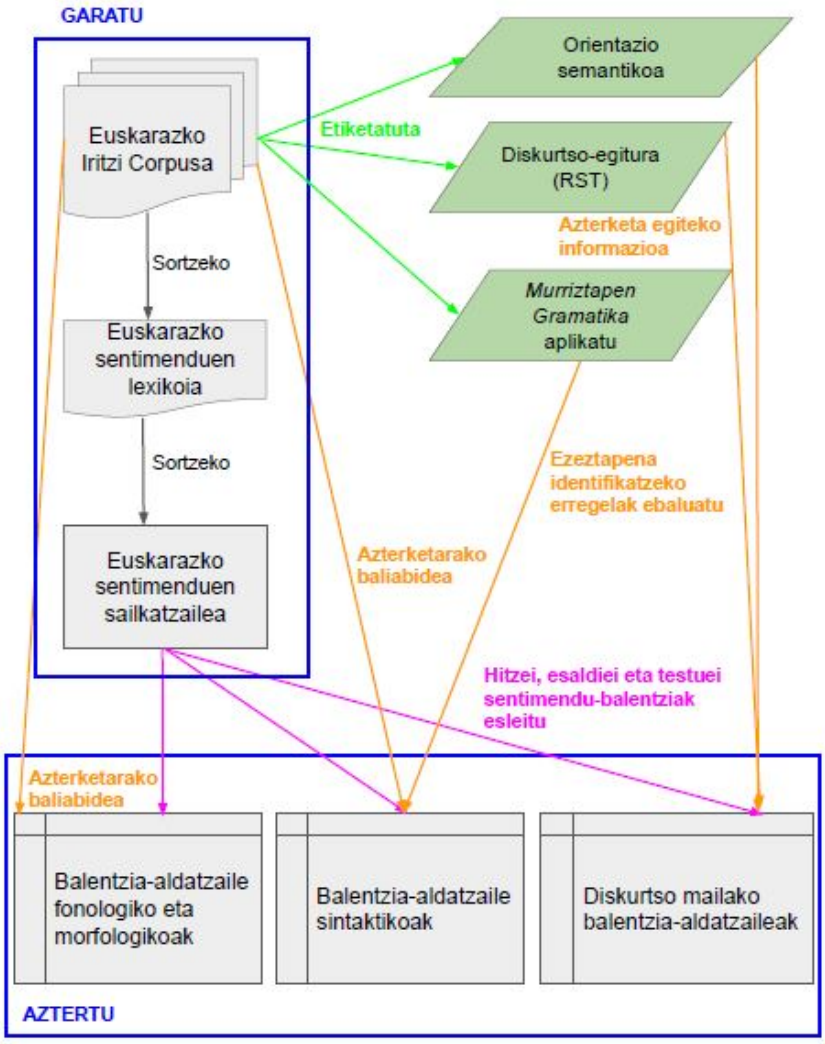
1. kapituluan zehaztutako helburuak betetzeko metodologia bat jarraitu eta zenbait baliabide eta tresna erabili ditugu. Kapitulu honetan horien inguruan arituko gara. 3.1 irudian, tesi-lan honen metodologiako urratsak laburbilduta ikus daitezke.

1- Ikerketa aurrera eramateko oinarrizko baliabideak sortu. Zehazki, Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016), euskarazko sentimenduen lexikoia eta euskatrazko sentimenduen sailkatzaile bat sortu ditugu.

   Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) 240 iritzi-testuz osatuta dago eta etiketatuta ere badago. Alde batetik, iritzi-testu osoaren orientazio semantikoa positiboa edo negatiboa den etiketatu dugu. Beste aldetik, corpuseko 70 iritzi-testuren diskurtso-egitura etiketatu dugu *Egitura Erretorikoaren Teoria* (RST) (Mann eta Thompson, 1988) erabiliz. Azkenik, corpuseko 48 iritzi-testuei *Murriztapen Gramatika* (MG) (Karlsson *et al.*, 1995) aplikatu diegu.

   Euskarazko sentimenduen lexikoia ere sortu dugu eta, horretarako, besteak beste, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarritu gara. Azkenik, euskarazko sentimenduen sailkatzailea ere garatu dugu eta bere oinarrian sortu dugun sentimenduen lexikoia dago.

2- Balentzia-aldatzaileak identifikatu. Bigarren urratsa, hizkuntza maila ezberdinetan, sentimendu-balentzia duten hitzetan eragin dezaketen fenomeno linguistikoak aztertu eta beren eragina neurtzea izan da.



3.1 irudia: Tesi-lanaren metodologia.

Balentzia-aldatzaile fonologikoak aztertu ditugu eta azterketa egiteko Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) eta euskarazko sentimenduen sailkatzailea erabili ditugu. Balentzia-aldatzaile sintaktikoetan ezeztapen-markak aztertu ditugu. Ezeztapen-marken adibideak Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) aurkitu ditugu eta, ondoren, ezeztapen-marka horiek identifikatzeko erregelak *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortu eta Euskarazko Iritzi Corpu-

sari(Alkorta *et al.*, 2016) *Murriztapen Gramatika* (Karlsson *et al.*, 1995) aplikatuta, erregela horiek ebaluatu ditugu.

Azkenik, diskurtso mailan, erlaziozko diskurtso-egitura eta unitate zentrala aztertu ditugu eta balentzia-aldaketan zer-nolako eragina duten zehaztu dugu. Hori egiteko Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) (diskurtso-egituraz eta orientazio semantikoz etiketatuta) eta euskarazko sentimenduen sailkatzailea erabili ditugu.

Jarraian, metodologiaren deskribapen zabalagoa egin dugu.

3.1 Sentimenduen analisirako baliabide eta tresnen sorkuntza

3.1.1. Euskarazko Iritzi Corpusa

Euskarazko Iritzi Corpusa sortzeko, lehenik eta behin, zer-nolako corpusa sortu nahi dugun erabaki dugu. Corpuseko testuek jarraian azaltzen diren ezaugarriak izatea erabaki dugu:

- Iritzi-testuak balorazio argi bat izatea. Testu batzuek iritzi argirik ez dutela oharturik, aldeko edo aurkako iritzi argia duten testuak hobetsi ditugu. Iritzi argia adierazten duten iritzi-testuak biltzen baditugu, tesi-lan honetan garatutako sentimenduen sailkatzailea hobeto ebaluatuko dugu.
- Iritzi-testuak egitura sintaktiko aberatsa eta balorazioa adierazteko baliabideak edukitzea. Iritzi-testuek egitura sintaktiko ezberdinak eta balorazioa egiteko molde ezberdinak edukitzea nahi izan dugu. Modu horretan, hizkuntza mailetako eta molde ezberdinetako balentzia-aldatzaileak identifikatzeko aukera izango dugu.

Irizpideak zehaztu ondoren, euskarazko iritzi-testuak dauden webguneak bilatu ditugu. Webgune horiek espezializatuak dira edota aldizkari eta egunkarienak dira. Labur azalduta, hiru iturri ezberdin erabili ditugu: i) egunkariak (hala nola, *Berria*, *Argia*, *Naiz* eta *El Correo*), ii) webgune espezializatuak (*Kritiken Hemeroteka*, *Entzun.com Zinea.eus* eta *Aizu*) eta, azkenik, iii) hainbat blog (horien artean, Joxe Landaren bloga, Baleikeko eguraldia atala eta EiTBko txirrindularitza bloga).

Euskarazko Iritzi Corpusa osatzeko webguneak aurkitu ostean, corpusak eduki beharreko ezaugarriak finkatu ditugu. Ezaugarriak hiru dira:

- Corpusak tamaina handia izango du. Corpusa 240 iritzi-testuz osatuta egongo da. Gure ustez, tamaina honetako corpusa nahikoa da gure helburuak betetzeko eta sentimenduen analisiko beste corpusek ere antzeko tamaina dute.

3.2 irudia: *Kritiken Hemerotekaren* webgunea.

- Corpusa domeinu anitzekoa izango da. Iritzi-testuak sei domeinutakoak izango dira: eguraldia, politika, kirola, zinema, musika eta literatura. Corpusa domeinu askotakoa izatea nahi dugu, modu horretan, domeinuari loturiko berezitasunak edo balentzia-aldatzaile zehatzak identifikatzeko aukera izango baitugu.

- Corpuseko testuek balorazio orekatua izango dute. Domeinu bakoitza testuek erakusten duten balorazioari dagokionez orekatua izango da. Hau da, domeinu bakoitzean balorazio positiboko 20 testu eta balorazio negatiboko beste 20 testu egongo dira. Corpusa orekatua izatea erabaki dugu eredutzat hartu dugun *SFU Review Corpus* (Taboada, 2008) halakoa delako.

Hurrengo urratsean, bildutako iritzi-testuekin datu-base bat sortu dugu. Datubase horretan, iritzi-testu bakoitzaren ezaugarri hauek zehaztu ditugu:

- Kodea. Iritzi-testu bakoitzari kode bat jarri diogu. Kode horrek beti *GAIA-zenbakia_balorazioa* egitura du.
  - Gaia: liburua (LIB), musika (MUS), pelikula (ZIN), politika (POL), kirola (KIR) eta eguraldia (EGU).

  - Zenbakiak: 1etik 40 arterainoko zenbakiak jarri dizkiegu iritzi-testuei, gaiko 40 iritzi-testu bildu baititugu.
  - Balorazioa: positiboa (POS) eta negatiboa (NEG).

- Izenburua. Testu bakoitzaren izenburua izan da zehaztu dugun beste ezaugarrietako bat.

- Helbidea. Iritzi-testua Interneten non dagoen ere zehaztu dugu, iturria onartzeko eta edozeinek lortu ahal izateko.

- Balorazioa. Datu-basean, iritzi-testuetako bakoitza positiboa edo negatiboan den ere zehaztu dugu.

- Tamaina. Iritzi-testuak duen hitz-kopurua ere zehaztu dugu. Hitz-kopurua kontatzeko, *Analhitza* tresna (Otegi *et al.*, 2017) erabili dugu.

Iritzi-testuetako bakoitzari loturiko informazioa datu-basean nola egituratuta dagoen ikus daiteke 3.1 taulan.

Domeinua	**Kodea**	**Izenburua**	**Iturria**	**Balorazioa**	**Hitz-kopurua**
Literatura	LIB29_pos	Zangotraba	Aizu	Positiboa	175
Kirola	KIR15_neg	Vigotik esku hutsik itzuli da Reala	Berria	Negatiboa	169

3.1 taula: Datu-basearen antolaketaren bi adibide.

Bestetik, iritzi-testuak formatu egokian ere jarri ditugu Hizkuntza Prozesamenduko tresnekin prozesatu ahal izateko. Testuak *txt* eta UTF-8 karaktere kodifikazio formatuan jarri ditugu.

Sortuko dugun corpusa ikerketarako egokia den ere neurtu nahi izan dugu eta horregatik, corpusa ingelesezko beste corpus subjektibo batekin (*SFU Review Corpus* (Taboada, 2008)) eta ingelesezko eta euskarazko bi corpus objektiboekin (corpus subjektiboen gai berak lantzen dituzten Wikipediako artikuluen multzoa) alderatu dugu. Honako aspektuak aztertu ditugu:

- Lehen pertsonaren presentzia. Ingelesean pertsona izenordainetan eta euskaran aditzetan, lehen pertsonak zenbateko presentzia duen neurtu dugu. Pertsonek beren esperientziak lehen pertsona singularrean edo pluralean azaltzen dituztelako aztertu dugu (Li *et al.*, 2011; Villarroel Ordenes *et al.*, 2017). Euskararen kasuan, 3.3 irudian agertzen diren instantzia guztietatik zenbat ageri diren lehen pertsona singularrean edo pluralean kontatu dugu.

```
Emaitza indibidualak
NI_ZURI 10
NR_GU   74
NR_HI   34
NK_NIK  185
NR_ZUEK 4
NR_HURA 4930
NR_HAIEK    1266
NI_ZUEI 3
NI_NIRI 77
NO  3
NI_GURI 149
NR_NI   73
NK_ZUK  22
NK_GUK  300
NI_HAIEI    77
NI_HARI 273
NK_HAIEK-K  635
NK_HIK-TO   1
NK_HIK-NO   32
NR_ZU   7
NK_HARK 2077
NK_ZUEK-K   12
NI_HIRI-NO  6
```

3.3 irudia: Euskarazko aditzetan dagoen lehen pertsonaren erabilearen neurketa.

- Adjektiboen presentzia. Gramatika-kategoria guztietatik, adjektiboek zer pisu duten neurtu dugu bina corpus objektibo eta subjektibotan. Pang *et al.*ek (2002) diotenez, adjektiboak izan dira gramatika-kategoriarik erabilienak sentimenduen sailkapeneko atazetan eta doitasun ona lortu dute. Horregatik aztertu dugu adjektiboen presentzia.

- Ezeztapen-markak. Zenbait lanen arabera (Dadvar *et al.*, 2011; Wiegand *et al.*, 2010; Jia *et al.*, 2009) ezeztapen-markek hitzen edota sin-

tagmen sentimenduen balentzian eragin dezaketelako eta hori euskaran aztertu nahi dugulako, euskarazko iritzi corpusean zenbat ezeztapen-marka dauden zenbatu dugu. Ezeztapen-marken kasuan ere, zenbaketa ordenagailuak eskaintzen dituen baliabideen bidez egin dugu.

Bukatzeko, Euskarazko Iritzi Corpusa diskurtso- eta sujektibitate-informazioz etiketatu dugu. Lehenik eta behin, corpuseko 240 testuetatik 70 iritzi-testu *Egitura Erretorikoaren Teoria* (RST) hurbilpena baliatuz etiketatu ditugu. Beste lanetan etiketatu den adina testu etiketatu dugu, eta, ondorioz, ez dugu corpus osoa etiketatu. Bestalde, RST erabili dugu euskaran gehien landu den hurbilpena delako. 3.2 taulan ikusten den moduan, guztira 70 testu (corpusaren % 29,16) etiketatu ditugu eta horietatik 19 testuek (etiketatutakoaren % 27,14) etiketatze bikoitza dute.

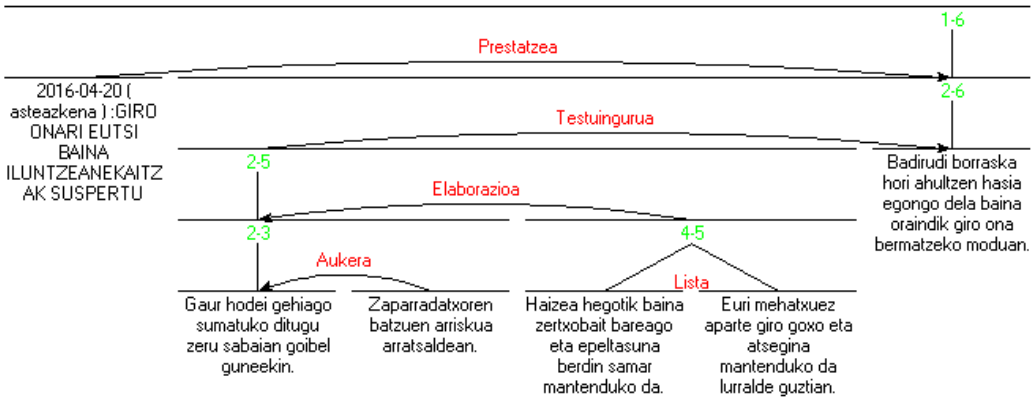
	A1	**A2**	**Etiketatuak guztira**	**Etiketatze bikoitza**
Zinema	30	9	30	9
Eguraldia	15	5	15	5
Literatura	5	25	25	5
Guztira	**50**	**39**	**70**	**19**

3.2 taula: Bi anotatzailek (A1 eta A2) anotatutako iritzi-testuen kopurua.

Anotazioa burutzeko, bi anotatzaileen Das eta Taboadaren (2018) irizpideei jarraitu behar izan diete. Domeinutik domeinura ezberdintasun batzuk egon dira etiketatzeko prozesuan. Esaterako, anotatzaileek eguraldiaren domeinuko testu bat anotatzeko 20 minutu inguru igaro behar izan dituzte eta literaturakoak anotatzeko, aldiz, ordu bete inguru. Horren arrazoiak testuen luzera eta konplexutasuna dira, eguraldikoak testu laburragoak eta sinpleagoak baitira, literaturakoen aldean. 3.4 irudian, eguraldiaren domeinuko iritzi-testu bat (EGU04) anotatuta ikus daiteke. Bi anotatzaileen arteko adostasunaren emaitzak 3.3 taulan daude. Zehazki, erlaziozko diskurtso-egitura ondo etiketatu den neurtu dugu.

Corpusaren etiketatzea bi modutan ebaluatu dugu. Alde batetik, eskuz unitateen esanahi erretorikoa ebaluatu dugu eta, bertan, azpiosagai zentralaren[1]

[1]Azpiosagai zentrala *span*eko unitaterik garrantzitsuena da (Egg eta Redeker, 2010).



3.4 irudia: EGU04 iritzi-testuaren diskurtso-egituraren anotazioa.

posizioa ez da aintzat hartu. Beste aldetik, ebaluazio kualitatiboa egin dugu. Kasu horretan, unitateen esanahi erretorikoaz gain, nuklearitatea, lotura eta osagaik hartu dira kontuan. Halaber, eskuzko ebaluazioan ez bezala, azpiosagai zentralaren posizioa aintzat hartu da ebaluazioa egiteko momentuan.

3.3 taulako emaitzek, eskuzko ebaluazioaren emaitzek, erakusten dutenez, erlaziozko diskurtso-egitura etiketatzean bi pertsonen arteko adostasuna % 9,81ekoa da. Ezbedintasunak daude domeinutik domeinura. Eguraldiaren domeinuan, adostasuna % 43,59ra igotzen da eta zinemaren domeinuan, aldiz, adostasuna % 37,73ra jaisten da. Literaturaren domeinuan, adostasuna % 41,67koa da.

Domeinua	**Adostasuna (%)**
Eguraldia	43,59 (17/39)
Literatura	41,67 (70/168)
Zinema	37,73 (83/220)
Guztira	**39,81 (170/427)**

3.3 taula: Bi etiketatzaileen arteko adostasun-maila iritzi-testuak RST hurbilpenaz etiketatzean (eskuzko ebaluazioa).

Bi pertsonen arteko adostasuna neurtu dugu ebaluazio kualitatiboa[2] egiten duen tresna erabiliz ere. Emaitzak 3.4 taulan daude ikusgai.

[2]Iruskieta *et al.*ek (2015) aipatzen duten moduan, ebaluazio mota honi kualitatiboa

	Osagaiak		Lotura		Nuklearitatea		Erlazioa	
Domeinua	**Bat egin**	**F1**	**Bat egin**	**F1**	**Bat egin**	**F1**	**Bat egin**	**F1**
Eguraldia	20/37	0,54	9/37	0,24	22/37	0,59	15/37	0,41
Literatura	84/155	0.54	67/155	0,43	105/155	0,68	48/155	0,31
Zinema	112/221	0,56	88/221	0.40	147/221	0,67	68/221	0,31
Guztira	**216/413**	**0,52**	**164/413**	**0,40**	**274/413**	**0,66**	**131/413**	**0,32**

3.4 taula: Etiketatzaileen arteko adostasunari buruz tresna automatikoak egindako ebaluazio kualitatiboa.

Eskuzko neurketan ez bezala, ebaluazio kualitatiboan[3] erlaziozko diskurtso-egitura motaz gain, beste zenbait aspektu neurtu ditugu. Diskurtsoen erlazio motari dagokionez, adostasuna 0,32koa da, eskuzkoa baino 0,08 puntu baxuagoa. Ezberdintasun horren atzean, eskuzko ebaluazioan ez bezala, tresnak azpiosagai zentralaren (ingelesez, *central subconstituent*) posizioa kontuan hartzen duela dago. Ebaluatu diren beste aspektuetan, adostasuna handiagoa da. Loturaren kasuan, adostasuna 0,40ekoa da eta osagai eta nuklearitate aspektuetan adostasuna 0,50etik gorakoa da, 0,52 eta 0,66koak baitira, hurrenez hurren.

Diskurtso-egitura etiketatu ondoren, corpus bera subjektibitatearen ikuspegitik etiketatu da. Zehazki esanda, testuetako eta erlaziozko diskurtso-egiturako zenbait osagairen orientazio semantikoa izan da etiketatu dena.

Bi anotatzaileek erlaziozko diskurtso-egituretako hiru osagai etiketatu behar izan dituzte: i) erlaziozko diskurtso-egitura bera, ii) haren nukleoa edo nukleoak (KONTRASTEA erlazioaren kasuan) eta iii) haren satelitea. Esleitu beharreko etiketak ere hiru dira: i) orientazio semantiko positiboa, ii) orientazio semantiko negatiboa eta iii) orientazio semantiko neutrala.

3.5 taulan, anotatzaile bat erlaziozko diskurtso-egitura etiketatzen ikus daiteke. Anotatzaileak lehen lerroan nukleoa den testu-zati bat du eta orientazio

deritzo hizkuntza ezberdinetan edo/eta pertsona ezberdinek garatutako diskurtso-egiturak alderatzeko aukera ematen duelako. Orain arteko ebaluatzeko moduak kuantitatiboak izan dira, baina honen ezberdintasuna da EDUak, testu-zatiak, nuklearitatea eta erlaziozko diskurtso-egiturak hartzen dituela kontuan.

[3]Ebaluazio kualitatiboak aspektu hauek aztertzen ditu: osagaia (oinarrizko diskurtso-unitateak), lotura (unitateek erlazioekin duten lotura), nukleartasuna (erlazioa N(ukleo)-S(atelite), S(atelite)-N(ukleo) edo N(ukleo)-N(ukleo) den) eta erlazioa (unitateen esanahi erretorikoa).

semantiko positiboa esleitu dio. Bigarren zutabean, satelitea den testu-zatia dago eta horri orientazio semantiko neutrala esleitu dio. Azkenik, hirugarren lerroan, erlaziozko diskurtso-egitura osoari orientazio semantiko positiboa esleitu dio.

Kodea	Erlazioa	Esaldia	Orientazio semantikoa	N-S
SENTARG01-A1.rs3	AHALBIDERATZEA	Aukera ederra da irakurlearentzat, Mirandek gaztetik hasi eta ia hil arte eutsi zion ohitura horri esker, modu kronologikoan emana gainera.	POS	N
		Froga bat egin liteke, herrien izena eta batez ere bere izena zein hizkuntzatan idazten duen begiratzea, une horretan bere herriarekiko sentimenduak eta harremanak nolakoak diren jakiteko.	NEU	S
		Aukera ederra da irakurlearentzat, Mirandek gaztetik hasi eta ia hil arte eutsi zion ohitura horri esker, modu kronologikoan emana gainera. Froga bat egin liteke, herrien izena eta batez ere bere izena zein hizkuntzatan idazten duen begiratzea, une horretan bere herriarekiko sentimenduak eta harremanak nolakoak diren jakiteko.	POS	N-S

3.5 taula: Erlaziozko diskurtso-egitura baten orientazio semantikoaren esleiketa.

Guztira, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden 240 iritzi-testuetatik 28 iritzi-testu erabili ditugu eta orientazio semantikoaren etiketatzea iritzi-testu horietako 384 erlaziozko diskurtso-egituretan egin dugu. 3.6 taulan ikusten den moduan, etiketatutako erlaziozko diskurtso-egiturek esanahi erretoriko ezberdinak dituzte[4]. Etiketatzaile batek 384 erlaziozko diskurtso-egitura etiketatu ditu eta beste batek horien % 40 etiketatu, etiketatzearen kalitatea neurtzeko. Hain zuzen ere, hori izan da etiketatzearen kalitatea neurtzeko erabili dena.

Horretarako, gidalerro batzuk sortu ditugu. Gidalerro horietan etiketatzaile bakoitzak egin beharrekoa azaltzen da. Kasu gatazkatsuenen berri ere ematen da, tartean esaldi metaforikoena, eta horiek nola anotatu behar diren ere esplikatzen da.

(31) (...) eta “egilearen intentzioa” $[\text{usna}]_+$ liteke. (SENTBER04)

[4]Orientazioa semantikoa etiketatuta duten iritzi-testuak *Basque RST Treebank* baliabidean (Iruskieta *et al.*, 2013) eskuragarri daude.

Erlaziozko diskurtso-egitura	**Instantziak**
AHALBIDERATZEA/MOTIBAZIOA	6
Baldintza taldea	18
KONTRASTEA	26
EBIDENTZIA/JUSTIFIKAZIOA	53
KONTZESIOA/ANTITESIA	70
Kausa taldea	71
EBALUAZIOA/INTERPRETAZIOA	140
Guztira	**384**

3.6 taula: Etiketatutako erlaziozko diskurtso-egiturak.

Esate baterako, (31) adibidean, *usnatu* aditza ageri da. Ez da berak duen berezko zentzuan ageri eta, adibide horretan, *usnatu* "egileak zer egin nahi duen antzematea" esatearen parekidea da. Horregatik, mota horretako adibideen orientazio semantikoa etiketatzea erabaki dugu. Etiketatzaileen etiketatze-adostasunaren emaitzek erakusten dutenez, Cohenen kapparen koefizientea 0,58 da. Beraz, *neurrizko adostasuna* (Landis eta Koch, 1977) da.

Bi anotatzaileen arteko desadostasunik handienak orientazio semantiko neutralarekin lotuta daude, 3.7 taulan ikus daitekeen moduan.

		A2			
		NEG	**NEU**	**POS**	**Guztira**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Guztira	92	120	181	393

3.7 taula: Erlaziozko diskurtso-egituren sentimendu balentzia anotazioari dagokion bi anotatzaileen arteko kontingentzia-taula.

Etiketatzaile batek orientazio semantiko neutrala esleitzen duenean, beste anotatzaileak ez du esleitzen eta alderantziz ere gertatzen da. Hainbat etiketatze-proba egin ostean eta gidalerroak hobetu ostean, etiketatzaile batek 384 erlaziozko diskurtso-egituren gainerako % 60aren orientazio semantikoa etiketatu du.

3.1.2. *Sentitegi* izeneko euskarazko sentimenduen lexikoiaren sorkuntza eta ebaluazioa

Euskarazko sentimendu lexikoia sortzerakoan, lehenik eta behin, horretarako zer-nolako baldintzak dauden aztertu dugu. Hiru aspektu nabarmentzen dira:

- Denbora. Sentimenduen lexikoia lanaren muina izan arren, ezin izan dugu denbora guztia horretan baliatu. Sentimenduen lexikoitik abiatuz, sentimenduen analisia atazaren hainbat alderdi (balentzia-aldatzaileak eta dokumentu mailako sentimenduen sailkatzailea, adibidez) lantzea dugu helburu. Horrek esan nahi du denbora banatu egin behar dugula sentimenduen lexikoia sortzearen eta sentimendu-analisiko hainbat atal lantzearen artean. Ondorioz, euskarazko lexikoia sortzeko denbora mugatua dugu.

- Tresnak eta baliabideak. Lexikoia sortzeko izan dugun beste mugetako bat eskuragarri dauden euskarazko tresna eta baliabideek dituzten ezaugarriak dira. Euskarazko hainbat sentimendu lexikoi sortu izan dira hurbilpen ezberdinak baliatuta (Chen eta Skiena, 2014; Cruz *et al.*, 2014; Barnes *et al.*, 2018b; Saralegi *et al.*, 2013; Vicente eta Saralegi, 2016), baina hitzei sentimendua esleitzeko moldea guretzat ez da egokia hainbat arrazoirengatik. Batzuetan, hitzari esleitutako sentimenduen balentzia ez dago eskala batean eta, horrenbestez, ezin da hitz batean gerta daitekeen intentsitatearen aldaketa neurtu balentzia-aldatzaileen eraginez. Beste batzuetan, berriz, lexikoian, hitzen sentimenduen balentziak bi eskaletan daude (positiboan eta negatiboan) hitz horiek duten polisemia tarteko eta, azkenik, eskala bera egokia ez den kasuak ere badaude. Hori horrela izanik, sentimendu lexikoi bat euskarara itzultzea erabaki dugu.

- Kalitatea. Gure helburua ahalik eta kalitate handieneko sentimenduen lexikoia sortzea da. Aurreko puntuan aipatu bezala, tresna eta baliabideek kalitatean eragin dezakete. Bestalde, ebaluatzeko modukoa den eta gerora hobetu daitekeen lexikoi bat sortu nahi dugu. Kalitatea eta bere ezaugarriak neurtzeko, aurretik dauden lexikoien antzeko

ezaugarriak dituzten lexikoi bat eratu nahi dugu. Ondorioz, sortutako lexikoiaren kalitatea neurtzeko, dagoen lexikoi baten antzekoa sortzeko beharra ikusi dugu.

Hurrengo urratsean, aurretik aipatutako baldintzak aintzat hartuta sentimenduen lexikoia nola sortu erabaki dugu. SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia jarraian azalduko dugun metodoari jarraituz itzultzeko erabakia hartu dugu. SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoia gaztelaniazko lexikoiko hitzari euskarazko ordaina emateko ere erabiltzea erabaki dugu. Honako abantailak ikusi ditugu, erabakia hartzerakoan:

- SO-CAL tresnako (Taboada *et al.*, 2011) lexikoien ezaugarriak. Hizkuntzako fenomeno ezberdinak kontuan hartzeko lexikoiko hitzen balentziak, sentimendua adierazten duten balioak, −5 eta +5 artean daude. Gure ustez, balentziak eta balentzien eskala egokiak dira guk sentimenduen analisian aztertu nahi ditugun alderdietarako, hain zuzen ere, balentzia-aldatzaileek eragiten dutena eskala horretan neur baitaiteke.
- Itzulpenak egiteko baliabideak. Lehen aipatu dugun bezala, sentimenduen lexikoia euskaraz sortzeko baliabideak lortzea zaila da eta daudenak dituzten ezaugarriengatik ez dira egokiak. Baina hiztegigintzan euskaran hainbat eta hainbat baliabide daude, eta hori dela eta, horiek erabiltzea egokia da. Besteen artean, *Elhuyar* hiztegia (Elhuyar, 2013) eta *Zehazki* hiztegia (Sarasola, 2005) sarean daude.
- Konparatzeko aukera eta ebaluazioa. Garatuko dugun sentimenduen lexikoiak beste sentimenduen lexikoien ezaugarriak baditu, elkarren artean konparatzeko aukera dago eta horrek lexikoia ebaluatzeko aukera ere ematen digu.

Ondoren, sentimenduen lexikoia garatzeko baliabideak eta tresnak bildu ditugu. Guztira bost baliabide edo tresna erabili ditugu: i) SO-CAL tresnaren gaztelaniako bertsioko lexikoia (Taboada *et al.*, 2011), ii) sarean dauden *Elhuyar* (Elhuyar, 2013) hiztegi eleanitza eta *Zehazki* (Sarasola, 2005) hiztegi

elebiduna, iii) SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko bertsioaren lexikoia, iv) Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) eta v) KWIC izeneko teknika:

- SO-CAL tresnaren gaztelaniako bertsioko lexikoia. Lexikoi hau izan da lanaren abiapuntua, itzuli den lexikoia baita. Guztira 4,880 sarrera ditu lexikoi honek, bost gramatika-kategorietan banatuta: izenak, adjektiboak, aditzak, adberbioak eta intentsifikatzaileak. 3.8 irudian, izenen zerrenda ikus daiteke eta bertako hitzek −5 eta +5 arteko sentimendu-balentzia dute.

Hitza	**Sentimendu-balentzia**
normalidad	2
imprudencia	-3
idiota	-3
envidia	-1
necesidad	-1
xenofobia	-4
hermosura	4
cumplimiento	3
violencia	-5
favorito	3

3.8 taula: SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko bertsioaren lexikoiaren zati bat.

- *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak. Gaztelaniazko SO-CAL lexikoia itzultzeko sarean dauden bi hiztegiak erabili ditugu: Elhuyar[5] eta Zehazki[6].

- Ingelesezko SO-CAL bertsioko lexikoia (Taboada *et al.*, 2011). Ingelesezko bertsioko lexikoiak 6.610 hitz ditu eta bost gramatika-kategorietakoak dira: izenak, adjektiboak, aditzak, adberbioak eta intentsifikatzaileak.

[5]Elhuyar hiztegiaren esteka: `https://hiztegiak.elhuyar.eus/eu_en`.
[6]Zehazki hiztegiaren esteka: `http://www.ehu.eus/ehg/zehazki/`.

3.9 irudian, lexikoiaren zati bat ikus daiteke. Euskarazko lehen bertsioko sarrerak hiztegi honetan agertzen diren ikusi dugu eta baita zer sentimendu-balentzia duten ere. Ondoren erabaki dugu hitzari zuen sentimenduaren balentzia mantendu edo ingelesekoa esleitu.

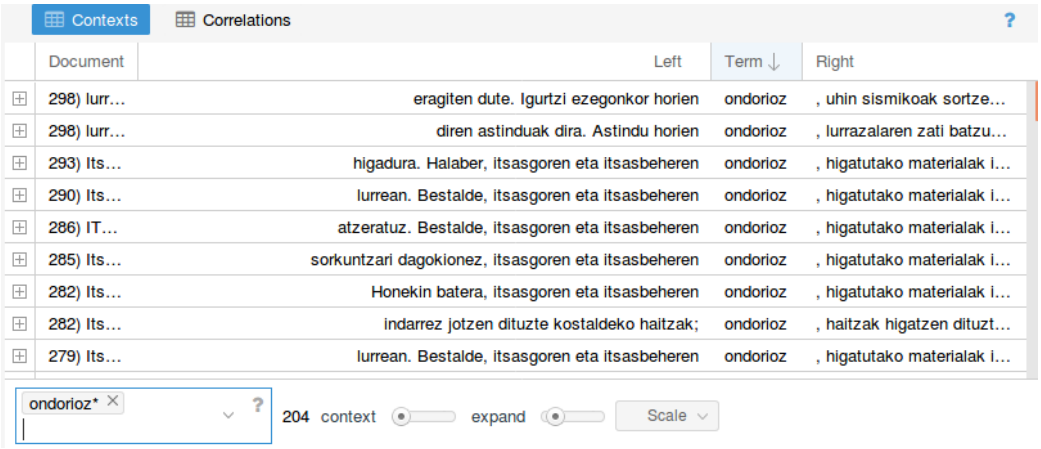
Izena	**Sentimendu-balentzia**
perfection	5
beauty	4
heroism	3
compassion	2
compromise	1
accommodation	-1
complication	-2
brutality	-3
atrocity	-4
monstrosity	-5

3.9 taula: SO-CAL tresnaren ingelesezko bertsioaren lexikoiaren zati bat.

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016). Itzulpen prozesuan erabili den beste baliabideetako bat izan da. Corpusak 240 iritzi-testu biltzen ditu sei domeinutakoak: kirola, eguraldia, politika, musika, zinema eta liburuak. Corpusa lexikoiko hitzen sentimenduen balentzia domeinuetara egokitzeko erabili da.

- KWIC teknika. Gaztelaniatik euskaratua izan den hitza corpusean bilatzeko eta hitzaren testuingurua ezagutzeko erabili da (ikus, 3.5 irudia). Modu horretan, gaztelaniatik euskarara itzulitako hitzari domeinuari loturiko sentimenduaren balentzia esleitu diogu.

Baliabideak eta tresnak bildu ostean, sentimenduen lexikoia itzuli eta sortu dugu. Itzulpen-prozesuan hiru urrats nagusi daude: i) gaztelaniazko lexikoia euskarara itzultzea, ii) euskarazko lexikoiaren garbiketa eta iii) euskarazko lexikoiaren ebaluazioa.



3.5 irudia: *Ondorioz* hitza eta bere testuinguruak.

Fenomenoa	GAZT	GAZT multzoka	EUS	ING	Azken balioa
F1	desacreditar	desacreditar −2	ospea_kendu −2 izena_kendu −2 sona_kendu −2	-	-
F2	atrofiar	atrofiar −1	atrofiatu −1	-	-
F3	amago	amago −1 cicatriz −2	seinale −1	-	-
F4	franquismo	franquismo −2	frankismo −2	-	−2
F5	correcto	acertado +3 correcto +3 decente +2	zuzen +3	right +1 correct +3	+3

3.10 taula: Itzulpen-prozesuaren azalpeneko adibideak.

Itzulpen-prozesuaren azalpena 3.10 taulako adibideetan oinarrituz egin dugu[7].

1- Itzulpena. Urrats nagusi honetan, SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskarara itzuli dugu.

   i) Itzulpen automatikoa gaztelaniatik euskarara. Lehenik eta behin, gaztelaniazko lexikoia automatikoki euskarara itzuli dugu *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak erabiliz. Gaztelaniazko hitz batek euskarazko ordain bat baino gehiago dituenean, ordain horietako bakoitzak hartu dugu aintzat, 3.10 tau-

[7]Euskarazko sentimenduen lexikoiaren sorkuntzari buruzko informazio osagarria 4.2.2 atalean dago.

lan ikus daitekeen moduan. Adibidez, F1 fenomenoan, *desacreditar* hitzaren euskarazko hiru ordain aintzat hartu ditugu: *ospea_kendu*, *izena_kendu* eta *sona_kendu*. Egoera horretan, gaztelaniazko hitzak duen balentzia; bere itzulpen-moduetako bakoitzak oinordekotzan hartu du.

ii) Iragazketa eta multzokatzea. Urrats honetan, euskaraz lortu diren ordainak iragazi eta euskarazko ordain berdinak multzokatu egin ditugu. Hau da, gaztelaniazko lexikoiko hitzei euskarazko ordaina ematean, errepikatuta dauden euskarazko ordainak batera jarri ditugu eta, beren ondoan, jatorrizko gaztelanizko hitzak eta gaztelaniazko hitz horien sentimendu-balentziak bildu ditugu.

Horren adibide dugu, 3.10 taulan, hirugarren zutabean, F3 fenomenoan, *seinale* hitza. *Seinale* ordainaren jatorrian gaztelaniako bi hitz daude: *amago* eta *cicatriz* eta horiek, euskarazko ordain bera dutenez (ordaina errepikatuta dagoenez), multzo berean jarri ditugu, baita bere jatorrizko gaztelaniazko hitzak eta gaztelaniazko hitzen sentimendu-balentziak ere ere. F5 fenomenoko *zuzen* itzulpenaren jatorrian, berriz, gaztelaniako *acertado*, *correcto* eta *decente* daude eta euskarazko ordainak eta beren jatorrizko gaztelaniazko hitzak bildu ditugu.

iii) *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrerak diren ordainak bakarrik aukeratu. Ordainak bananbanan aztertu ditugu eta aipaturiko bi hiztegietako sarrerak diren aztertu dugu. Lexikoia itzuli eta sortzerakoan, hiztegiko sarrera diren ordainak bakarrik hartuko ditugu aintzakotzat.

Esaterako, 3.10 taulan, F1 fenomenoan, *sona_kendu*, *izena_kendu* eta *ospea_kendu* ez dira hiztegi horietako sarrerak eta horrenbestez, ez ditugu aintzat hartu. *Atrofiatu*, aldiz, hiztegiko sarrera da eta kontuan hartu dugu.

iv) Sentimenduen balentziaren aukeraketa. Hiztegi bateko sarrerak ez diren ordainak kendu ostean, geratzen diren ordainetan, euskarazko ordainari dagokion gaztelaniazko hitza eta bere sentimendu-balentzia aukeratu dizkiogu.

Euskarazko ordainari dagokion sentimenduen balentzia eta gaztelaniako esanahia aukeratzerakoan, prozedura honi jarraitu diogu:

- Euskarazko ordainak itzulpen (eta balentzia) bakarra badu jatorrian gaztelaniaz, itzulpena eta hari dagokion balentzia hautatu dugu. Hori gertatu da 3.10 taulako F2 eta F4an.
- Euskarazko ordainak jatorrian hainbat gaztelaniazko hitz eta haien balentziak baditu, hitzari esleituko zaion sentimenduen balentzia (eta gaztelaniazko esanahia) Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarrituta egin dugu. Hori izan da 3.10 taulako F3 eta F5en kasua.
- Kasuak egon dira non lortu den ordaina ez den corpusean agertzen eta jatorrian gaztelaniazko hainbat dituen. Egoera horretan, ordainak duen esanahirik erabiliena edo ohikoena hobetsi da.

2- Garbiketa. Bigarren urrats honetan, sentimenduen lexikoiaren lehen bertsioa garbitu egin dugu eta bigarren bertsioa sortu dugu. Zehazki esanda, lexikoia domeinu jakin batzuetara moldatu dugu eta SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoia baliatuta aberastu eta zuzenketak egin ditugu.

v) Domeinu- eta corpus-moldaketa. Bigarren bertsioa sortzeko helburuz, lexikoiko hitzek Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) duten agerpena kontuan hartu dugu.

Adibidez, 3.10 taulan, F2 fenomenoan, *atrofiatu* Osasungintzako kontzeptu bat da eta domeinu hori ez dagoenez gure corpusean, kontzeptu hori lexikoitik ezabatu dugu.

vi) Lexikoiko sarrera bakoitza berraztertzea eta hobetzea. Lexikoiko euskarazko hitz bakoitzaren ingeleseko ordaina aurkitu dugu *Elhuyar* (Elhuyar, 2013) hiztegia erabilita eta, ondoren, hitz hori SO-CALen ingelesezko bertsioko lexikoian agertzen den aztertu dugu. Euskarazko sarrera ingelesezko lexikoian agertzen bada, euskarazko sarrerari ingelesezko sarreraren sentimendu-balentzia esleitu diogu. Euskarazko sarrera ingelesezko bertsioko lexikoian

ez bada agertzen, aldiz, sarrera hori kendu edo mantendu egin dugu hitzaren egokitasuna aintzat hartuz.

Esaterako, 3.10 taulan, F5 fenomenoan, euskarazko *zuzen* hitzaren baliokideak SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko lexikoian *correct* eta *right* dira eta sentimendu-balentzia ezberdinak dituzte.

3- Lexikoiaren ebaluazioa[8]. Azken urratsean, garatutako euskarazko sentimenduen lexikoia ebaluatu dugu. Horretarako, lehendabizi, bi pertsonen anotaziotik urre-patroi bat sortu da eta ataza batean urre-patroiak eta lexikoiak ematen dituzten emaitzak alderatu ditugu.

vii) Corpuseko maiztasunik handieneko hitzen urre-patroiaren anotazioa.

Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) maiztasunik handieneko 400 hitz erauzi ditugu (gramatika-kategoria bakoitzeko 100 hitz), *Analhitza* tresna (Otegi *et al.*, 2017) erabiliz, eta bi pertsonek hitz horiei −5 eta +5 arteko sentimenduen balentzia esleitu diete.

3.11 taulan ikusten den moduan, anotatzaile batek *ahul* hitzari −3 sentimendu balentzia esleitu dio eta *argi* hitzari +5. Ondoren, bien anotazioan oinarrituz, urre-patroia eratu ditugu.

viii) Urre-patroian erabilitako 400 hitzei euskarazko sentimenduen lexikoiaren bigarren bertsioa aplikatu zaie.

ix) Urre-patroiak eta lexikoiak emandako sentimenduen balentzia esleipenak alderatu ditugu Pearson korrelazioa (Benesty *et al.*, 2009) erabiliz. Pearson korrelazioa erabiliz, adostasun neurketa bi modutan egin da:

* Pearson 1. Neurketa honetan, 400 hitzeko zerrendan, bi pertsonek etiketatutako hitzak bakarrik hartu ditugu kontuan. Hitz bat pertsona bakar batek etiketatuta badago, hitz hori ez da kontuan hartu.

[8]Euskarazko sentimendu lexikoiaren ebaluazioari buruzko informazio gehiago 4.2.3 atalean dago.

Zenbakia	Adjektiboa	Orientazio semantikoa	Sentimendu-Balentzia
1	ageri		
2	ahul	Negatiboa	3
3	antisozial	Negatiboa	5
4	apur	Negatiboa	1
5	ar		
6	argi	Positiboa	3
7	aspergarri	Negatiboa	3
8	ausart	Positiboa	4
9	bakar	Positiboa	5
10	bakoitz		

3.11 taula: 400 hitzeko zerrendaren zati bat etiketatzaile batek bertako hitzei sentimendu-balentzia esleituta.

* Pearson 2. Neurketa honetan, zerrendako 400 hitzak erabili ditugu. Hitzen bat pertsona batek edo bik etiketatu gabe baldin badago 0 balentzia esleitu zaio, hitz hori kontuan hartu ahal izateko.

3.1.3. Lexikoian oinarritutako dokumentu mailako euskarazko sentimenduen sailkatzailea

Lan honen hirugarren urrats nagusia lexikoian oinarritutako dokumentu mailako euskarazko sentimenduen sailkapena garatzea izan da. Helburu horrekin, SO-CAL tresnaren[9] (Taboada *et al.*, 2011) euskarazko lehen bertsioa garatu nahi izan dugu, euskarazko sentimenduen sailkapena tresna horretan oinarritu baita. Hori dela eta, *Eustagger* tresna (Aduriz *et al.*, 2003) (3.1.3.1 atala) eta *SentiTegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) (3.1.3.2 atala) erabili ditugu euskarazko bertsioa sortzeko.

[9]SO-CAL tresnaren ingelesezko bertsioaren ezaugarriak 4.3.1 atalean deskribatuta daude.

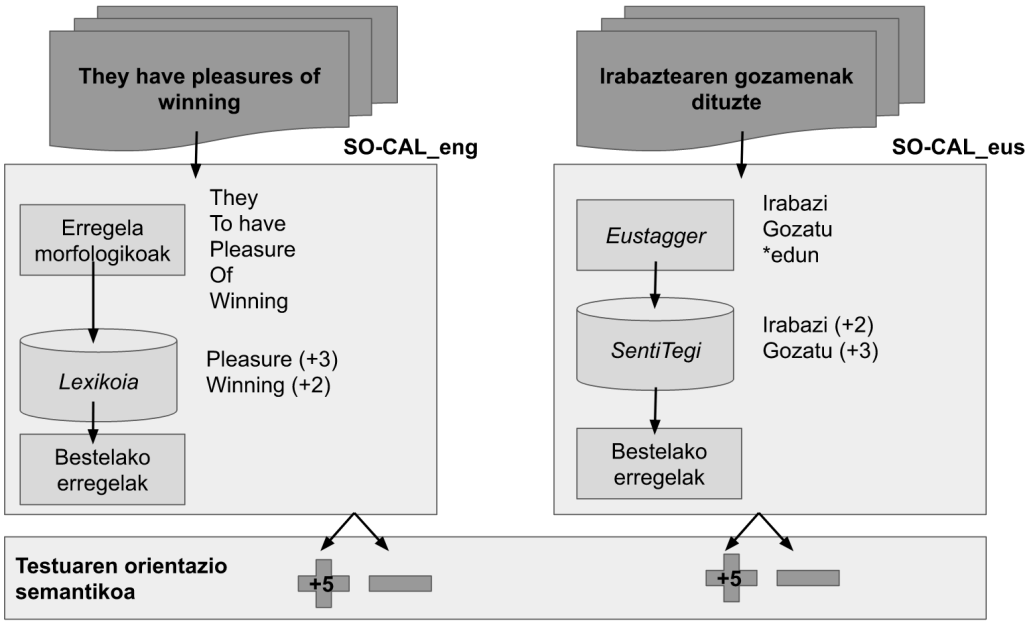
3.1.3.1. *Eustagger* tresna integratzea

Lexikoian oinarritutako sentimenduen sailkatzailea den SO-CALen (Taboada *et al.*, 2011) ingelesezko bertsioan, lematizatutako sentimendu-balentziadun hitzen lexikoia eta flexioari lotutako zenbait erregela daude. Baina euskaran, hori baliagarria izan arren, ez da modurik egokiena lexikoian oinarritutako sentimenduen sailkatzaile bat garatzeko; izan ere, tresnak testua prozesatzekotan flexioari lotutako erregela asko eduki beharko lituzke eta hori ez da bideragarria lan konplexua delako. Beste aukera bat flexioak eragindako hitz baten aldaera guztiak lexikoian sartzea izango litzateke baina, kasu horretan, lana konplexuagoa izango litzateke eta denbora gehiago eskatuko luke. Hori da euskarazko sentimenduen sailkatzailea sortzerakoan izan dugun zailtasuna.

Zailtasun horri aurre egiteko konponbidea, lexikoiak testuko hitzei sentimenduen balentzia esleitu aurretik, testuak lematizatzeko *Eustagger* tresna (Alegria *et al.*, 2002) integratzea erabaki dugu. Modu horretan, lehenik eta behin, tresnak testua lematizatuko du eta ondoren, tresnak lematizatutako hitz hori tresnaren lexikoietan badagoen aztertuko du eta baldin badago, testuko hitzari sentimenduen balentzia esleituko dio. Beraz, lematizatzaile bat tresnan integratuta, tresna gauza izango da lehenik eta behin, testua lematizatzeko; ondoren, lematizatutako hitza lexikoian dagoen aztertzeko eta, azkenik, baldin eta hitza lexikoian badago, hitzari sentimenduen balentzia esleitzeko.

Egindako aldaketa eta ingelesezko eta euskarazko SO-CAL tresnaren egituren alderaketa ageri dira, 3.6 irudian.

Ikusten den moduan, ingelesezkoan, erregela morfologikoak daude eta, modu horretan, testuetako hitzak lema bilakatzen dituzte. Euskarazkoan, aldiz, aberastasun morfologikoa tarteko, *Eustagger* lematizatzailea (Alegria *et al.*, 2002) integratu da helburu bera lortzeko asmoz. Ondoren, sentimenduen lexikoiak datoz eta lematizatutako hitzak lexikoian baldin badaude, tresnak hitz horiei sentimendu-balentzia esleitzen die. Azkenik, beste zenbait erregela daude ingeleserako, euskararako ere berak direnak (orientazio semantiko negatibodun hitzei pisu gehiago esleitzea, galde-perpausetako hitzei sentimendu balentzia ez esleitzea, etab. egiten dituztenak) eta bi hizkuntzetarako baliagarriak dira.



3.6 irudia: SO-CAL tresnaren ingelesezko eta euskarazko bertsioen egituren alderaketa.

3.1.3.2. *Sentitegi* sentimenduen lexikoia integratzea

Iritzi-testuak lematizatzeko *Eustagger* tresna (Alegria *et al.*, 2002) integratu ondoren, euskarazko sentimenduen lexikoia gauza da testuak prozesatzeko; ez, ordea, hitzei sentimendu-balentzia esleitzeko. Horretarako, tesi-lan honetan garatutako metodologiaren ondorioz sortutako *Sentitegi* euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018) integratu dugu tresna horretan.

Tresnak berak modulu bat du sentimenduen lexikoiei bideratuta; hala ingelesezko lexikoiak kendu eta euskarazkoak gehitu ditugu. Tresnak funtziona dezan, beharrezkoa da lexikoiak *txt* formatuan egotea eta bertan, lerro bakoitzeko sarrera bat eta bere sentimendu balentzia egotea. Halaber, gramatika-kategoria bakoitzak bere fitxategia eduki behar du. Izan ere, *Eustagger* lematizatzaileak (Alegria *et al.*, 2002) testuko hitza lematizatu eta bere gramatika-kategoria identifikatuko du eta ondoren, tresnak hitz hori bere gramatika-kategoriako lexikoian bilatuko du. Guk lau gramatika-kategorietako lexikoiak integratu ditugu: izenena, adjektiboena, adberbioena

eta aditzena. Egin dugun aldaketa 3.12 taulan ikus daiteke.

Ingelesa		**Euskara**	
Thriving	+3	Bikain	+5
Record-setting	+3	Maximo	+1
Leading	+3	Orokor	+3
Industrious	+3	Min	-2
Best-selling	+3	Polit	+4
Upset	+3	Gogor	-1
Clean	+2	Ahul	-2
Capacious	+2	Behar	-1
Cogent	+2	Txar	-3
Confident	+2	Zail	-2

3.12 taula: Ingelesezko eta euskarazko sentimenduen lexikoien zati bat.

Ezkerrean, ingelesezko lexikoia ageri da, hitz-zerrenda batekin eta beren sentimenduen balentziekin. Eskuinean, berriz, euskarazko lexikoiaren zati bat dago eta horrekin ordezkatu dugu ingelesezkoa. Bi lexikoiak *txt* formatuan dauden fitxategian daude eta *SentiTegi* lexikoia integratzeko, fitxategi bat bestearekin ordezkatu dugu.

3.1.3.3. Sentimenduen sailkatzailearen ebaluazioa

Sentimenduen sailkatzailean *Eustagger* lematizatzailea (Alegria *et al.*, 2002) eta *Sentitegi* sentimendu lexikoia (Alkorta *et al.*, 2018) integratu ondoren, sailkatzailea bera ebaluatu dugu[10].

Tresna ebaluatzeko Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 48 iritzi-testu erabili ditugu, sentimenduen analisiarekin lotutako azterketa eta lanketa ez delako iritzi-testu horietan oinarrituta egin. Metodologia atal honetan aipatu bezala, corpuseko iritzi-testuen orientazio semantikoa etiketatuta dago eta sentimenduen sailkatzaileak ematen dituen emaitzak horiekin konparatu ditugu, 3.13 taulan ikusten den moduan.

[10]Euskarazko sentimenduen sailkatzailearen ebaluazioari buruzko informazio osagarria 4.3.3 atalean dago.

Iritzi-testua	**Orientazio semantikoa**	**Sailkatzaileak esleitutako sentimendu-balentzia**
EGU32	Positiboa	0,80 (Positiboa)
LIB30	Positiboa	1,09 (Positiboa)
ZIN39	Negatiboa	0,57 (Positiboa)

3.13 taula: Euskarazko sentimenduen sailkatzailearen ebaluazioa.

Tresnak emaitzak zenbakiz ematen ditu eta ebaluatzeko orduan, zenbakizko emaitza horretan zeinua hartu dugu kontuan, hau da, emaitza horiek zeinu positiboa edo negatiboa duten.

3.2 Balentzia-aldatzaileen identifikazioa

Ikerketa aurrera eramateko oinarrizko balibideak eta tresnak garatu ondoren, euskaran dauden testuinguruko balentzia-aldatzaileak identifikatzen hasi gara. Urratsez urrats, hizkuntza mailetako balentzia-aldatzaileak zein diren eta nolako eragina duten aztertu dugu.

3.2.1. Fonologia eta morfologia: bustidura adierazkorra eta hizkiak

Lehenik eta behin, fonologia eta morfologia mailako balentzia-aldatzaileak identifikatzen hasi gara. Helburu horrekin, lehen urratsean, euskararen morfologia eta fonologia lantzen duten hainbat iturri bibliografikoren bilaketa egin dugu eta euskarak dituen hizkien (aurrizkiak, artizkiak eta atzizkiak) zerrenda bat osatu dugu. Zerrenda horretan, hizki bakoitza zer gramatika-kategorirekin agertzen den, bere esanahi semantikoa zein den eta adibide bat jarri dugu. 3.14 taulako zerrenda[11] osatzeko erabili ditugun iturri bibliografikoen artean Mujika (1982), Euskara Institutua (2011) eta Oñederra (1990) aipa ditzakegu.

Hizkia	**Izena**	**Adjektiboa**	**Aditza**	**Sailkapen semantikoa**	**Adibidea**
-zale	X			Zaletasuna	Ardozale
ez-	X	X	X	Ezezkotasuna	Ezberdin
-z- → -x-				Hurbiltasuna/ txikitasuna	Gazte → gaxte zahar → xahar

3.14 taula: Hizkiei buruz bildutako informazioa.

Iturri bibliografikoetatik bildutako hizki eta bustidura adierazkorren adibideak ageri dira 3.14 taulan. Lehenengo hizkia *-zale* atzizkia da, izenekin agertzen da eta zaletasuna adierazten du. Bigarrena, berriz, aurrizki bat da (*ez-*); izen, adjektibo nahiz aditzekin ager daiteke eta ezezkotasuna adierazten du. Bukatzeko, azken adibidean bustidura adierazkor[12] bat dago: [z] bilakatzea

[11] 3.14 taulako zerrenda lanaren zati bat da.

[12] Bustidura adierazkorra erabilera ez-estandarrean ohikoena den arren, bere zenbait kasu aurkitu ditugu corpuseko testu formaletan.

[x]. Bustidura adierazkor guztiak bezala, ez dago murriztapen gramatikalik eta hurbiltasuna edo txikitasuna adierazteko erabiltzen da. Prozedura bera jarraitu dugu beste hizki guztiekin.

Hurrengo urratsean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 192 iritzi-testu hartu ditugu eta *Analhitza* tresna (Otegi *et al.*, 2017) bidez aztertu ditugu. Tresna pasa ondoren, testuetako hitz guztiak maiztasunaren arabera zerrendatzea lortu ditugu. 3.15 Taulan, *Analhitza* tresnak (Otegi *et al.*, 2017) emandako hitz-zerrenda ageri da eta maiztasunean oinarritzen da. Ikusten denez, *-txo* hizkia daraman *zertxobait* hitza hamaika aldiz agertzen da corpusean.

Hitza	**Instantzia-kopurua**
zati	11
zertxobait	11
zerua	9

3.15 taula: Corpuseko hitzen zerrendaren zati bat maiztasunean oinarrituta.

Ondoren, euskarak dituen hizkiak eta bustidura adierazkorra egiteko moduak corpusean nola agertzen diren aztertu ditugu. Horretarako atal honetako lehen urratsean sortu dugun zerrenda erabili dugu eta, hori oinarritzat hartuz, corpusean agertzen diren hizki eta bustidura adierazkorra egiteko modu guztiekin zerrenda bat osatu dugu. Zerrendan, banan-banan, beren ezaugarriak deskribatu ditugu 3.16 taulan ikusten den moduan.

Adb.	**Hitza**	**Hizkia**	**Balentzia**	**Balentzian eragina**	**Semantika**	**Iz.**	**Adj.**	**Adi.**
(1)	istorio							
(2)	erasotxo	-txo	−2	Ahuldu	Txikitasuna	X	X	
(3)	kuttun	[t] → [tt]	+2	Indartu	Hurbiltasuna			

3.16 taula: Corpusean agertutako hizkien eta bustidura adierazkoren azterketaren lagin bat

3.16 taulan, (1) adibidean, bustidura adierazkorrik edo hizkirik ez da agertzen. (2) adibidean, *-txo* atzizkia ageri da eta honek *eraso* hitzaren sentimendu balentzia (−2) ahultzen du. (3) adibidean, azkenik, bustidura adierazkorra

dago [t] [tt] bilakatu baita eta aldaketa horrek adierazkortasuna indartu duenez, duen sentimenduaren balentzia are indartsuagoa bilakatzen da. Guztira 59 atzizki, 7 aurrizki eta 13 bustidura adierazkorraren instantzia aurkitu ditugu Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) eta horiekin guztiekin prozedura bera jarraitu dugu.

3.2.2. Sintaxi maila: ezeztapen-markak

Hitzen balentzian eragin dezakeen hizkuntzaren beste alderdi bat sintaxia da. Baina kasu honetan, balentzia aldatzearen ondorioa jasoko duena ez da hitz bat baizik eta hitz multzo batek osaturiko sintagma edo esaldi bat izan daiteke. Hau da, morfologiako eta fonologiako balentzia-aldatzaileek hitzari eragiten dioten moduan, sintaxiko balentzia-aldatzaileek hitzi ez ezik, sintagma edota esaldiari ere eragin diezaiokete.

Sintaxian, ezeztapen-markek esaldien sentimendu balentzia nola aldatzen duten aztertzeko asmoz, lehenik eta behin, euskararen ezeztapen-marken zerrenda osatu dugu eta zerrenda hori Altuna *et al.*en (2017) eta Altuna *et al.*en (1985) lanetan oinarrituta osatu dugu. Honako hauek izan dira ikerketarako erabili ditugun ezeztapen-markak: *ez*, *ezin*, *gabe*, *ezik*, *salbu*, *ezta* eta *ezean*.

Hurrengo urratsean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 96 testutan, ezeztapen-markaren bat duten esaldiak bilatu ditugu eta horiekin esaldi-zerrenda bat osatu dugu, 3.17 taulan ikus daitekeen moduan. Ez ditugu testu guztiak erabili, Altuna *et al.*en (2017) lanean ageri diren ezeztapen-marka gehienak lortu ditugulako eta testu gehiago aztertu arren, ezeztapen-marka berririk ez delako agertu. Guztira 359 ezeztapen-marken instantzia lortu dugu. *Analhitza* tresnak (Otegi *et al.*, 2017) emandako datuen arabera, 359 ezeztapen-marka horiek 320 esalditan banatuta daude. Horrenbestez, badaude esaldiak non ezeztapen-marka bat baino gehiago dauden. Bestalde, lortu dugun ezeztapen-marken azpicorpus honek 5.515 hitz ditu.

Ondoren, 320 esaldi horietako hitzei nahiz esaldi osoei sentimendu balentzia esleitu diegu. Hori egiteko, hitzei kategoria gramatikala edota marka lexikala esleitzen dien *Eustagger* tresnari (Aduriz *et al.*, 2003) *Sentitegi* (Alkorta *et al.*, 2018) sentimenduen hiztegiko hitzak gehitu dizkiogu; lexikoian hitzak

TESTUA	ITEMA
EGU23	Ez da espero euri asko egitea, baina giroa hezea eta freskoa izango da.
EGU23	Asko ez du egingo, baina zerbait bai.
EGU24	Zumaian ez du elurrik egingo baina elur-kota 400-500 metrora jaitsiko da, beraz kontuz errepideetan.
EGU24	Oraintxe bertan ematen du Frantzia iparraldera iritsiko dela eta hortik haize indartsua sortuko du gure lurraldean, baina ez da izango lehengokoa bezain indartsua.
EGU25	Abuztua hasi berri dugun honetan, ez dugu urte sasoi honetarako espero izan ohi dugun eguraldirik izango.
EGU25	Goizean berriz aterri mantenduko du eta ez du ia euririk egingo.

3.17 taula: Ezeztapen-markak aztertzeko sortutako azpicorpusaren zati bat.

kategoria gramatikalean oinarrituta banatuta baitaude.

(32) Pogostkinak ezin [hobeki]$_{+2}$ atera zituen. (MUS20)

(33) [Irabazi]$_{+2}$ ezinik jarraitzen du Eibarrek. (KIR17)

(34) Ikuspuntu [politikotik]$_{-1}$ ez ezik, [ekonomikotik]$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(32), (33) eta (34) adibideetan, etiketatzaileak eta hari gehitutako sentimendu lexikoiak lau hitzei esleitu die sentimendu balentzia: *hobeki* adjektiboari, *irabazi* aditzari, *politiko* adjektiboari eta *ekonomiko* adjektiboari. Lehen bi hitzei +2 balentzia esleitu die eta adjektiboei −1 eta −3, hurrenez hurren. Modu honetan, ezeztapen-markak eta sentimendu balentziadun hitzak edukita, analisi linguistikoa egiteko ingurunea prest utzi dugu. Eta horrenbestez, analisi linguistikoa egitea da hurrengo urratsa.

Ezeztapen-marken analisi linguistikoa egiterakoan, aintzat hartu duguna izan

da ezeztapen-markak eragiten duen aldaketa orientazio semantikoan eta zehazki, sentimendu-balentzian. Hots, ezeztapen-markak sentimendu-balentziadun hitz eta hitz-multzoetan, haren irismenean eta ondorioz, baita esaldiko sentimendu-balentzian ere, zer ondorio uzten duen aztertu nahi izan dugu. Azterketa eskuz egin dugu eta jarraian azaltzen den prozedura jarraitu dugu.

Azterketan, corpuseko esaldiak banan-banan aztertu ditugu eskuz. Esaldietako ezeztapen-markak identifikatu ondoren, haren irismena zein den adierazi dugu. Azkenik, irismen-eremuko sentimendu-balentzian nahiz esaldi osokoan utzi duen ondorioa aintzat hartuta, esaldia bera eta haren ezeztapen-marka multzokatu egin ditugu.

Azter ditzagun, berriz, aurretik aipaturiko hiru adibideak.

(35) Pogostkinak [ezin hobeki$_{+2}$] atera zituen. (MUS20)

(36) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)

(37) [Ikuspuntu politikotik$_{-1}$ ez ezik], [ekonomikotik$_{+3}$] ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(35) adibidearen kasuan, ezeztapen-marka *ezin* da eta bere irismena *hobeki* adjektiboraino iristen da. Irismen-eremuan hitz bakarra dagoenez eta balentziaduna denez, irismen-eremuaren balentzia +2 da, baita esaldiarena ere. Kasu honetan, ezeztapen-markaren eragina aztertuta, ezeztapen-markak *hobeki* balentziadun hitza indartu egiten duela ohartu gara, intentsitatez *ezin hobeki hobeki* baino indartsuagoa baita. Ondorioz, esaldia bera eta ezeztapen-marka balentzia indartzen duten taldean sailkatu ditugu.

(36) adibidean, aldiz, nahiz eta ezeztapen-marka bera den (*ezin*), ikusi dugu haren eragina ezberdina dela. Hau da, *ezin* ezeztapen-markaren irismena *irabazi* da, +2 balentzia duena eta honi indarra ahultzen dio. Hots, *irabazi ezinik irabazi* baino ahulagoa da sentimendu-balentziari dagokionez. Ondorioz, ezeztapen-marka eta esaldia bera beste multzo batean sartu ditugu, ahultze bat egon delako. Bi adibide hauekin ikusi den moduan, posible da ezeztapen-marka berak bi funtzio izatea.

Azkenik, (37) adibidean, *ez ezik* ezeztapen-markak −1 sentimendu-balentzia

duen *ikuspuntu politikotik* hitz multzoa ezeztatzen du, baina sentimendu-analisiaren ikuspegitik, hitz multzoaren sentimendu-balentzian ez da aldaketarik gertatzen. Izan ere, kasu honetan, ezeztapena ondotik informazioa gehitzeko erabiltzen da eta ez ezeztapenaren irismena den *ikuspuntu politikotik* ezeztatzeko. Beraz, *ez ezik* egiturak bere irismenaren parte diren hitzen sentimendu-balentzian ez du eragiten.

Erregeletan oinarritzen diren SO-CAL moduko (Taboada *et al.*, 2011) tresnetan ezeztapenari buruzko informazioa gehitu nahi bada, beharrezkoa da ezeztapen-markak eta haren irismena identifikatzea eta, horregatik, ezeztapen-markak eta beren irismena identifikatzeko erregelak sortu eta ebaluatu ditugu[13], 3.18 taulan ikus daitekeen moduan.

3.18 taulako, erregelak deskribatu aurretik, erregela horiek nola sortu diren azalduko dugu jarraian, (38) eta (39) adibideen bitartez. Bertan bosgarren erregelako bi aldaerak landuko ditugu. (38) adibidean, *ezin* ezeztapen-markaren ondoren, aditz laguntzailea (*da*) eta aditz nagusia (*baztertu*) ageri dira eta, azkenik, izen-sintagma bat den subjektua (*ekaitz zaparradaren bat izatea*). Beraz, esaldi honetarako lortutako egitura honelakoa izango litzateke: *ezin* + aditz laguntzailea + aditz nagusia + IS. Dena kako artean dago eta horrek ezeztapen-markaren irismena zein den adierazten du.

(38) (...) [ezin da baztertu$_{-1}$ ekaitz zaparradaren bat izatea.] (EGU35)

(39) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek (...) (KIR22)

(39) adibidea aztertzen badugu, ikusiko dugu *ezin* ezeztapen-markaren aurretik *irabazi* aditz nagusia agertzen dela eta bera dela ezeztapen-markaren irismena. Corpusean ageri diren *ezin* ezeztapen-markaren beste 36 instantziekin prozedura bera jarraitu ondoren, 3.18 irudiko (1) eta (5) adibideetan azaltzen diren erregelak lortu ditugu.

> Bildu den informazioetako bat erregela bakoitzak duen egitura sintaktikoa da. 3.18 taulako erregeletako egiturek honelako ezaugarriak di-

[13]Ezeztapen-markak eta beren irismena identifikatzeko *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako eta ebaluatutako erregelei buruzko informazio gehiago 5.2.3 atalean dago.

Adb.	Ezeztapen-marka	Erregelen egitura
(1)	ezin	PM *ezin* + [adjektiboa/adberbioa] (+ atzizki konp.) PM
(5)	ezin	PM [(IS) + aditza + *ezin*] PM PM [*ezin* (+ ad. lag.) (+ IS) + aditza (+ IS)] PM
(10)	ez	PM [IS] *ez ezik* PM

3.18 taula: Sentimendu-balentzian eragin ezberdinak dituzten ezeztapenak identifikatzeko proposatu diren erregeletako batzuk.

tuzte: kakoek [] ezeztapen-markaren irismena adierazten dute, parentesiek () sintagma horren agerpena hautazkoa dela adierazten dute. *Letra etzanak* ezeztapen-marka edo egitura lexikalizatuen osagaiak adierazten dituzte eta barrak /, azkenik, ezeztapen-markaren irismenean, aukeran aipaturiko osagai ezberdinek daudela adierazten du. Erregeletan ageri diren beste elementuak hitz-multzo eta gramatika-kategoriei dagozkie. ISk eta ASk izen- eta aditz-sintagma adierazten dute, hurrenez hurren. Erregeletan ageri diren beste elementuak dira adjektiboa, adberbioa, aditza, aditz laguntzailea eta sintagma. Bukatzeko, bada beste elementu bat erregeletan azaltzen dena eta hori puntuazio-markaren murriztapena (PM) da. Murriztapen honen helburua erregeletan ezeztapen-marken irismena esaldi barnean kokatzea eta beste esaldietako elementuak ez hartzea da. Izan ere, irismena ezeztapen-markaren aurrean nahiz atzean ager liteke eta aurreko edo ondorengo esaldietako hitzak ezeztapen-markaren irismeneko hitz moduan tratatzeko arriskua ikusten dugunez, erregelei puntuazio-marka murriztapena gehitzea erabaki dugu. Egitura lexikalizatuen kasuan, ez dago horrelako murriztapenik. Egitura lexikalizatuek egitura egonkorra dutenez (errepikatzen diren hitz-multzoak direnez) ez da beharrezkoa. Egitura lexikalizatuaren adibide bat "baino ez". Bi hitzak elkarrekin uste baino maiztasun handiagoz azaltzen dira eta bi hitzen baturak berezko esanahi bat sortzen du, "bakarrik" hitzak duen esanahiaren parekoa dena.

Aurreko urratsetan analisi linguistikoan bildutako informazioa modu egituratuan antolatu ondoren (3.18 taulan ikusten den moduan), ezeztapen-markak eta beren irismena identifikatzeko erregelak Karlsson *et al.*en (1995) *Mu-*

rriztapen Gramatika hurbilpenean sortu eta ebaluatu ditugu. 3.7 irudian erregelak nolakoak diren ikus daiteke[14]. Erregela horiek honela osatuta daude:

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT PUNT_GALD PUNT_ESKL
    PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = ``ez'';
LIST BESTERIK = ``beste'';

# (2)
# besterik ez egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)    (1C EZ);

(...)
```

3.7 irudia: Ezeztapen-markak eta beren irismen-eremua identifikatzeko erregelen adibide bat *Murriztapen Gramatika* (Karlsson *et al.*, 1995) ingurunean.

- LIST zerrendak. Bertan hitz jakinak edo bestelako elementuak, puntuazio-markak adibidez, zerrendatzen dira. 3.7 irudian, puntuazio-markak (LIST_PUNTUAZIOA) eta ezeztapen-markak (LIST_EZ, LIST_BESTERIK) zerrendatu ditugu.

- MAP komandoa (islapen-erregelak). Hauen bidez erregeletan adierazitako egiturak corpusean bilatzen dira eta aurkitzen badira, etiketa bat esleitzen zaio corpusean bilatu den egiturari. Ezaugarri hauetazz osatuta dago erregela:

  - Esleituko den etiketa. Etiketak aurretik ! ikurra du, 3.7 irudian.
  - Etiketa esleituko zaion hitza. Adibidez, *!besterikezHAS* etiketa determinatzaile bati (DET) esleituko zaio.
  - Erregelaren baldintzak. Esaterako, *!besterikezHAS* etiketa esleitzeko, 0 posizioan BESTERIK zerrendak, hau da, *beste* hitzak, (0C BESTERIK) egon behar du eta 1 posizioan EZ zerrendak (kasu honetan, *ez* ezeztapen-markak), erregelak funtziona dezan.

```
"<,>""<PUNT_KOMA>"
        PUNT_KOMA
"<batere>"
          "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG %SINT
"<harrotu>"   S:278/0
          "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG
             %ADIKAT S:278 !gabeAUR
"<gabe>"   S:141/0
          "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> %SINT S:141 !
             gabe
"<$.>""<PUNT_PUNT>"
        PUNT_PUNT
```

3.8 irudia: *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelek esleitutako ! etiketa.

*Murriztapen Gramatika*n (Karlsson *et al.*, 1995) idatzirik erregela bakoitzak bere etiketa uzten du corpuseko lerroan. 3.8 irudian, esaterako, Murriztapen Gramatikan sortu ditugun erregelek bi hitzi etiketa ezarri diete (*!gabeAUR* etiketa *harrotu* hitzean eta *!gabe* etiketa *gabe* hitzean).

Guk sorturiko erregelek corpusean ! ikurra duen etiketak esleitzen dituztela begiztatu ondoren, erregela horiek ebaluatu ditugu.

Ebaluazioa egiteko, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 48 testu erabili ditugu. 48 testu horiek prozesatu egin behar izan ditugu *Murriztapen Gramatika*ren ingurunera (Karlsson *et al.*, 1995) ekartzeko eta horretarako *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) oinarritzen den euskarazko desanbiguatzaile morfosintaktikoa erabili dugu (Aduriz *et al.*, 1997).

*Murriztapen Gramatika*ren (Karlsson *et al.*, 1995) ingurunean sorturiko 75 erregelak corpusaren test zatiko 144 testuetan aplikatu ondoren, erregela horiek ebaluatu egin ditugu. Lehenik eta behin, bi personek ebaluazioa egiteko duten adostasuna neurtu dugu corpusaren zati txiki bat erabiliz. Adostasuna neurtzeko kappa neurria (Cohen, 1960) erabili da. Emaitzek erakusten dutenez, kappa neurria 0,60koa da. Landis eta Kochen (1977) arabera, *neurrizko adostasuna* da. Etiketatzaileen arteko adostasuna neurtu ondoren, etiketatzaile batek ebaluazio osoa egin du.

[14]Ezeztapen-markak eta beren irismena identifikatzeko sortutako erregela guztiak A eranskinean daude zerrendatuta.

Ebaluatzeko, erregelek corpusean esleitutako etiketak ebaluatu ditu etiketatzaileak. Etiketatzaileak hiru etiketa erabili ditu: ETIK_ONDO, erregelaren etiketa zuzena denean; ETIK_FALTA, corpuseko hitzak etiketa behar zuenean eta erregelak jarri ez dionean eta, azkenik, ETIK_GAIZKI, erregelak corpuseko hitzari etiketa jarri dionean eta hitzak etiketa behar ez duenean.

Ebaluazioaren adibide bat ikus daiteke, 3.9 irudian. Bertan hiru hitz daude eta hirurek egon beharko lukete etiketatuta. Hiruretatik bi (*harrotu* eta *gabe*) etiketatuta daude eta ETIK_ONDO etiketa jarri zaie, lerroen hasieretan. Baina, hori ez da hala *batere* hitzaren kasuan. Nahiz eta ezeztapen-markaren irismen-eremuko parte izan, *harrotu* aditzari eragiten diolako, ez dago etiketatuta. Ondorioz, hitz horri ETIK_FALTA etiketa jarri zaio, kasu honetan ere, lerroaren hasieran. Aipaturiko prozedura hau jarraitu da bi pertsonen arteko adostasuna nahiz corpusa ebaluatzerakoan.

```
``<,>''``<PUNT_KOMA>''
        PUNT_KOMA
ETIK_FALTA ``<batere>''
           ``batere'' ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG %SINT
ETIK_ONDO ``<harrotu>''    S:278/0
           ``harrotu'' ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG
              %ADIKAT S:278 !gabeAUR
ETIK_ONDO ``<gabe>''    S:141/0
           ``gabe'' ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> %SINT S:141 !
              gabe
``<$.>''``<PUNT_PUNT>''
        PUNT_PUNT
```

3.9 irudia: Erregelak ebaluatzeko etiketatzaileak hitzei esleitutako etiketak.

3.2.3. Diskurtsoa: erlazio diskurtso-egiturak, beren osagaiak eta unitate zentrala

Diskurtso mailan, RST jarraituz[15] (Mann eta Thompson, 1988), bi alderdi landu ditugu. Alde batetik, balentzia-aldaketak erlaziozko diskurtso-egituretan aztertu ditugu. Beste aldetik, unitate zentraletan, hots, testuko

[15] *Egitura Erretorikoaren Teoria*ri (RST) buruzko azalpenak 5.3.2 azpiatalean daude.

diskurtso-unitate garrantzitsuenetan, orientazio semantikoa duten hitzek zer gramatika-kategoritakoak diren landu dugu.

3.2.3.1. Erlaziozko diskurtso-egiturak

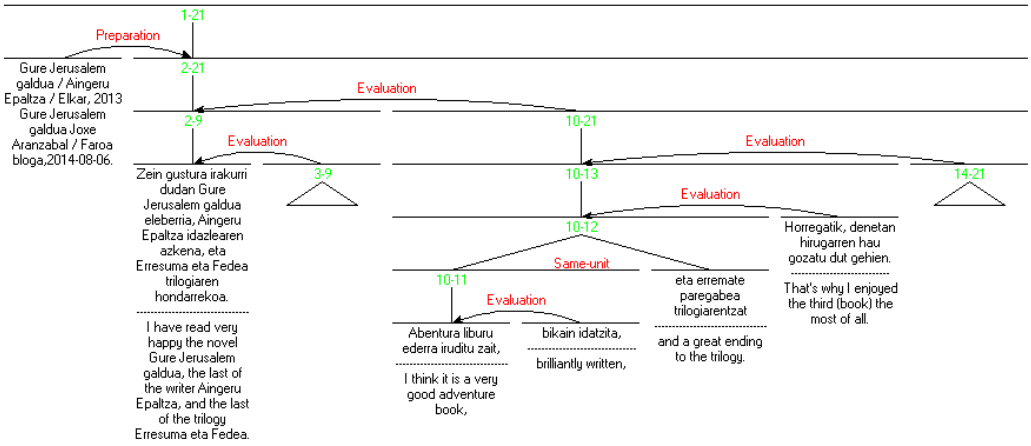
Erlazio erretorikoetan, lehenik eta behin, gure helburuak zeintzuk diren zehaztu ditugu. Guztira hiru helburu finkatu ditugu:

- Erlaziozko diskurtso-egituren eta bere osagaien arteko orientazio semantikoaren arteko adostasuna neurtu nahi dugu.
- Erlaziozko diskurtso-egiturek unitate zentraletik duten distantzia aintzat hartuz, erlazio erretorikoek eta testu osoak zer distantziatan duten orientazio semantikoaren adostasunik handiena neurtu nahi dugu.
- Iritzi-testuetan, esanahi erretoriko ezberdineko erlazioek testuetan beren kokagune zehatza baduten aztertu nahi dugu. Erlazio erretorikoen kokagunea unitate zentralarekiko distantzian oinarrituz finkatuko dugu.

Erlaziozko diskurtso-egituretan ditugun helburuak finkatu ostean, zenbait urrats bete ditugu. Jarraian, metodologia urratsez urrats deskribatzen da:

1- Diskurtso mailako osagai ezberdinei orientazio semantikoa esleitu diegu.
   - 240 iritzi-testuri.
   - Erlaziozko diskurtso-egitura hauei (29 iritzi-testuetatik hartutakoak): EBALUAZIOA/INTERPRETAZIOA (140), KAUSA taldea (71), KONTZETSIOA/ANTITESIA (70), EBIDENTZIA/-JUSTIFIKAZIOA (53), KONTRASTEA (26), BALDINTZA taldea (18), AHALBIDERATZEA/MOTIBAZIOA (6). Guztira, 384 instantziei esleitu zaie orientazio semantikoa.
   - Aurreko erlazio erretorikoen osagaiei: nukleoa eta sateliteari edota erlazioko lehen eta azken osagaiari esleitu zaie orientazio semantikoa.

2- Erlaziozko diskurtso-egituretako parametroen analisia egin dugu. Literaturaren domeinuko 28 testuen erlazio erretorikoak datu-base batean sartu ostean, jarraian zerrendatuta dauden parametroak aztertu ditugu. 3.10 irudiko EBALUAZIOA erlaziozko diskurtso-egitura (EDU 10-12) adibide moduan erabiliko dugu parametroak esplikatzeko:



3.10 irudia: SENTFAR-01 testuaren RST-zuhaitza.

- Nuklearitatea. Erlaziozko diskurtso-egitura nukleo bakarrekoa edo nukleo anitzekoa izan daiteke. 3.10 irudian, EBALUAZIOA erlaziozko diskurtso-egitura N(ukleo)-S(atelitea) motakoa dela zehaztu dugu.
- Erlaziozko diskurtso-egituren eta EDUen orientazio semantikoa. Erlaziozko diskurtso-egiturari nahiz EDUei hiru motako orientazio semantikoa esleitu diegu: positiboa, negatiboa eta neutrala. Lehenik eta behin, EDUei esleitu diegu orientazio semantikoa eta ondoren, erlazio erretoriko osoari. 3.10 irudiko EBALUAZIOA erlazioan (10-11 eta 12), bi EDUei eta erlazio osoari orientazio semantiko positiboa esleitu diegu.
- Unitate zentralarekiko distantzia. Erlaziozko diskurtso-egituren eta unitate zentralaren arteko distantzia bien artean dauden erlaziozko diskurtso-egituren kopurua zenbatuz neurtu dugu. Aztertzen

ari garen EBALUAZIOA erlazioren kasuan (EDU 10-12), distantzia +2 dela adierazi dugu, unitate zentrala 3.10 irudian ezkerretik hasita bigarren EDUa baita eta tartean bi erlaziozko diskurtso-egitura baitaude.

– Erlaziozko diskurtso-egituren esanahi erretorikoa. Erlazioen esanahi erretorikoa izan da aztertutako azken parametroa. Adibidetzat hartu dugun erlazioa EBALUAZIOA motakoa da. Horrek 10-11 EDUek egoera bat aurkeztu eta 12 EDUak egoera horri buruzko ebaluaziozko aipamen bat egiten du.

3- Erlaziozko diskurtso-egituretan eta beren osagaiei orientazio semantiko etiketatzaileen artean dagoen adostasuna neurtu dugu. Eskuzko ebaluazioa egin dugu, F-neurria erabiliz.

3.2.3.2. Unitate zentrala

Bestetik, unitate zentralean zer gramatika-kategoritako hitzak diren ohikoenak jakiteko eta orientazio semantikoa duten hitzek nolako banaketa duten aztertzeko, jarraian zerrendatuta dauden urratsak burutu ditugu:

1- Corpusetik unitate zentralak atera. Lehenik eta behin, pertsona batek Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 192 testu hartu eta testu bakoitzaren unitate zentrala hautatu du. Test zatiko testuak ez dira erabili. Unitate zentrala hautatu aurretik, testuak segmentatu egin dira RST hurbilpeneko Das eta Taboadaren (2018) gidalerroei jarraituz. Bestalde, testuen segmentazioa egin ahal izateko, *RSTTools* (O'Donnell, 2000) programa erabili dugu.

3.11 irudian, EGU01 testua segmentatuta ikus daiteke XML formatuan. Segmentu horietako bakoitza oinarrizko diskurtso-unitate bat da eta anotatzailearen zeregina, hurrengo urratsarekin lotura eginez, diskurtso-unitate horietako zein den unitate zentrala, hots, iritzi-testuko diskurtso-unitaterik garrantzitsuena, identifikatzea da.

2- Unitate zentralaren hautaketako adostasuna neurtu. Pertsona horrek unitate zentrala ondo hautatu duen frogatzeko, beste pertsona batek

```
<rst>
  <header>
    <relations>
    </relations>
  </header>
  <body>
    <segment id=''1''>Gaurko eguraldia. HEGO EKIALDEKO HAIZEA ETA HODEI
        BATZUREKIN EPELDU EGINGO DA.</segment>
    <segment id=''2''> Giro ona tokatuko zaigu gaur ere hego haize
        epelarekin.</segment>
    <segment id=''3''> Borraska pixka bat hurbiltzeak haizea apur bat
        mugituko du</segment>
    <segment id=''4''> baina </segment>
    <segment id=''5''>tenperatura igoaz, </segment>
    <segment id=''6''>giroa goxo mantenduko da leku gehienetan.</segment>
    <segment id=''7''> Hodei zirrinta mehe batzuk agertuko dira zeru sabaian
        </segment>
    <segment id=''8''> baina eguzkia apur bat lausotu arren,</segment>
    <segment id=''9''> astro handia bistan edukiko dugu</segment>
    <segment id=''10''> eta dotoreziak ez du bat ere galduko.</segment>
    <segment id=''11''> Baditeke iluntze aldera ekaitz hodei batzuk garatzea
         eta euri zarrastaren batzuk botatzea agian. </segment>
    <segment id=''12''>Akats txiki horiek gora behera eguraldi bikaina oro
        har.</segment>
  </body>
</rst>
```

3.11 irudia: Bi pertsonek egin beharreko unitate zentralaren aukeraketa EGU01 iritzi-testuan.

ere testu horien unitate zentralak hautatu ditu. Bi pertsonen arteko adostasuna neurtzeko ez dira unitate zentral guztiak erabili; izan ere, guztira 78 iritzi-testuren (corpus osoaren % 32,5) unitate zentrala hautatu behar izan dute bi pertsonek. Adostasuna kalkulatzeko F-neurria erabili dugu.

Adostasunaren neurketak 3.19 taulan erakusten dira, 78 iritzi-testuetatik 51etan, bi pertsonek diskurtso-unitate bera jo dute unitate zentraltzat. Beste 44 iritzi-testuetan, aldiz, diskurtso-unitate ezberdinak lotu dituzte unitate zentralarekin. Beraz, 78 iritzi-testuren unitate zentrala aukeratzean egon den adostasuna 0,65ekoa izan da.

3- Unitate zentraletako hitzen gramatika-kategoriaren azterketa. Hurrengo urratsean, unitate zentral bakoitzean zer hitz-mota agertzen den aztertu nahi izan dugu, baita hitzen bat edo batzuk maiztasun erre-

Neurketa	Instantziak	F-neurria
Adostasuna	51	0,65
Desadostasuna	27	0,35

3.19 taula: Bi pertsonen arteko adostasun-maila 78 iritzi-testuetan unitate zentrala aukeratzerakoan.

gular batez azaltzen diren ere. Horretarako, *Analhitza* tresna (Otegi *et al.*, 2017) erabili dugu. Aipaturiko *Analhitza* tresna horretara, hautatuak izan diren unitate zentral guztiak igo ditugu. Unitate zentral horiek domeinuka aztertu ditugu tresnaren bidez. Unitate zentralak aztertzeko, honako alderdiak hartu ditugu kontuan:

i) Hitzen maiztasuna. Gramatika-kategoria bakoitzeko, gehienez maiztasun handieneko 30 hitz aztertu ditugu.

ii) Hitzen gramatika-kategoria. Aintzat hartutako hitzen besten ezaugarri bat beren gramatika-kategoria izan da. *Analhitza*k (Otegi *et al.*, 2017) zuzenean hitzak gramatika-kategoria aintzat hartuz sailkatzen ditu; horrenbestez, lana erraztu digu.

3.20 taulan, *Analhitza* tresnak Otegi *et al.* (2017) eguraldiaren domeinuko hitzak zein gramatika-kategoriatan sailkatu dituen ikus daiteke, Izen-kategoriako hitzak dira gehienak (% 32,65).

Gramatika-kategoria	Kopurua	%
Izenak	80	32,65
Adjektiboak	25	10,20
Aditzak	69	28,16
Adberbioak	19	7,76
Determinatzaileak	10	4,08
Juntagailuak	35	14,29
Preposizioak	0	0,00

3.20 taula: *Analhitza* tresnak (Otegi *et al.*, 2017) eguraldiaren domeinuko hitzei egindako gramatika-kategoriaren sailkapena.

Gainera, *Analhitza* tresnak (Otegi *et al.*, 2017) gramatika-kategoria

bakoitzeko hitzak zerrendatuta ematen ditu, 3.21 taulan ikusi daitekeen moduan.

Adjektiboak
epel
eder
dotore
goxo
eguzkitsu
atsegin
txar
oker

3.21 taula: Eguraldiaren domeinuko unitate zentraleko hitzen zerrenda.

iii) Hitzen domeinua. Maiztasunik handieneko hitz horiek domeinuari oso lotutakoak diren edo ez ere aztertu dugu. Domeinu bakoitzean, hitzak bi multzotan sailkatu ditugu: hitzak semantikoki domeinukoak diren edo ez. Eguraldikoaren kasuan, hitzak hiru multzotan sailkatu ditugu: i) eguraldi domeinuari loturiko hitza, ii) eguraldiaren domeinuarekin zerikusirik ez duen, iii) denbora-adierazpen bat den. Eguraldia domeinuan, denborak garrantzia duela uste dugulako egin ditugu hitzen sailkapenean hiru multzo.

Esaterako, eguraldiaren domeinuko maiztasunik handieneko hitzetan, *negu* hitza domeinuko moduan sailkatu dugu eta *itxura*, berriz, ez; ez dagoelako zuzenean domeinu horri lotuta. Azkenik, *asteburu* hitzak denbora adierazten duenez, denborazkoetan sailkatu dugu.

4- Unitate zentralei eta horien hitzei sentimenduen balentzia esleipena. Aztertu dugun beste ezaugarrietako bat unitate zentraletako gramatika-kategorien sentimendu balentzia izan da. Sentimendu-balentziaren esleipena egiteko sortu dugun euskarazko sentimenduen sailkatzailea erabili dugu.

(40) adibidean, euskarazko sentimenduen sailkatzaileak eguraldiaren domeinuko unitate zentral bati sentimendu-balentzia nola esleitu dion ikus daiteke. *Ezegonkor* adjektiboari esleitu dio sentimendu-balentzia eta gramatika-kategoria bakoitzean zenbat hitzei esleitu zaien sentimendu-balentzia zenbatu dugu.

(40) giro $[ezegonkor]_{-1,5}$ nagusi izan izan etorri egun. (EGU01)

3.22 taulan, eguraldiaren domeinuko gramatika-kategorietan, zenbat hitzek duten sentimendu-balentzia zenbatzen da. Adjektibo-kategoriako hitzen dute sentimendu-balentzia gehien.

EGURALDIA	**Sentimendu-balentziadun hitzen kopurua**
Izena	7
Aditza	12
Adjektiboa	22
Adberbioa	3
Guztira	44

3.22 taula: Eguraldiaren domeinuan, gramatika-kategoria bakoitzeko sentimendu-balentziadun hitzen kopurua.

3.3 Laburpena

Kapitulu honetan, tesi-lan honen metodologia deskribatu dugu. Lehenik eta behin, balentzia-aldatzaileak identifikatzeko eta beren eragina neurtzeko baliabide eta tresnak sortu ditugu: Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) (diskurtso-egituraren eta orientazio semantikoaren aldetik etiketatuta) eta *SentiTegi* (Alkorta *et al.*, 2018) izeneko euskarazko sentimendu lexikoia.

Bigarren urratsean, balentzia-aldatzaileak identifikatu eta euren eragina neurtu dugu. Balentzia-aldatzaile fonologikoak eta morfologikoen instantziak corpus zati batetik eta sentimendu-balentzia esleitu diegu. Gero, instantzia horiek hitzean eragiten duten neurtu dugu.

Ezeztapen-markak ere antzeko modu batean landu ditugu. Ezeztapen-markak corpus zati batean bilatu eta ezeztapen-marka duten esaldiei sentimendu-balentzia esleitu diegu. Beren eragina neurtu ondoren, ezeztapen-markak eta beren irismena identifikatzeko erregelak garatu eta, azkenik, *Murriztapen Gramatika* ingurunera egokitu ditugu horiek ebaluatzeko asmoz.

Diskurtso-mailan, RST hurbilpenaz etiketatutako corpus zatitik esanahi erretoriko ezberdinetako erlazio lortu ditugu. Ondoren, eskuz nahiz *SentiTegi* (Alkorta *et al.*, 2018) erabiliz orientazio semantikoa eta sentimendu balentzia esleitu diegu eta erlazio erretorikoak sentimenduen analisiaren ikuspegitik landu ditugu. Bestalde, unitate zentrala ere landu dugu.

4. KAPITULUA

Sentimenduen analisirako baliabideak

Kapitulu honetan, tesi-lan honen ondorioz garaturiko baliabideak aurkeztuko ditugu. Guztira hiru baliabide garatu ditugu: Euskarazko Iritzi Corpusa (4.1 atala), *Sentitegi* izeneko euskarazko sentimenduen lexikoia (4.2 atala) eta, azkenik, dokumentu mailako euskarazko sentimenduen sailkatzailea (4.3 atala).

4.1 Euskarazko Iritzi Corpusa

Atal honetan, euskarazko egunkari eta webgune espezializatuetik bildutako 240 iritzi-testuez osaturiko corpusa azalduko dugu. Lehenik eta behin, 3.1.1 atalean jarraituriko metodologiaz sortu dugun iritzi-testuen corpusaren ezaugarriak aipatuko ditugu. Ondoren, berriz, metodologiaren garapenean izandako zailtasunen berri emango dugu.

4.1.1. Euskarazko Iritzi Corpusaren ezaugarriak

Sortu dugun Euskarazko Iritzi Corpusak guztira 240 iritzi-testu ditu sei domeinutakoak: kirola, politika, musika, zinema, liburuak eta eguraldia.

Corpusaren beste ezaugarrietako bat bertako iritzi-testuetan agertzen den balorazioaren oreka da. Domeinuko 40 testu daude eta erdiek balorazio positiboa dute eta beste hainbestek, berriz, negatiboa.

Analhitza tresnak (Otegi *et al.*, 2017) ematen dituen emaitzen arabera, corpusak 52.092 token eta 3.711 esaldi ditu. Corpusa beste hizkuntzetako iritzi-

testuen corpusekin alderatuz, tamainaz zertxobait txikiagoa dela esan daiteke, 4.1 taulan agertzen den moduan. Beste hizkuntzetan, oro har, iritzi-testu kopurua handiagoa da, baina, tokenen kasuan, euskarazko corpusaren eta beste corpusen arteko ezberdintasuna are handiagoa da.

Corpusa	**Testuak**	**Hizkuntza**	**Tokenak**
Boldrini *et al.* (2010)	300		
Hu eta Liu (2004)	113		81.855
Rushdi-Saleh *et al.* (2011)	500	Arabiera	215.948
Wilson (2008)	535	Ingelesa	265.000
Taboada (2008)	400	Ingelesa	289.270
Quan eta Ren (2009)	1.487	Txinera	878.164

4.1 taula: Sentimenduen analisirako eskura dauden iritzi-testuen corpusak.

4.1.2. Euskarazko Iritzi Corpusaren garapena

Euskarazko Iritzi Corpusa sortzerakoan, bi zailtasun izan ditugu. Alde batetik, guk nahi genituen moduko iritzi-testuak bilatzea zaila suertatu da (4.1.2.1 atala). Beste aldetik, domeinu batzuetan, orientazio semantiko jakin batzuetako iritzi-testuak biltzea ere asko kostatu da (4.1.2.2 atala). Azkenik, 4.1.3.2 atalean, corpusaren baliagarritasuna neurtzeko hautatutako irizpideak zergatik hautatu ditugun azalduko dugu.

4.1.2.1. Ezaugarri jakin batzuetako iritzi-testuak bilatzeko zailtasunak

Gure lana eta lanaren emaitza konparatu ahal izateko, badagoen baliabide (kasu honetan, corpusa) baten antzeko bat egin nahi izan dugu. Hau da, hitz edo sintagmen balentzia aldatzen duten fenomeno linguistikoak aztertzeko moduko iritzi-testuak bildu nahi ditugu. Horregatik, Euskarazko Iritzi Corpusa sortzerakoan, *SFU Review Corpus*[1] (Taboada, 2008) hartu dugu erreferentziatzat.

[1]Corpusa eskuragarri dago `https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html` webgunean.

*SFU Review Corpus*ak (Taboada, 2008) zortzi gaietako 400 iritzi-testu biltzen ditu. Gai bakoitzeko 50 testu daude, erdiak iritzi positibodunak dira eta beste erdiak, berriz, negatibodunak. Iritzi-testuek lantzen dituzten gaiak hauek dira: liburuak, autoak, ordenagailuak, sukaldeko tresneria, pelikulak, musikak eta telefonoak.

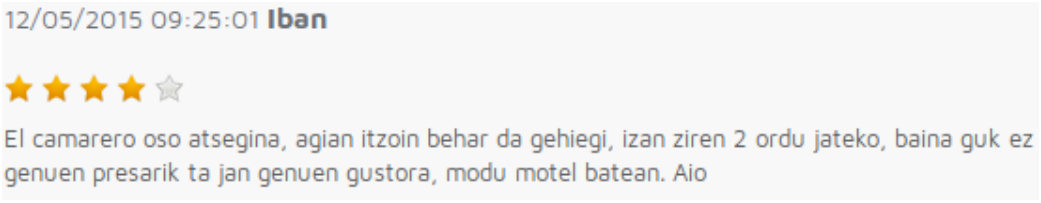
Gure asmoa gai horiei buruzko iritzi-testuak biltzea izan da, baina euskara hizkuntza gutxitua denez, eta horrek zailtasunak areagotu egiten dituenez, asmoa birplanteatu egin behar izan dugu. Izan ere, liburu, musika eta pelikulei buruzko iritzi-testuak aurkitu ditugu, baina besteen kasuan, kalitate baxukoak edo desegokiak diren iritzi-testuak aurkitu ditugu edo ez ditugu aurkitu.

4.1 irudian, *Tripadvisor* webgunetik hartutako iritzi-testu bat ikus daiteke. Hiru hizkuntzetan idatzita dago iritzi-testu bakarra eta hiruretan esaten dena ez da guztiz berdina. Hizkuntza bat baino gehiago agertzen diren iritzi-testuak aurkitzea oso ohikoa izan da gure iritzi-testuen bilaketan. Badirudi, iritzi-emaile euskaldunek beren iritziek ahalik eta irismen handiena izatea nahi dutela eta, horregatik, euskaraz gain beste hizkuntza bat erabiltzen dutela.

4.1 irudia: Iritzi-testu hirueledun baten adibidea.

Halaber, aurkitu ditugun euskarazko iritzi-testu gehienak labur samarrak dira eta horrek gure ikerketa-lana mugatu dezake, aztertu nahi ditugun hizkun-

tza-gertakariak maiztasun txikiagoz agertu direlako. Euskarazko iritzi-testu labur bat ageri da 4.2 irudian.



4.2 irudia: Euskarazko iritzi-testu labur baten adibidea.

Ikusten denez, ortografia zuzena ez izateaz gain, egitura sintaktikoa ez da aberatsa eta bertatik diskurtsoaren egitura lantzea ere zailagoa bilakatzen da, testuaren tamaina ez delako handia. Egoera horren ondorioz, zenbait aldizkari eta komunikabide aztertu ditugu eta aurkitzen errazagoak diren beste gai batzuk dituzten iritzi-testuak biltzea pentsatu dugu. Hala, liburuei, musikari eta pelikulei buruzko iritziak biltzeaz gain, eguraldiari, politikari eta kirolari buruzko iritzi-testuak biltzea erabaki dugu. Gainera, ahalik eta kalitate hobereneko iritzi-testuak biltzeko, komunikabide eta aldizkari espezializatuetara jotzea erabaki dugu.

Nahiz eta iritzi-testuen gaiak batzuetan *SFU Review Corpus*ekoetatik (Taboada, 2008) urrundu diren, beste ezaugarrietan antzekoak izan daitezen saiatu gara. Horregatik, erabaki dugu Euskarazko Iritzi Corpusak domeinu bakoitzeko 40 testu izatea eta domeinu bakoitzeko, iritzi-testuen erdiak positiboak izatea eta besteak negatiboak izatea. Beraz, gai bakoitzeko 20 testu positibo eta beste hainbeste negatibo bildu ditugu.

4.1.2.2. Orientazio semantiko jakin bateko iritzi-testuak bilatzeko zailtasunak

Iritzi-testuak balorazioen aldetik orekatuak izateak beste zailtasun eta arazo batzuk ekarri dizkigu. Esaterako, literatura-liburuen kasuan, zailtasunak izan ditugu balorazio txarreko iritziak aurkitzeko. Zailtasun horren jatorriaren berri Egañak (2013) ematen du.

Egañak (2013) 1975tik 2005ra bitarteko 2.300 euskal literatura kritika aztertu ditu eta haren ikerketaren arabera, literatura kritiken % 84k iruzkin positiboa

edo oso positiboa dute. *Berria* egunkarian egindako elkarrizketan dioenez[2], "euskal literaturak berak ere bere burua legitimatzeko arazoak" ditu eta horren ondorio da iruzkin positiboen kopuru handia. Halaber, "Euskal literaturan oso jende gutxi da kritikari profesionala" eta "subjektuak eta legitimatua den objektuak elkarrengandik urrun samar" ez daudenez, legitimatzeko arazoak sortzen dira. Hots, euskal literatura komunitate txikia denez, eta denen arteko loturak beste literatura-komunitateekin alderatuta estuagoak direnez, iruzkin negatiboak gutxiago egiten dira. Egoera hori aintzat hartuta, balorazio negatiboak bereziki atzerriko lanei buruzko iritzi-testuetan aurkitu ditugu. Politikako iritzi-testuetan, berriz, aurkakoa gertatu da. 2008tik aurrera dagoen krisi ekonomikoa tarteko, iritzi gehienak iruzkin negatibokoa izan da eta iruzkin positibokoak aurkitzea zaila gertatu zaigu.

4.1.3. Euskarazko Iritzi Corpusaren baliagarritasunaren neurketa

Euskarazko Iritzi Corpusa osatu ondoren, corpus horrek subjektibotasunik baduen eta ezeztapen-marken instantzia ugari badauden eta, horrenbestez, sentimendu-analisirako baliagarria den neurtu dugu. Halaber, gure zenbait ikerketa-lerrotarako (hizkuntza maila ezberdinetako balentzia-aldatzaileak identifikatzeko eta euskarazko sentimenduen lexikoi bat sortzeko) beharrezkoak diren hizkuntza-fenomenoak ere agertzen ote diren aztertu dugu. Aztertzea erabaki ditugun aspektuak (lehen pertsonaren erabilera, adjektiboen agerpena eta ezeztapena) iturri bibliografikoetan oinarrituta hautatu ditugu. Jarraian, i) corpusaren baliagarritasuna neurtzeko erabili ditugun baliabideak eta irizpideak eta ii) neurketen emaitzak azalduko ditugu.

4.1.3.1. Baliagarritasuna neurtzeko baliabideak eta irizpideak

Euskarazko Iritzi Corpusaren subjektibotasuna neurteko lehen pertsonaren erabilera eta adjektiboen agerpena neurtu dugu eta, horretarako baliabideak 4.2 taulako corpusak izan dira. Corpus horiek alderatu egin ditugu.

[2]Elkarrizketa esteka honetan dago eskuragarri: `www.berria.eus/paperekoa/1744/024/001/2015-02-04/literatura_batek_sendotzeko_gatazka_behar_du_bere_barruan.htm`, *Berria*, 2015-02-14.

	Subjektiboa	**Objektiboa**
Euskara	Euskarazko Iritzi Corpusa	Euskarazko Wikipediako artikuluak
Ingelesa	SFU Review Corpus	Ingeleseko Wikipediako artikuluak

4.2 taula: Lehen pertsonaren eta adjektiboaren agerpena neurtzeko erabilitako corpusak.

- *SFU Review Corpus* (Taboada, 2008). Gure erreferentziazko corpusa da. Biek hainbat domeinutako iritzi-testuak biltzen dituzte eta, horrenbestez, eduki subjektiboa dute. 16.705 esaldi ditu corpusak.
- Ingelesezko Wikipediako zenbait artikulu. *SFU Review Corpus*aren (Taboada, 2008) domeinuekin lotura duten artikuluak hartu ditugu. Guztira 8 artikulu bildu ditugu eta corpus horrek 64.258 hitz ditu.
- Euskarazko Wikipediako zenbait testu. Ingeleseko Wikipedian bezala, kasu horretan ere, Euskarazko Iritzi Corpuseko domeinuekin bat datozen artikuluak erabili ditugu. Euskarazko Wikipediaren kasuan, 28 artikulu bildu ditugu eta corpus horrek 37.964 hitz ditu.

Wikipedia entziklopedia bat izanik, artikuluek ikuspegi neutroa edukitzea eskatzen da[3]. Testurik objektiboenak horiek direla uste dugulako, ingelesezko eta euskarazko corpus objektiboak Wikipedian oinarrituta sortu ditugu.

Lehen pertsonaren eta adjektiboaren agerpenaren neurketa aipaturiko corpusetan horrela eta arrazoi horiengatik gauzatu dugu:

1- Lehen pertsonaren agerpena. Iritziak, normalean, lehen pertsonan adierazi ohi dira (Halliday *et al.*, 2014) lehen pertsona zenbateko maiztasunez agertzen den aztertu dugu gure corpusean. Gure corpuseko emaitzak konparatzeko 4.2 taulako beste hiru corpusak erabili ditugu.

   Lehen pertsonaren agerpena neurtzerakoan, zailtasun batekin egin dugu topo. Euskaran, aditz laguntzailean agertzen dira aditzari lotutako marka morfologikoak eta, horrenbestez, baita pertsona eta nume-

[3]Esteka honetan azaltzen dira Wikipediaren ezaugarriak: `https://eu.wikipedia.org/wiki/Wikipedia`.

roari buruzko informazioa ere. Ingelesean, aldiz, hirugarren pertsonaren kasuan izan ezik, aditz-forma guztiak berdinak dira eta horrek lehen pertsona neurtzea eragozten du. Konponbide moduan, ingelesean pertsona-izenordainen presentzia neurtzea erabaki dugu. Beraz, lehen pertsonaren agerpena neurtzeko, euskaran aditz laguntzaileak erabili ditugu eta ingelesean, ordea, pertsona-izenordainak. Neurketa egitean, lehen pertsona hori singularra edo plurala den kontuan hartu dugu.

2- Adjektiboen agerpena. (Liu, 2010, 32 orr.) lana gramatika-kategoriez (*Part-Of-Speech*, POS) ari denean, adjektiboak iritzien adierazle garrantzitsuak direla adierazten du. Horregatik, adjektiboek gramatika-kategoria guztien artean duten presentzia neurtu dugu.

Euskarazko Iritzi Corpusean ezeztapen-marka asko ageri diren zenbatu dugu; izan ere, ezeztapena nolako balentzia-aldatzailea den landutako hizkuntza-fenomenoetako bat izan da. Kasu horretan, bere agerpen asko dauden interesatzen zaigu eta ez beste corpusekin alderatzea; izan ere, ezeztapena lehen pertsona eta adjektiboen erabilera ez bezala ez dira subjektibitatearen adierazle.

3- Ezeztapena. Ezeztapena garrantzitsua da esaldien eta testuen sentimendu-balentzia aldatzeari dagokionez, Polanyi eta Zaenenen arabera (2006) testuinguruko balentzia-aldatzailea da. Hori dela eta, corpusean agertzen diren ezeztapen-markak zenbatzea erabaki dugu; instantzia kopuru minimo bat behar baitugu azterketa ondo gauzatzeko.

   Laburbilduz, lehen pertsonaren eta adjektiboen agerpen kopurua beste corpusetako (euskarazko eta ingelesezko Wikipediako corpusak eta *SFU Review Corpus* (Taboada, 2008)) agerpen kopuruarekin alderatu dugu. Agerpen kopuruaren alderaketa eginez Euskarazko Iritzi Corpusa tesi-lanerako baliagarria zaigun zehaztu dugu. Ezeztapenaren kasuan, aldiz, beren agerpena Euskarazko Iritzi Corpusean zenbatu egin dugu, aztertu dugun balentzia-aldatzaileetako bat baita.

4.1.3.2. Neurketen emaitzak

Corpusaren baliagarritasuna neurtzeko aztertu diren ezaugarriak hauek izan dira: lehen pertsonaren agerpena (neurketa eta alderaketa), adjektiboaren agerpena (neurketa eta alderaketa) eta ezeztapena (kontaketa).

- Lehen pertsonaren agerpena. Corpus horietan agertzen den lehen pertsona kopurua 4.2 taulako baliabideetakoekin alderatu dugu. Euskaran, lehen pertsonaren erabilera (singularra eta plurala) aditz laguntzailean neurtu dugu, eta ingelesean, pertsona-izenordainetan.

Corpusa	Lehen sg.	Lehen pl.	Guztira
Euskarazko Wikipedia	% 0,12	% 1,09	**% 1,21**
Ingelesezko Wikipedia	% 0,03	% 0,09	**% 0,12**
Euskarazko Iritzi Corpusa	% 3,27	% 5,10	**% 8,37**
SFU Review Corpus	% 10,10	% 1,70	**% 11,80**

4.3 taula: Lehen pertsonaren agerpena lau corpus ezberdinetan.

4.3 taulan agertzen den moduan, euskarazko nahiz ingelesezko Wikipedietatik hartutako testuekin osatutako corpusetan, lehen pertsonaren agerpena oso baxua da. Ingelesezko Wikipediaren kasuan, azaltzen diren pertsona-izenordain guztien % 0,12 lehen pertsonari dagokio. Joera bera erakusten du euskarazko Wikipediak, non bertako aditzetako marka morfologikoetan, % 1,21a lehen pertsonakoak diren.

Corpus subjektiboetan, aldiz, lehen pertsonaren agerpena handiagoa da. Euskarazko Iritzi Corpusean, aditzetako marka morfologikoetan, lehen pertsonaren agerpena % 8,37koa da. *SFU Review Corpus*ean (Taboada, 2008), lehen pertsonaren agerpena handiagoa da; pertsona-izenordainen % 11,80 baitago lehen pertsonan. Bestalde, aipagarria da corpus subjektiboetan, lehen pertsona plurala dela nagusi euskararen kasuan eta lehen pertsona singularra, berriz, ingelesaren kasuan.

- Adjektibo-kopurua. Corpusaren baliagarritasuna neurtzeko aintzat hartu dugun beste alderdi bat adjektiboen kopurua izan da. Adjektiboak oinarrizko elementuak dira maila lexikoan subjektibotasuna erakusterakoan, beraz, corpus subjektiboetan adjektiboen agerpen handiagoa

espero liteke corpus objektiboetan baino. Hala ere, 4.4 taulako emaitzek ez dute hori erakusten.

Corpusa	Guztira
Euskarazko Wikipedia	**% 8,50**
Ingelesezko Wikipedia	**% 9,09**
Euskarazko Iritzi Corpusa	**% 9,82**
SFU Review Corpus	**% 8,35**

4.4 taula: Adjektiboen agerpena corpus objekibo eta subjektiboetan.

Emaitzen arabera, adjektiboen agerpena antzekoa da corpus guztietan, bai objektiboa edo subjektiboa izanda, baita euskarazkoa edo ingelesezkoa izanda ere. *Analhitza*k (Otegi *et al.*, 2017) emandako datuen arabera, Euskarazko Wikipedian eta ingelesezko Wikipedian, adjektiboak gramatika-kategoria guztietatik % 8,50 eta % 9,09 dira, hurrenez hurren. Euskarazko Iritzi Corpusean adjektiboak gramatika-kategoria guztien % 9,82 dira eta *SFU Review Corpus*ean (Taboada, 2008) % 8,35. Ondorioz, badirudi adjektiboen agerpen handiagoa ez dela corpus subjektiboen ezaugarri bat. Baliteke horren ordez, adjektibo-mota izatea testu objektibo eta subjektiboen arteko ezberdintasuna.

- Ezeztapena. Corpusean neurtu dugun azken ezaugarria ezeztapena eta zehazki, ezeztapen-markak izan dira. Hiru ezeztapen-marka moten kontaketa egin dugu.

  Alde batetik, perpaus osoari (adibidez, *gaur dendak ez daude zabalik*) eragiten dion *ez* ezeztapen-marka 718 aldiz agertzen da. Izen sintagmari (adibidez, *arrakastarik gabe*) eragiten dino *gabe* ezeztapen-marka 107 aldiz agertzen da eta azkenik, menpeko perpausei eragiten dien *ezean* lau aldiz zenbatu dugu. Kasu horietan, ezeztapen-marka kopuru handia aurkitu dugu corpusean eta horrek horiek ikertzeko aukera eman digu.

4.2 Euskarazko sentimenduen lexikoia

Atal honetan, hasteko, euskarazko sentimenduen lexikoiaren bi bertsioek dituzten ezaugarriak aurkeztuko ditugu. Ondoren, euskarazko lexikoiaren garapeneko zenbait alderdi (ordainen kasuistika eta hartutako erabakiak, besteak beste) aipatuko ditugu eta, azkenik, lexikoiaren ebaluazioa azalduko dugu.

4.2.1. Euskarazko sentimenduen lexikoiaren ezaugarriak

Gaztelaniazko SO-CAL lexikoia (Brooke *et al.*, 2009) itzultzeko jarraitu dugun prozesuaren ondorioz, euskarazko sentimenduen lexikoi beraren bi bertsio lortu ditugu. 4.5 taulak sentimenduen lexikoiaren bi bertsioen emaitzak erakusten ditu.

	V1.0		V2.0	
Gramatika-kategoria	**Sarrerak**	**%**	**Sarrerak**	**%**
Izenak	2.282	28,06	461	37,27
Adjektiboak	3.162	38,85	446	36,05
Aditzondoak	652	7,98	54	4,36
Aditzak	1.657	20,36	276	22,32
Intentsifikatzaileak	387	4,75		
Guztira	**8.140**	100	**1.237**	100

4.5 taula: Euskarazko sentimenduen lexikoiaren bi bertsioen ezaugarriak.

Sentimenduen lexikoiaren lehen bertsioa (V1.0) bigarren bertsioa (V2.0) baino handiagoa da. Horren atzean zenbait arrazoi daude. Hasteko, lehen bertsioan *Zehazki* (Sarasola, 2005) eta *Elhuyar* (Elhuyar, 2013) hiztegien bidez lorturiko euskarazko ordain guztiak hartu dira kontuan. Bigarren bertsioan, aldiz, bi hiztegi horietako sarrerak diren euskarazko ordainak bakarrik hartu dira kontuan. Horrenbestez, itzulpen bidez lortu diren lexikoiaren sarrerek aniztasun handia erakusten dute: batzuek izenlagun hizkia daramate (adibidez, *garrantzizko*), beste batzuk kolokazioak[4] dira, etab. Halaber, intentsifi-

[4]Kolokazioak uste baino maiztasun handiagoaz batera agertzen diren hitzak dira. *Adarra jo* horren adibidea da.

katzaileak ere lexikoiaren lehen bertsioan badaude, bigarrengoan ez bezala.

Lexikoi horien arteko berdintasunak eta ezberdintasunak aintzat hartuta, lexikoiaren lehen bertsioak 8.140 sarrera ditu, bigarren bertsioak 1.237 dituen bitartean. Batetik bestera asko jaisten da sarrera kopurua; izan ere, bigarrengo bertsioa domeinuari lotuta dago. Bi bertsioetan izenak eta adjektiboak dira gramatika-kategoria nagusienak. Aditzak zerbait gutxiago dira eta aditzondoak, berriz, are gutxiago. Bigarren bertsioan, intentsifikatzailerik ez dago arrazoi praktikoengatik. Gure ustez, intentsifikatzaileek beste hitzei eragiten dietenez, gertuago daude sintaxi mailatik lexiko mailatik baino. Gainera, kontuan hartu behar da intentsifikatzaileek % −100 eta % +50 arteko balioa dutela biderketa eragiketan, beste gramatika-kategoriek −5 eta +5 arteko balioak (gehiketa eta kenketa eragiketak) dituzten bitartean.

Lexikoiaren bi bertsioen beste ezaugarri bat lexikoi paraleloak direla da. Hots, lexikoietako sarrerekin datu-base bat sortuta dago eta euskarazko sarrerarekin gaztelaniazko eta ingelesezko ordainak (bigarren bertsioan) paraleloki eta modu ordenatuan daude jarrita. Gaztelaniazko eta ingelesezko ordainek ere beren sentimendu-balentzia badute, 4.6 taulan ikus daitekeenez. Bertan, lexikoia paraleloa dela erakusten duten lau adibide daude. Adibideak adjektiboak dira eta beren sentimendu-balentziak eskalan daude.

Hitza lexikoian	**Balentzia**	**SPA**	**Balentzia**	**ENG**	**Balentzia**
bikain	+5	excepcional	+5	excellent	+5
on	+2	buen	+2	-	-
eskas	−1	escaso	−2	insufficient	−1
txar	−3	adverso	−3	bad	−3

4.6 taula: Lexikoi paraleloaren adibide batzuk.

Ikusten den moduan, batzuetan zerrendetako sentimendu-balentziak ez datoz bat (*eskas* (−1), *escaso* (−2) eta *insufficient* (−1), esaterako), gaztelaniazko eta ingelesezko lexikoiak ez direlako modu bere-berean sortu. Baina, euskarazko lexikoiko sarreretako sentimendu-balentziak beti ingeleseko edota gaztelaniako sarreretako sentimendu-balentzia batekin bat egin behar du.

Behin euskarazko lexikoiaren euskarazko lehen bertsioa euskarazko sentimenduen sailkatzailean integratu ondoren, sorturiko euskarazko sentimen-

duen lexikoia gauza da hitzei nahiz esaldiei sentimendu-balentzia automatikoki esleitzeko. Sentimenduen lexikoiak hitzei esleitzen die sentimendu-balentzia, eta ondoren, sentimenduen sailkatzaileak eragiketa egiten du esaldiaren sentimendu-balentzia kalkulatzeko. Hurrengo adibideek hori erakusten dute.

(41) Halere, pentsa litekeenaren aurka, gaien urritasunak eta diskurtso [errepikakorrak]$_{-6}$ ez dakarte ñabardura aberastasunik, are gutxiago argumentu-mailako sakontasunik.[–6] (LIB18)

(42) Arazo [nagusia]$_{+2}$, nire ustez, gaien [emankortasun]$_{+4}$ zalantzazkoan eta ekintzaren bilakaera [eskasean]$_{-3}$ datza.[+3] (LIB18)

(43) (...) Emaitza [ezustekorik]$_{-1.5}$ gabeko istorio bat da, irakurlea [epel]$_{-1.5}$ uzteko arrisku dezente duen tonu arras moderatu batean emana.[–3] (LIB18)

Goiko adibideetan, hitzek eskuinetara sentimendu-balentzia bat dute, euskarazko sentimenduen sailkatzailean inplementatu dugun sentimenduen lexikoiarenak. Esaldien amaieran, esaldi osoaren sentimendu-balentzia ageri da eta esaldian dauden sentimendu-balentzi guztien batura da. Lehen adibidean, esaterako, lexikoiaren arabera hitz batek, *errepikakorrak* hitzak, bakarrik du sentimendu-balentzia –6 eta ondorioz, esaldiarena ere halaxe da[5]. Hurrengo bi adibideetan ere, lexikoiak funtzionamendu bera du euskarazko sentimenduen sailkatzailearen baitan.

4.2.2. Lexikoiaren sorkuntza

Euskarazko sentimenduen lexikoia sortzeko metodologian, hartu ditugun erabakiak azalduko ditugu (4.2.2.1 atala); ondoren, gaztelaniazko lexikoia euskaratzerakoan agertu zaigun kasuistikaren berri emango dugu (4.2.2.2 atala);

[5]Gogoan hartu behar da lexikoiko sarreren sentimendu-balentzia –5 eta +5 artekoa dela, baina SO-CAL tresnan badaude zenbait eragiketa hizkuntzarekin zerikusia dutenak eta hizkuntza guztietan aplikatzen direnak eta horiek hitzaren sentimendu-balentzia indar edo ahul dezakete. Kasu horretan, *errepikakorrak* hitzaren sentimendu-balentzia –6raino indartu da.

jarraian, gaztelaniazko lexikoi horren euskarazko ordainak ematen sentimenduen balentziak aukeratzeko zein irizpide hartu ditugun kontuan azalduko dugu (4.2.2.3 atala).

4.2.2.1. Metodologian hartutako erabakiak

Sentimenduen lexikoiko hitzek eta beren sentimendu-balentziek elkarren artean koherentzia izan dezaten, metodologian zehar erabaki hauek hartu ditugu:

- Nahiz eta gaztelaniazko lexikoia euskaratu dugun, kasu gehienetan, euskarazko ordainei ingelesezko lexikoiaren sentimendu-balentziak esleitu dizkiegu. Lan honetan gaztelania edota ingelesa izan zitezkeen hastapeneko hizkuntzak. Guk gaztelania hobetsi dugu hastapeneko hizkuntza moduan arrazoi hauengatik: alde batetik, sentimenduen lexikoiaren gaztelaniako bertsioaren batez besteko doitasuna % 71,81ekoa da, ingelesekoarena % 76,65 den bitartean. Beste aldetik, gaztelaniako hitzei euskarazko ordaina emateko baliabide gehiago daude ingelesetik baino eta, gaztelaniatik ordaina ematea errazagoa izango litzateke.

  Halaber, beste arrazoi bat badago eta hori hizkuntzari lotuta dauden ezaugarri soziokulturalak dira. Euskarak eta gaztelaniak ezaugarri soziokultural gehiago partekatzen dituzte, ingelesak eta euskarak baino. Horren adibide dugu 4.2.2.2 azpiataleko 4. Fenomenoa, *frankismo* hitzaren kasua, hain zuzen ere. Euskarak eta gaztelaniak partekatzen duten ale hitz bat da. Horrelako kasu gehiago egon dira euskarazko ordainak ematean. Hizkuntzen artean gertatzen diren itzulpeneko kasu hauek Meng *et al.*en (2012) lanean aipatzen dira eta Mohammad *et al.*ek (2016) itzulpenetan gertatzen diren sentimenduen aldaketak ere aipatzen ditu.

  Bestalde, ingelesezko lexikoiak gaztelaniazko lexikoia baliatuta euskaraz sortu dugun lexikoian esleitu diren balentziak egokiak diren aztertzeko eta ez badira zuzentzeko balio izan digu. 3.10 taulako 3. Fenomenoan (*seinale*) gertatu den moduan, euskarazko ordain batek aukeran gaztelaniazko lexikoiko ale hitzaren (*cicatriz*) eta ingelesezko

lexikoiko ale hitzaren (*signal*) sentimedu-balentziak dituenean, ingeleseko sentimendu-balentziaren alde egin dugu, batez besteko doitasuna handiagoa duelako (% 76,65). Euskarazko sarrerak ingeleseko lexikoian ordainik ez duenean, aldiz, kasu batzuetan euskarazko sarrera hori lexikoitik kendu egin dugu, gaztelaniako lexikoiaren doitasun baxuagoak (% 71,81) euskarazko lexikoiaren kalitatea jaitsi dezakeelakoan. Ondorioz, 576 gaztelanizko hitzei, hasiera batean, euskarazko ordaina eta sentimendu-balentzia esleitu diegun arren, horiek euskarazko sentimenduen lexikoitik kendu egin ditugu, ingelesezko lexikoian agertzen ez direlako.

Beraz, batetik, lexikoiaren gaztelaniazko bertsioaren estaldura eta ezaugarri soziokulturalak eta, bestetik, lexikoaren ingelesezko bertsioaren doitasuna (hitzen balentzia) uztartu ditugu.

- Gaztelaniazko lexikoiko hitz baten euskarazko hainbat ordain kontuan hartzea. Gaztelaniazko lexikoiko hitzei euskarazko ordaina ematean izan dugun beste zalantzetako bat euskarazko ordain guztiak kontuan hartu behar diren izan da. Izan ere, gaztelaniazko hitzak batzuetan euskarazko hainbat ordain izan ditzke esanahi bera mantenduz. Esaterako, gaztelaniazko hitz batek euskarazko hainbat izan ditzake euskaran, 4.7 taulan ikusten den moduan. Halaber, hitzak polisemikoak ere izan daitezke, eta horrek ere gaztelaniazko hitz bati euskarazko hainbat ordain ematea eragiten du; nahiz eta esanahia kasu horretan ezberdina den. Egoera horren aurrean, hau da, gaztelaniazko hitz bati euskarazko hainbat ordain ahal zaizkionean, euskarazko ordain posible guztiak aintzat hartzea erabaki dugu sortuko den lexikoiak ahalik eta estaldura handiena izan dezan.

  4.7 taulan bi adibide daude non ordain posible guztiak kontuan hartu direla erakusten den. Gaztelaniazko lexikoian *enfermedad* eta *horror* hitzak daude eta horiek *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegiak erabiliz euskarazko ordainak lortu dira eta hitz horietako bakoitzak euskaran hainbat ordain posible ditu, hirugarren zutabean agertzen den moduan.

- Hitz polisemikoen domeinuaren egokitzapena. Gaztelaniako lexikoiko

Gaztelaniaz jatorrian	Balentzia	Ordain posible guztiak euskaraz
enfermedad (izena)	−1	gaixotasun, gaitz, eritasun, afekzio, ezontsa
horror (izena)	−4	laztura, izu, lazgarrikeria, izugarrikeria

4.7 taula: Euskarazko ordain posible bat baino gehiago dituzten adibideak.

hitzei euskarazko ordaina ematean eta sentimendu-balentzia esleitzean, testuinguruaren arabera esanahi ezberdinak eta zeinu ezberdinetako (+ edo −) sentimendu-balentziak dituzten ordainak agertu dira, 4.8 taulako modukoak. Egokiena euskarazko ordain guztiak (eta sentimendu-balentziak) kontuan hartzea izango litzateke, nahiz eta euskarazko ordainek esanahi ezberdinak izan, baina horrek lexikoiaren konplexutasuna asko handituko luke. Horretaz gain, lexikoia inplementatuta egongo litzakeen sistemak hitzen adiera-desanbiguaziorako gaitasuna eduki beharko luke.

Hori konpontzeko, esanahi ezberdinak eta zeinu ezberdinetako sentimendu-balentziak dituzten ordainetan esanahi bat bakarrik hartzea erabaki dugu. Hautaketa egitean, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen diren testuak eta domeinuak erabili ditugu.

4.8 taulan, hitz polisemikoetan balentzia hautaketa nola egin den esplikatzen da. *Deigarri* eta *lehiatsu* adjektiboek esanahi ezbedinak (*deigarri*ren kasuan, *aparatoso* −3, *llamativo* +3 eta *lehiatsu*ren kasuan, *ansioso* −3, *apasionado* +4 eta *competitivo* +3) dituzten hitzen ordainak dira eta, metodologiako laugarren urratsean, euskarazko ordain horiei sentimendu-balentzia esleitu behar diegu. Aukeran sentimendu-balentzia bat baino gehiago daudenez (*deigarri*ren kasuan, −3 eta +3 eta *lehiatsu*ren kasuan, −3, +4 eta +3), euskarazko ordain horien testuingurua aztertu dugu Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) gako-hitzak testuinguruan (*Key Word In Context*, KWIC) erabiliz eta ordainen corpuseko testuinguruekin hobekien datorren sentimendu-balentzia aukeratu dugu. *Deigarri* eta *lehiatsu* ordainetan, +3 sentimendu-balentzia izan da hautatu duguna.

- Koherentziaren sendotasuna. Euskarazko ordaina lortu ondoren, ordai-

Euskal ordaina	Jatorriz gaztelaniaz eta balentzia	Bal. aukeratua
deigarri (adj.)	aparatoso −3 , llamativo +3	+3
lehiatsu (adj.)	ansioso −3 , apasionado +4 , competitivo +3	+3

4.8 taula: Hitz polisemikoen tratamendua.

nei sentimendu-balentzia esleitu diegu, balentziak bat zetozen kasuetan, balentzietan koherentzia mantentzen saiatu gara.

Irizpidea	Euskara	Balentzia	Euskara	Balentzia
A	errukigabe	−4	errukigabeko	−4
B	tonto	−3	tuntun	−3
C	arduradun	+2	arduragabe	−2

4.9 taula: Itzulpen-prozesuaren koherentzia erakusten duten adibideak.

4.9 taulan ageri dira zenbait adibide. Bertan agertzen diren kasuak euskarazko ordaina ematerakoan laugarren urratsekoak dira, balentzia hautatu behar den mementokoak. Guztira hiru kasu ezberdin antzeman ditugu:

A- Izenlagunarekin agertzen diren euskarazko ordainak. 4.9 taulan, A kasuan, *errukigabe* lema ageri da. *Errukigabe* eta *errukigabeko* izenondoak dira, baina, bigarrenak izenlagunaren *-ko* atzizkia ere badu. Bi hitzei balentzia hautatzeko mementoan, balentzia bera eman diegu; gramatika-kategoria gorabehera, esanahi bera baitute. Aipagarria da horrelako kasu asko egon direla balentzia hautaketan zehar. Izenlaguna daraman euskarazko ordainaren beste kasu bat *berehala/berehalako* hitz parea izan da eta biei +2 balentzia esleitu diegu. Kontuan hartu behar da, kasu hori euskarazko lexikoiaren lehen bertsioa sortzerakoan gertatu dela, lexikoiaren bigarren bertsioan hiztegietako sarrerak diren hitzak bakarrik hartu baititugu kontuan.

B- Esanahi bera, baina genero ezberdina. Kasu gutxi batzuk egon dira non bi hitz ezberdin erabiltzen diren generoarengatik, bai-

na esanahi bera dutenak. Horrelako kasuetan ere bi hitzetan sentimendu-balentzia bera jartzen ahalegindu gara; aukeran zeuden sentimendu-balentziek uzten zuten heinean. Hala, 4.9 taulako B kasuan, gizonezkoentzat *tonto* eta emakumezkoentzat *tuntun* hitzak ditugu, aipaturiko egoera islatzen dutenak. Ezberdintasun bakarra hitzek erreferentzia egiten dioten pertsonen generoa denez, biei sentimendu-balentzia bera (−3) esleitu diegu.

C- Hizkiak dituzten hitzak. Gaztelaniako lexikoiko hitzei euskarazko ordaina ematean agertu den beste hitz-multzo mota da hizkiak dituztenena. Hizkiak erabiltzearengatik, aurkakotasuna adierazten duten hitzen bikoteak agertu dira ordaina ematean. Hitz bat aurkakotasuna adierazten duen hizkiarekin eta hori gabe agertzea nahiko ohikoa izan da. Aurkakotasuna adierazteko gehien agertu diren hizkiak ezeztapenarekin lotutako hauek izan dira: *des-*, *ez-*, *-ezin* eta *-gabe*. Hitz bera aurkakotasun hizkiarekin eta gabe agertu izan denean, bi egoeretako intentsitate bereko sentimendu-balentzia bera jarri diegu, baina zeinu ezberdinarekin. 4.9 taulan, C kasuan, hitz bera ageri da (*ardura*), baina batean *-dun* atzizkia du, jabetza adierazten duena eta bestean, berriz, *-gabe* atzizkia du, gabezia adierazten duena.

- Ordain *ez-zuzenak*. Euskarazko ordainak ematean agertu den beste fenomenoetako bat ordain *ez-zuzen*ena da. Ordain horiek berez zuzenak dira, baina gure ikerketa-lanerako ez dira baliagarriak testuinguru zehatz batzuetan erabiltzen direlako edota Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) domeinuekin ez dutelako loturarik.

  Adibidez, gaztelaniazko lexikoian agertzen den *provinciano* (−1) hitza, *atzeratu* moduan itzulita zuzena eta erabilgarria izango litzateke guretzat; izan ere, mespretxuzko esanahi bat da eta horiek erabilgarriak dira sentimenduen analisian (Sorby, 2008). Hala ere, *Elhuyar* hiztegia (Elhuyar, 2013) erabilita *Araba, Bizkaia edo Gipuzkoako herritar (gipuzkoarra bereziki)* izan da lortu den ordainetako bat. Ordain hori gure lanerako ez da erabilgarria. Horrenbestez, ordain hori ez dugu aintzakotzat hartu.

- Hizkuntza figuratiboa. Aurkitu dugun beste ordain-mota bat hizkuntza figuratiboa da. Bertan, hitzek edo hitz-multzoek ez dute berez duten esanahia. Hiztunak esanahia moldatu egiten du, bigarren mailako esanahia emanez (Ghosh *et al.*, 2015). Bigarren mailako esanahi horiek zuzenak dira, baina testuinguru jakin eta mugatu batean erabiltzen direnez, horrelako ordainak kontuan ez hartzea erabaki dugu.

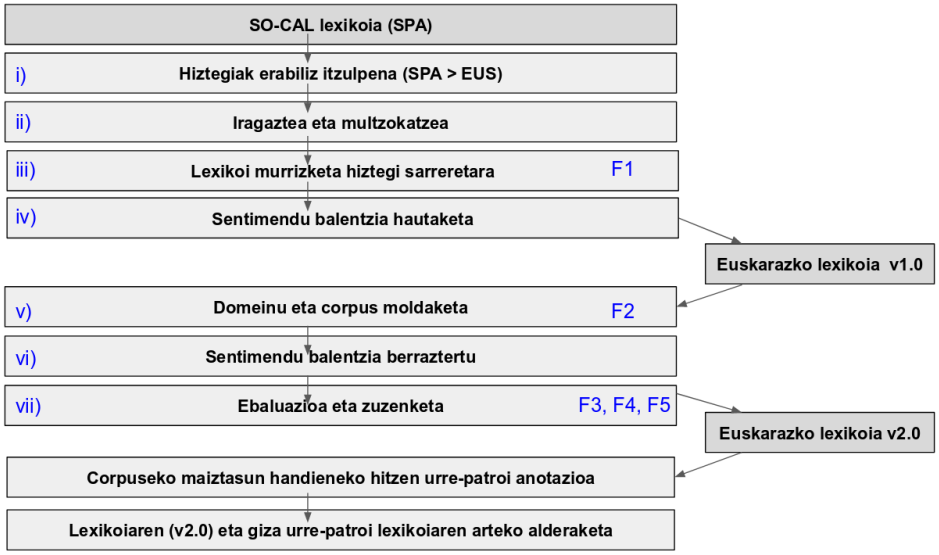
  Adibidez, gaztelaniazko lexikoiko hitzari euskarazko ordaina ematean *beltza* (−2) azaldu da. Hitz horrek bi esanahi ditu: i) beltza, kolorea eta ii) tristea, zoritxarrezkoa. Azken hori esanahi figuratiboa da, bigarren mailako esanahia. Hitzari sentimendu-balentzia aukeratzeko urratsean, euskarazko ordainak gaztelaniazko hitzaren −2 balentzia du jaso du; gaztelaniazko hitzak esanahi figuratiboari egiten diolako erreferentzia. Nahiz eta, hori egokia izan arren, gure lexikoian ez sartzea erabaki dugu: alde batetik, hitzaren erabilera hori ez delako ohiko esanahiarena baino handiagoa eta, beste aldetik, gure corpusera begira ere ez da egokia.

4.2.2.2. Gaztelaniazko lexikoiko hitzari euskarazko ordaina ematean egondako kasuistikak

SO-CAL tresnaren (Taboada *et al.*, 2011) gaztelaniazko lexikoia euskaratzerakoan, 3.1.2 ataleko metodologiari jarraitu dugu. Hala ere, gaztelaniazko hitzei euskarazko ordaina ematean, gaztelaniazko hitz guztiek ez dute amaiera bera izan. Hau da, euskarazko ordaina eman beharreko hitzaren ezaugarrien arabera, urrats gehiago edo gutxiago bete behar izan dira gaztelaniazko hitzei ordaina emateko.

4.3 irudian azaltzen dira SO-CALen gaztelaniako bertsioko lexikoiko hitzei euskarazko ordainak emateko urratsak[6] eta euskarazko ordaina ematean agertu diren bost fenomenoak. "Fenomeno"[7] hitza aipatzean, gaztelaniazko lexikoiko hitzek urratsei dagokienean hasiera bera, baina amaiera ezberdina izan dutela adierazi nahi dugu. F1-F5 bidez dago adierazita euskarazko ordainetako bakoitzak non bukatzen duen bere ibilbidea.

[6] 4.3 irudian, urratsak zenbaki erromatarren bidez adierazita daude.
[7] 4.3 irudian, fenomenoak Fren bidez adierazita daude.



4.3 irudia: Gaztelaniazko lexikoiko hitzei euskarazko ordaina emateko gauzatutako urratsak.

Jarraian, bost fenomeno horietako bakoitza esplikatuko dugu egitura honi jarraituz: lehenik eta behin, fenomenoa zertan den azalduko dugu; ondoren, fenomeno horren atzean dagoen arrazoia aditzera emango dugu eta, azkenik, fenomeno horren adibide bat emango dugu.

- Hiztegietako sarrerak ez diren euskarazko ordainak (F1).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu da, baina ordaina ez denez *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat ez dugu aintzakotzat hartu.
  - Arrazoia. Sortzen ari garen lexikoiko sarrerek ezaugarri berak izatea nahi dugu eta, horregatik, lexikoian zer sartu erabakitzerakoan muga bat jarri behar izan dugu eta muga hori euskarazko ordaina hiztegiko sarrera bat izatea da.
  - Adibidea. Gaztelaniako *desacreditar* hitzari hiru euskarazko ordain (*ospea_kendu*, *izena_kendu* eta *sona_kendu*) eman dizkiogu eta

ordainetako bakoitzak jatorrizko hitzaren balentzia −2 oinordekotzan hartu du. Baina, ordain horiek ez dira hiztegietako sarrerak eta, horrenbestez, lexikoitik kendu egin ditugu.

- Corpusean agertzen ez diren euskarazko ordainak (F2).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, euskarazko ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, baina ordaina ez da Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen. Horrenbestez, hitz hori ez dugu kontuan hartu euskarazko lexikoiaren bigarren bertsioan[8].
  - Arrazoia. Corpuseko domeinuekin loturarik ez duten hitzak lexikoitik kendu egin ditugu. Modu horretan, lexikoiak corpusarekiko koherentzia izatea nahi dugu. Erabaki horren ondorioz, lexikoia lehen bertsioan 8.140 sarrera izatetik bigarren bertsioan 1.813 sarrera izatera igaro da.
  - Adibidea. Esaterako, *atrofiatu* hitzak −1 balentzia du, baina hitz hori ez da Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) azaltzen. Izan ere, aztertzen ari garen hitza ez da corpusean dauden domeinuetan ohikoa. Hitza corpusean agertzen ez denez, lexikoitik kendu egin dugu; (corpuseko) domenuei loturiko lexikoia eratu nahi baitugu.
- Balentziaren egokitasuna zalantzazkoa duten euskarazko ordainak (F3).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzei euskarazko ordaina eman diegu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da eta ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016)

[8] 4.3 irudian, euskarazko lexikoiaren lehen bertsioa iv) eta v) urratsen artean ageri da eta bigarren bertsioa vii) urratsaren ondoren. Euskarazko lexikoiaren lehen bertsioan, gaztelaniako lexikoiko hitz guztiei eman diegu euskarazko ordaina, baina euskarazko lexikoiaren bigarren bertsioan, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen ez diren euskarazko ordainak kendu ditugu.

agertzen da. Baina, ordainak duen balentziaren egokitasuna zalantzazkoa da eta zalantza argitzeko SO-CAL tresnaren (Taboada *et al.*, 2011) ingeleseko bertsioaren lexikoian agertzen den egiaztatu da eta ez da azaltzen. Ondorioz, lexikoiaren bigarren bertsioan, hitz hori kendu egin dugu, baina lehen bertsioan mantentzen dugu.

- Arrazoia. Euskarazko sarreraren baliokiderik ez dago SO-CALen (Taboada *et al.*, 2011) ingelesezko lexikoian eta gaztelaniazko bertsiotik lortutako balentzia ez da egokitzat jotzen, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden domeinuak kontuan hartuta. Hori dela eta, euskarazko sarrera lexikoitik kentzea erabaki dugu.
- Adibidea. Fenomeno horren adibidea da *seinale* hitza. SO-CALen gaztelaniako lexikoitik (Brooke *et al.*, 2009) −1 balioa hartu du zaio, baina ez dugu egokitzat jo, nahiz eta balentzia intentsitate txikikoa izan. Arrazoi horregatik, hitza lexikoitik kentzea erabaki dugu.

- Arrazoi soziokulturalengatik aintzat hartutako euskarazko ordainak (F4).
  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen da, baina ordain hori ez da SO-CALen (Taboada *et al.*, 2011) ingeleseko lexikoiko hitz bat. Kalitatea zalantzazkoa ez denez, nahiz eta ingeleseko lexikoian ez agertu, lexikoiaren lehen bertsioan nahiz bigarrenean jaso dugu.
  - Arrazoia. Euskarazko sarreraren baliokiderik ez dago SO-CALen (Taboada *et al.*, 2011) ingeleseko lexikoian eta gaztelaniazko bertsiotik lortutako balentzia egokitzat jotzen da, Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) dauden domeinuak aintzakotzat hartuz. Hau da, hitz hori ingelesean ez da hain ohikoa arrazoi soziokulturalengatik, baina bai gaztelanian. Horregatik eta kali-

tatea zalantzazkoa ez denez, euskarazko ordaina aintzakotzat hartzen da. Ondorioz, hitz bera aurreko fenomenoaren egoera berean egon arren, emaitza ezberdina da eta hitza lexikoian mantentzea erabaki dugu.

- Adibidea. Esaterako, euskarazko lexikoiko sarrera bat *frankismo* da eta −2 balioa hartu du gaztelaniazko bertsiotik. Hitz hori espainiar politikari lotutakoa da eta, horregatik, gaztelaniazko bertsioan agertzen da, baina ingelesezkoan ez. Hitzaren ezaugarri hori kontuan hartu dugu eta, horregatik, euskarazko lexikoian mantentzea erabaki dugu.

- Gaztelaniazko eta ingelesezko lexikoietatik eta corpusetik elikatutako euskarazko ordainak (F5).

  - Fenomenoa. SO-CAL tresnaren gaztelaniako lexikoiko (Brooke *et al.*, 2009) hitzari euskarazko ordaina eman diogu, ordaina *Elhuyar* (Elhuyar, 2013) edota *Zehazki* (Sarasola, 2005) hiztegietako sarrera bat da, ordaina Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) agertzen da, baita SO-CALen (Taboada *et al.*, 2011) ingelesezko lexikoian ere. Egoera horretan, hitza euskarazko lexikoiaren bi bertsioetan dago.
  - Arrazoia. Euskarazko sarreraren baliokidea badago ingelesezko bertsioan ez ezik, gaztelaniazkoan ere badago, eta euskarazkoari balentzia esleitzerakoan ingelesezko bertsiokoa hobesten da.
  - Adibidea. 3.10 taulako kasuan, *zuzen* hitzak gaztelaniazkotik +3 balentzia oinordekotzan hartu du, eta ingelesekoan bi aukera daude: *right* +1 balentziarekin eta *correct* +3 balentziarekin. Gaztelaniazko hitzak (*correcto*) eta ingelesezko hitzak (*correct*) balentzia bera dutenez (+3), *zuzen* hitzak gaztelaniatik oinordekotzan hartutako balentzia mantentzea jo dugu egokitzat.

4.2.2.3. Euskarazko ordainari balentzia aukeratzeko irizpideak

Euskarazko ordainari dagokion sentimendu-balentzia eta gaztelaniako esanahia aukeratzerakoan, jarraian azaltzen den prozedura jarraitu dugu.

- Prozedura 1.
  - Azalpena. Baldin eta jatorrizko gaztelaniazko hitzak euskarazko ordain bakarra eta egokia badu, euskarazko ordainak haren sentimendu-balentzia hartuko du.
  - Adibidea. Hori gertatu da 2. eta 4. Fenomenoetan. Fenomeno horietan, ordain bakarra dago eta ordain bakar hori egokia da, ez delako ez esanahi figuratiboa, ez Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) domeinuekin bat ez datorren ordaina. Hori dela eta, *atrofiatu* hitzari −1 balentzia esleitu diogu eta *frankismo* hitzari −2 balentzia.
- Prozedura 2.
  - Azalpena. Baldin eta euskarazko ordainak jatorrian hainbat gaztelaniazko hitz eta haien balentziak baditu, hitzari esleituko diogun sentimendu-balentzia (eta gaztelaniazko esanahia) Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) oinarrituta egingo dugu. Lan hori corpusean oinarrituta dagoenez, gaztelaniazko esanahien (eta sentimendu-balentzien) artetik aukeraketa egin beharko dugu eta aukeratutakoak bat etorri edo egokia izan beharko Euskarazko Iritzi Corpusean (Alkorta *et al.*, 2016) hitz hori agertzen den testuinguruarekin. Hori gauzatu ahal izateko hitz-gakoa testuinguruan (*Key Word in Context*, KWIC) erabili dugu.
  - Adibidea. Prozedura hori jarraitu dugu *erakargarri* hitzaren kasuan. Lexikoiko gaztelaniako hitzei euskarazko ordaina eman ondoren, *erakargarri* hitzaren jatorrian hitz eta sentimendu-balentziak hauek daude: *apasionante* +5, *apetecible* +2, *apetitoso* +2 , *atractivo* +2, *fascinante* +5, *glamuroso* +4, *goloso* +3, *interesante* +4, *irresistible* +4, *seductor* +1 eta *tentador* +2. KWIC teknikaren bidez, *erakargarri* hitzaren testuingurua corpusean behatu ondoren, +2 balentzia esleitu diogu, corpuseko testuinguruarekin bat gehien berak egiten duelako.
- Prozedura 3.

– Azalpena. Euskarazko lexikoaren lehen bertsioa euskaraz sortzerakoan, kasuak egon dira non gaztelaniazko hainbat hitzek euskaraz ordain bera duten, eta euskarazko ordain hori corpusean ez den agertzen. Egoera horretan, euskarazko ordainak esanahirik zabalena duen gaztelaniazko hitzaren sentimendu-balentzia esleitu diogu.

– Adibidea. *Bat_ez_datorrena* hitzaren kasua prozedura horren adibidea da. Egitura hori ez da corpusean agertzen; eta, beraz, bere aukeretan (*descontento* −3, *discordante* −2, *insatisfecho* −2) esanahirik erabilienaren (*discordante*) balentzia (−2) hautatu dugu.

4.2.3. Euskarazko sentimenduen lexikoiaren ebaluazioa

Hasteko, 4.2.3.1 azpiatalean, ebaluazioa egiteko prozedura aurkeztuko dugu. Ondoren, 4.2.3.2 azpiatalean, ebaluazio horren emaitzak azalduko ditugu.

4.2.3.1. Ebaluazioa egiteko prozedura

Euskarazko sentimenduen lexikoia ebaluatzeko honako baliabideak erabili ditugu:

- 400 hitzeko zerrenda. *Analhitza* (Otegi *et al.*, 2017) tresna baliatuz, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) gramatika-kategoria bakoitzeko (izenak, adjektiboak, adberbioak eta aditzak) maiztasun handieneko 100 hitz lortu ditugu.
- Urre-patroia. Horretarako, bi etiketatzailek 400 hitzei banan-banan −5 eta +5 arteko sentimendu-balentzia esleitu die.

Euskarazko sentimendu lexikoiak zerrendako 400 hitzi sentimendu-balentzia esleitu die automatikoki eta, ondoren, sentimendu-balentzia horiek urre-patroiko sentimendu-balentziekin alderatu dira. Bi horien arteko adostasuna neurtzeko Pearson korrelazioa (Benesty *et al.*, 2009) erabiltzea erabaki dugu. Etiketatzaile edo lexikoi batek hitz bati −3 balentzia esleitzen dion bitartean, beste etiketatzaile edo lexikoi −1 edo +1 balentzia esleitzen badio, Pearson korrelazioan oinarrituta aintzat hartuko dugu −1 balentziak +1 balentzia baino

hurbiltasun handiagoa duela eta, horrenbestez, bi etiketatzeren arteko adostasuna handiagoa dela. Gainera, kalitatearen neurketa bi modutan egitea erabaki dugu jarraian azaltzen diren bi arrazoiengatik:

- Pearson 1: bi etiketatzaileek etiketatutako hitzak bakarrik kontuan hartuta egindako korrelazio neurketa da. Beraz, 400 hitz horietatik ez dira guztiak erabili neurketa egiteko mementoan.

  4.10 taulan, Pearson 1 deitu dugun korrelazioaren neurketa nola egin den ikus daiteke. Taulan 10 hitz daude, baina horietako hiru (*ageri*, *ar* eta *bakoitza*) ez daude bi etiketatzaileek etiketatuta. Horrenbestez, Pearson korrelazioaren koefizientea kalkulatzean, hiru hitz horiek ez dira kontuan hartu.

Adjektiboak	**Eti1**	**Etik2**	**Etik1**	**Etik2**
ageri	POS		1	
ahul	NEG	NEG	3	3
antisozial	NEG	NEG	1	5
apur	NEG	NEG	1	1
ar				
argi	POS	POS	2	3
aspergarri	POS	NEG	3	3
ausart	POS	POS	3	4
bakar	NEG	POS	1	5
bakoitz				

4.10 taula: Bi etiketatzaileren arteko korrelazioaren kalkulua, etiketatu gabeko hitzak aintzat hartu gabe (Pearson 1).

- Pearson 2: corpusetik ateratako hitz guztiak, 400 hitz guztira, erabili dira. Batzuetan, bi anotatzaileek hitzei sentimendu-balentzia jarri diete. Beste kasu batzuetan, aldiz, anotatzaileak edo anotatzaileek ez dio(te) hitzari sentimendu-balentzia jarri, haien ustez ez duelako sentimendu-balentziarik behar. Horrelako kasuetan, anotatu gabeko hitzei 0 sentimendu-balentzia esleitu diegu. Hori da Pearson 1ekiko ezberdintasun bakarra, anotatu gabeko hitzei 0 sentimendu-balentzia esleitzea. Modu horretan, hitz horiek ere neurketan erabili ditugu.

4.11 taulan, Pearson 2 deitu dugun korrelazioaren neurketa nola egin den azaltzen da. 4.10 taulan bezala, anotatzaileek hiru hitzi (*ageri*, *ar* eta *bakoitza*) ez diete sentimendu-balentzia esleitu. Baina horiek ere neurketan kontuan hartu nahi ditugu eta, horregatik, 0 sentimendu-balentzia esleitu diegu.

Adjektiboak	**Etik1**	**Etik2**	**Etik1**	**Etik2**
ageri	POS	0	1	0
ahul	NEG	NEG	3	3
antisozial	NEG	NEG	1	5
apur	NEG	NEG	1	1
ar	0	0	0	0
argi	POS	POS	2	3
aspergarri	POS	NEG	3	3
ausart	POS	POS	3	4
bakar	NEG	POS	1	5
bakoitz	0	0	0	0

4.11 taula: Bi etiketatzaileren arteko korrelazioaren kalkulua, etiketatu gabeko hitzak ere aintzat hartuz (Pearson 2).

4.2.3.2. Ebaluazioaren emaitza

Jarraian, egin ditugun bi ebaluazioak aurkeztuko dugu. Lehenik eta behin, bi etiketatzaileen arteko korrelazioa azalduko dugu. Ondoren, urre-patroiaren eta euskarazko sentimenduen lexikoiaren arteko korrelazioaz arituko gara. Zerrendako hitzei sentimendu-balentzia esleitzerakoan dagoen bi pertsonen arteko korrelazioa hurrengo 4.12 taulan ikus daiteke.

Gramatika-kategoria	**Pearson 1**	**Pearson 2**
Izena	0,87	0,59
Adjektiboak	0,71	0,60
Aditzondoak	0,93	0,82
Aditzak	0,87	0,76
Guztira	**0,79**	**0,73**

4.12 taula: Bi anotatzaileen arteko Pearson korrelazioaren neurketa.

Pearson 1 balioak erakusten korrelazio-koefizientea altua dela (0,79). Horrek bi anotatzaileek zerrendako hitz askotan hitzei esleitu dieten sentimendu-balentzia nahiko antzekoa dela esan nahi du. Gramatika-kategorien artean badaude ezberdintasun batzuk, korrelaziorik altuena 0,93koa delako eta baxuena, aldiz, 0,71koa. Korrelaziorik altuena aditzondoei dagokie eta baxuena adjektiboei. Pearson 2ri dagokionez, emaitzek erakusten dutenez, korrelazio-koefizienteak altua izaten jarraitzen du, nahiz eta Pearson 1ekin alderatuta zerbait baxuagoak diren. Pearson 2ren kasuan, gramatika-kategorietan koefizienteak 0,82 eta 0,60 artean kokatzen dira.

Kontigentzia-taulak emaitzak interpretatzeko informazio osagarria ematen digu 4.13 taulan. Kontigentzia-taula hori Pearson 2rena da; izan ere, zutabe eta zerrendetan 0 agertzen da, eta horrek adierazten du anotatzaileek hitzari ez diotela sentimendu-balentzia esleitu. Bertan ikusten denez, bi anotatzaileen arteko ezberdintasuna nagusiki egoera batean gertatzen da: anotatzaile batek hitzari balentzia esleitzen dionean besteak ez dio esleitzen, eta alderantziz. Egoera hori desadostasun guztien % 90,19ren jatorrian dago (ezberdintasuna dagoen 102 instantzietatik 92etan, hain zuzen ere).

Guztira kategoriak			
	0	**Negatiboa**	**Positiboa**
0	187	12	27
Negatiboa	14	42	5
Positiboa	39	5	69

4.13 taula: Bi anotatzaileen arteko kontigentzia-taula.

4.14 taulan, berriz, sentimenduen lexikoiaren eta urre-patroiaren arteko korrelazio-koefizienteak ageri dira. Urre-patroiaren eta sentimenduen lexikoiaren arteko korrelazio-koefizienteen neurketak ezberdintasun batzuk erakusten ditu aurretik egindako bi anotatzaileen arteko korrelazio koefizientearekiko.

Pearson 1i dagokionez, hau da, sentimendu-balentzia esleitu zaien hitzak bakarrik kontuan hartuta, korrelazio-koefizientea aurreko neurketatik gertu dago. Kasu horretan, koefizientea 0,76 da eta aurreko neurketan 0,79. Koefizientean ezberdintasun handiak daude gramatika-kategorien artean. Koefizienterik altuena izenetan dago (0,96) eta baxuena, berriz, aditzetan (0,69).

Gramatika-kategoria	Pearson 1	Pearson 2
Izena	0,96	0,59
Adjektiboak	0,78	0,56
Aditzondoak	0,75	0,47
Aditzak	0,69	0,54
Guztira	**0,76**	**0,54**

4.14 taula: Euskarazko sentimenduen lexikoa (V2.0) eta urre-patroiaren arteko Pearson korrelazioaren neurketa.

Baina, ezberdintasun esangurantsua Pearson 2n gertatzen da, bi anotatzaileen arteko korrelazio koefizientearekin alderatzen denean. Pearson 2an koefizientea 0,54 izan da; aurreko neurketan 0,73 izan denean (0,19ko ezberdintasuna). Koefizienterik altuena izenenak izaten jarraitzen du (0,59) eta baxuena, berriz, aditzondoena da (0,47).

Datu hauen gure interpretazioa hauxe da: lexikoiaren eta urre-patroiaren arteko Pearson 1 koefizienteak altua izaten jarraitzen du zerrendako hitzei esleitu zaien balentzia antzekoa izan delako bi kasuetan, batean, bi anotatzaileen artean eta bestean, urre-patroia eta lexikoiaren artean. Baina, ezberdintasuak daude balentzia esleitu zaien zerrendako hitzetan. Bi anotatzaileen kasuan, balentzia esleitu zaien hitzak kasik berak izan dira. Hemen, urre-patroiaren eta lexikoiaren arteko emaitzak kontuan hartuz, balentzia esleitu zaien hitzak beti ez dira berak izan.

Kontigentzia-taulak lexikoiaren eta urre-patroiaren arteko ezberdintasunak argitara ematen ditu 4.15 taulan. Bi etiketatzaileen artean ezberdintasunak balentzia esleitu beharreko hitzaren inguruan daude. Batek hitz bati balentzia esleitzen dionean, besteak ez egitea, eta alderantzizko kasua. Lexikoiak edo urre-patroiak hitz bati balentzia esleitzean besteak balentzia ez esleitzea ezberdintasun guztien % 89,83 (118tik 106 kasutan) izan da. Bi anotatzaileek hitzei sentimendu-balentzia esleitzean ere gertatu da anotatzaile batek hitzari balentzia esleitzean beste anotatzaileak ez esleitzea, baina askoz intentsitate baxuago batean, kasu horretan ez bezala. Baina, bi pertsonen arteko emaitzak alderatuta badago beste ezberdintasun bat. Bi pertsonen artean, batek balentzia esleitzean besteak ez egitea, eta alderantziz, nahiko

Guztira kategoriak			
	0	**Negatiboa**	**Positiboa**
0	195	2	15
Negatiboa	30	34	8
Positiboa	59	4	53

4.15 taula: Euskarazko sentimenduen lexikoa (V2.0) eta urre-patroiaren arteko kontigentzia-taula.

orekatua izan da. Hemen, ordea, joera bat nagusitzen da: kasu gehienetan, batek (urre-patroiak) zerrendako hitzari sentimendu-balentzia esleitzen dionean, besteak (lexikoiak) ez du egiten. Horrek lexikoia urre-patroiaren aldean kontserbadoreagoa eta zorrotzagoa dela aditzera ematen du. Lexikoiak askozaz ere hitz gutxiagori esleitzen dien balentzia, Pearson 2ko korrelazio koefizientea baxuagoa da.

Laburbilduz, Pearson 1 korrelazio-koefizientea altua eta antzekoa izan da bai bi anotatzaileen artean, bai lexikoiaren eta urre-patroiaren artean. Horrek hitzei esleitu zaien sentimendu-balentzia (−5etik +5era) antzekoa (0,79 vs. 0,76) izan dela esan nahi du. Pearson 2 korrelazio koefizientea, aldiz, altua izan da bi anotatzaileen artean, baina jaitsiera bat egon da lexikoiaren eta urre-patroiaren artean (0,73 vs 0,54). Lexikoiak urre-patroiarekin (eta bi anotzaileekin) alderatuz, hitz gutxiagori esleitu die sentimendu-balentzia eta, horregatik, da koefizientea baxuagoa.

1 Ikerketa hipotesia

1.2 ataleko ikerketa hipotesiari erantzunez, emaitzek erakusten dute posible dela gaztelaniatik euskara sentimenduen lexikoi bat itzultzea eta bera baliagarria izatea.
Pearson korrelazioek erakusten dutenez, itzulpenaren ondoren, urre-patroiaren eta sentimenduen lexikoiaren adostasuna 0,76 koefizientekoa hitzei sentimenduen balentzia esleitzekoan. Hala ere, badaude hobetzeko alderdi batzuk, besteak beste, garaturiko sentimenduen lexikoian sarrera gehiago gehitzea falta delako.

4.3 Euskarazko sentimenduen sailkatzailea

Lehenik eta behin, guk sortu dugun sentimenduen sailkatzailearen oinarrian dagoen SO-CAL tresnaren ezaugarriak deskribatuko ditugu (4.3.1 atala). Ondoren, euskarazko sentimenduen sailkatzailea sortzeko SO-CALen egindako moldaketak azalduko ditugu (4.3.2 atala). Azkenik, euskarazko sentimenduen sailkatzailearen ebaluazioaren emaitzak aurkeztuko ditugu (4.3.3 atala).

4.3.1. SO-CAL izeneko sentimenduen sailkatzailearen ezaugarriak

Guk sortu dugun euskarazko sentimenduen sailkatzailearen oinarrian SO-CAL tresna (Taboada *et al.*, 2011) dago. Tresna hori hasiera batean ingeleserako sortu bazen ere; gaur egun beste hizkuntzetako bertsioak ere badaude, esaterako txinerarena (Miao *et al.*, 2013).

Tresna horren oinarrian sentimenduen lexikoia dago. Sentimenduen lexikoiak lau gramatika-kategorietako hitzak (izenak, adjektiboak, aditzak eta adberbioak) biltzen ditu. Hitz horiek −5 eta +5 arte sentimendu-balentzia dute, 4.16 taulan ikusten den moduan.

Gramatika-kategoria horretako hitzez gain, intentsifikatzaileak ere badaude eta horiek hitzen sentimendu-balentzian aldaketak eragiten dituzte. Intentsifikatzaileen hitzen sentimendu-balentzia % +100 eta % −50 artean indartu edo ahuldu dezakete 4.17 taulan, intentsifikatzaile horietako batzuk ageri dira.

Intentsifikatzaileen eta sentimendu-balentziadun hitzen artean (4.17) adibidean bezalako eragiketako gertatzen dira. *Sleazy* ("zikin") hitzak −3 sentimendu-balentzia du baina *somewhat* ("zerbait") intentsifikatzailearen ondorioz, bere sentimendu-balentzia −3tik −2, 1era jaisten da.

(44) Sleazy: (−3), somewhat sleazy: −3 x (%100 − %30)= −2, 1. (Taboada *et al.*, 2011, 275 orr.)

SO-CAL tresnak (Taboada *et al.*, 2011) ezeztapena ere tratatzen du eta, ho-

Hitza	Sentimendu-balentzia
monstrosity	-5
hate (izena eta aditza)	-4
disgust	-3
sham	-3
fabricate	-2
delay (izena eta aditza)	-1
determination	+1
inspire	+2
inspiration	+2
endear	+3
relish (aditza)	+4
masterpiece	+5

4.16 taula: SO-CAL tresnako lexikoiaren zenbait sarrera (Taboada *et al.*, 2011).

Intentsifikatzailea	Modifikatzailea (%)
slightly	-50
somewhat	-30
pretty	-10
really	+15
very	+25
extraordinarily	+50
(the) most	+100

4.17 taula: SO-CAL tresnako zenbait intentsifikatzaile (Taboada *et al.*, 2011).

rretarako, 5.2.2 atalean azaldutako desplazamendu-ezeztapena izeneko hurbilpena erabiltzen du. (45) adibidean, tresnak ezeztapena nola laguntzen duen ikus daiteke. Ezeztapenak −4 balentzia du esaldiak orientazio semantiko positiboa duenean eta −4 esaldiak orientazio semantiko negatiboa duenean. (45) adibideko lehen zatian *terrific* (“‘bikaina”) hizak −5 sentimendu-balentzia du baina ezeztapenak sentimendu-balentzia hori +1era aldatzen du. Adibide bereko bigarren zatian, aldiz, *terrible* (“beldurgarria”) hitzak −5 sentimendu-balentzia du baina ezeztapenak balentzia hori +1ean bihur-

tzen du.

(45) He's not terrific (5 − 4 = 1) but not terrible (−5 + 4 = −1) either. (Taboada *et al.*, 2011, 277 orr.)

Sentimendu lexikoiaren beste ezaugarrietako bat da sentimenduen analisirako baliagarriak ez diren esaldiak identifikatzea eta horiei sentimendu-balentzia ez esleitzea. Sentimenduen analisirako baliagarriak ez diren esaldiei *irrealis* edo testuinguru ez-faktualak deritze. Ingelesean, testuinguru ez-faktuala hitzen hurrenkeraren, aditz modalaren edota aginteraren bidez agertzen da. SO-CAL tresnak (Taboada *et al.*, 2011) horrelako elementuak aurkitzen dituenean, esaldi horietako hitzen sentimendu-balentzia ez du kontuan hartzen. Komatxoen artean agertzen diren hitzen sentimendu-balentziak ere ez ditu kontuan hartzen.

Esaterako, (46) adibidean, *great* ("sekulako") hitzak +3 sentimendu-balentzia du baina esaldiak testuinguru ez-faktuala duenez, bertako hitzen sentimendu-balentzia ez aintzat hartzen.

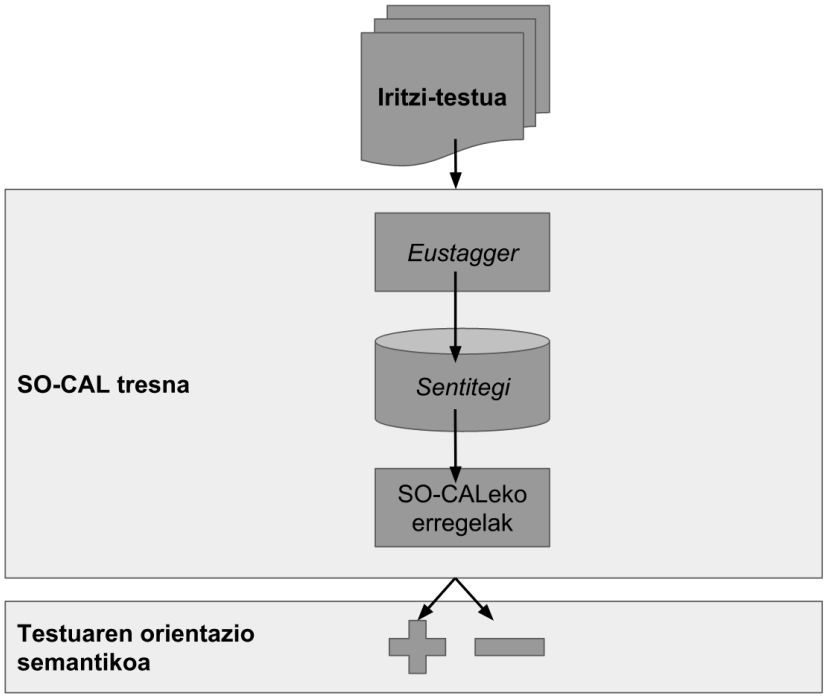
(46) This should have been a great movie. (+3 = 0). (Taboada *et al.*, 2011, 279 orr.)

Kennedy eta Inkpenek (2006) lexikoian oinarritutako sentimenduen sailkatzaileek orientazio semantiko positiboa esleitzeko joera antzeman zuten eta, horregatik, tresna honetan sentimendu-balentzia negatiboa duten hitzei pisua esleitu zaie. Hain zuzen, % +50eko pisua esleitzen zaie hitz horiei. Gainera, sentimendu-balentziadun hitz bat testu batean hainbat aldiz errepikatzen bada, errepikapen horietako bakoitzean sentimendu-balentziadun hitz horrek bere balentzia galtzen du ($1/n$ eragiketa, n= errepikapen-kopurua). Erregela horren adibidea euskaran 4.7 irudian ikus daiteke.

Azkenik, SO-CAL tresnaren (Taboada *et al.*, 2011) ingelesezko bertsioak zenbait erregela morfologiko ditu, testuetako hitzei flexio-morfemak kendu eta lema moduan uzten dituetan. Hitza lematizatuta dagoenez, tresnako lexikoiak hitzari sentimendu-balentzia esleitu diezaioke, baldin eta hitza lexikoian badago.

4.3.2. Sentimenduen sailkatzailaren arkitektura

SO-CAL izeneko sentimenduen sailkatzailearen (Taboada *et al.*, 2011) euskarazko bertsioa garatzeko zenbait aldaketa egin behar izan ditugu. Aldaketa horiek 4.4 irudian daude ikusgai.



4.4 irudia: SO-CAL sentimenduen sailkatzailearen euskarazko bertsioa.

Euskarazko bertsioa sortzeko egin dugun lehen aldaketa *Eustagger* (Aduriz *et al.*, 2003) integratzea izan da. Ingelesak, hizkuntza analitikoa denez, erregela morfologiko gutxi behar ditu flexio-morfemak kendu eta hitzak lematizatuta uzteko. Euskara, ordea, morfologikoki aberatsa da eta, horregatik, testuetako hitzak lematizatzeko *Eustagger* (Aduriz *et al.*, 2003) integratu dugu.

Gainera, ingelesezko sentimenduen lexikoiaren ordez, guk sortu dugun *Sentitegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) integratu dugu. Gure sentimenduen lexikoiak izen, adjektibo, aditz eta adberbio gramatika-

kategorietakoak biltzen ditu.

Azkenik, SO-CAL tresnak (Taboada *et al.*, 2011) berak dituen bestelako erregelak bere horretan utzi ditugu. Erregela horiek sentimendu-balentzia negatiboa duen hitzei pisua esleitzekoa, testu batean hainbat aldiz errepikatzen diren hitzei balioa kentzekoa edota komatxoen artean dauden hitzen sentimendu-balentziak kontuan ez hartzekoak dira.

Euskarazko sentimenduen lexikoia garatzeko aldaketa horiek egin ostean, jarraian azalduko dugun moduan ibiltzen da euskarazko sentimenduen lexikoia. 4.5 irudian, euskarazko sentimenduen sailkatzaileak aztertuko duen iritzi-testua ageri da. Ikusten den moduan, testua gordina da.

```
Asteburu polita orokorrean.
Ez dugu Penintsula hegoaldean dagoen beroa eta ezta nahi ere. Tenperatura
    atseginekin doa uda eta horrelaxe jarraituko dugu asteburuan ere.
    Gainera, oro har eguraldia ondo portatuko da. Hala ere, larunbata
    hodeiekin hasiko dugu eta baliteke zirimiri pixka bat egitea ere,
    bereziki goizaldean. Arratsalderako ordea, argituko du eta eguzkia
    ikusiko dugu. Iparraldeko haizea ibiliko da eta tenperatura maximoa 21 C
    -koa izango da, atsegina.
Igandean eguzkia nagusituko da berriro ere, baina udan askotan gertatzen den
     bezala, goizaldean hodei baxuak sortuko dira seguruenik eta hodei
    hoiekin esnatzea posible da, baina goizean zehar eta larunbatean baino
    azkarrago, ostarteak zabaltzen hasiko dira, eguraldi eguzkitsua geratuz
    eguerditik aurrera. Tenperatura pixka bat igo egingo da, maximoa 24 C-ra
     iritsiz.
Esandakoa gure betiko uda naturalarekin jarraituko dugu: hodei baxuak,
    eguzkia eta tenperatura atseginak, berorik gabe.
```

4.5 irudia: EGU40 iritzi-testua.

Euskarazko sentimenduen sailkatzaileari 4.5 irudiko iritzi-testua ematean, tresnak lehendabizi testua lematizatzen du, eta ondoren, bertako hitzei sentimendu-balentzia esleitzen die. 4.6 irudian, iritzi-testua lematizatuta dago eta esaldien eskuinetara esaldietako hitzei esleitu dien sentimendu-balentzia agertzen da.

Hurrengo urratsean, sentimendu-balentzia esleitu dien hitzei zenbait erregela aplikatzen dizkie, hala nola sentimendu-balentzia negatiboa duten hitzei pisua esleitzekoa, hainbat aldiz errepikatzen diren hitzei balioa kentzekoa eta komatxoen artean dauden sentimendu-balentziadun hitzak kontuan ez hartzekoa.

```
asteburu polit orokor . 7.0
tenperatura atsegin joan uda eta horrelaxe jarraitu *edun asteburu ere . 3.3
gain , oro har eguraldi ondo portatu izan . -1.5
hala ere , larunbat hodei hasi *edun eta *edin zirimiri pixka bat egin ere ,
     bereziki goizalde . 0
arratsalde ordea , argitu *edun eta eguzki ikusi *edun . 2.0
iparraldeko haize ibili izan eta tenperatura maximo 21 C-koa izan izan ,
    atsegin . 4.0
igande eguzki nagusitu izan berriro ere , baina uda asko gertatu izan bezala
     , goizalde hodei baxu sortu izan seguru eta hodei hoiekin esnatu
    posible izan , baina goiz zehar eta larunbat baino azkar , ostarte
    zabaldu hasi izan , eguraldi eguzkitsu geratu eguerdi aurrera . 5.8
tenperatura pixka bat igo egin izan , maximo 24 C-ra iritsi . 0.5
esan gu be uda natural jarraitu *edun : hodei baxu , eguzki eta tenperatu
    atsegin , bero gabe . 1.75
```

4.6 irudia: EGU40 iritzi-testua lematizatuta eta hitzei sentimendu-balentziak esleituta.

4.7 irudian, euskarazko sentimenduen sailkatzaileak iritzi-testuko sentimendu-balentziadun hitzei aplikatutak erregelak ikus daitezke. Ikusten denez, zenbait hitzek sentimendu-balentzia negatiboa dute eta horiei pisua esleitu die (NEGATIVE erregela). Ondorioz, esaterako, adjektiboetan *baxu* hitzaren sentimendu-balentzia -3tik $-4,5$era pasa da. Beste erregela batek, testu berean hainbat aldiz errepikatzen diren hitzei balioa kentzen die. *Atsegin* horren adibidea da. Lehen aldiz agertu denez, tresnak *atsegin* hitzari $+1$ sentimendu-balentzia esleitu dio, baina bigarren aldiz agertu denean, hitzari pisua kendu dio (REPEATED erregela) eta bere sentimendu-balentzia $+1$etik $+0,5$era igaro da.

Sentimendu-balentziadun hitzei erregela horiek aplikatu ostean, tresnak iritzi-testuari sentimendu-balentzia esleitzen dio, 4.8 irudian ikusten den moduan. Kasu honetan, iritzi-testuaren sentimendu-balentzia $+1,04$ da; beraz, iritzi-testuak balorazio positiboa egiten du.

4.3.3. Sentimenduen sailkatzailearen ebaluazioa

Atal honetan, sentimenduen sailkatzailaren ebaluazioaren emaitzen berri emango dugu. Ebaluazioaren emaitzak hainbat modutara emango ditugu: asmatutako eta ez asmatutako testuen orientazio semantikoak, orientazio semantiko

```
Text Length: 168
---------
Nouns:
-----
har -1.0  X 1.5 (NEGATIVE) = -1.5
-----
Average SO: -1.5
-----
Verbs:
-----
jarraitu *edun asteburu ere 1.0 X 1.3 (INTENSIFIED)  = 1.3
argitu 2.0  = 2.0
ibili 2.0  = 2.0
atsegin 1.0  = 1.0
nagusitu izan berriro ere 1.0 X 1.3 (INTENSIFIED)  = 1.3
sortu 2.0  = 2.0
jarraitu 1.0 X 1/2 (REPEATED)  = 0.5
atsegin 1.0 X 1/2 (REPEATED)  = 0.5
-----
Average SO: 1.325
-----
Adjectives:
-----
polit 4.0  = 4.0
orokor 3.0  = 3.0
atsegin 2.0  = 2.0
maximo 1.0  = 1.0
baxu -3.0  X 1.5 (NEGATIVE) = -4.5
seguru 1.0  = 1.0
posible 1.0  = 1.0
azkar 2.0  = 2.0
eguzkitsu 3.0  = 3.0
maximo 1.0 X 1/2 (REPEATED)  = 0.5
natural 2.0  = 2.0
baxu -3.0 X 1/2 (REPEATED)  X 1.5 (NEGATIVE) = -2.25
bero 1.0  = 1.0
-----
Average SO: 1.05769230769
-----
```

4.7 irudia: EGU40 iritzi-testuko hitzei sentimenduen sailkatzailearen erregelak aplikatuta.

```
---------
Total SO: 1.03863636364
---------
```

4.8 irudia: EGU40 iritzi-testuaren sentimendu-balentzia.

jakin bakoitzean sailkatzaileak duen asmatze-tasa eta, azkenik, sailkatzaileak corpuseko domeinu bakoitzean duen asmatze-tasa.

4.18 Taulako emaitzen arabera, sailkatzailearen asmatze-tasa 0,71koa da. Izan ere, corpuseko test zatian dauden 48 testuetatik 34 iritzi-testuren orientazioa semantikoa ondo esleitu baitu. Iritzi-testuek duten orientazio semantikoaren ikuspegitik emaitzak aztertuta, emaitzek bestelako interpretazioak egiteko aukera ematen dute.

Emaitzak	**Testu-kopurua**	%
Ondo	34	0,71
Gaizki	14	0,29

4.18 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak.

4.19 taulako emaitzetan, tresnaren funtzionamenduari buruzko emaitza argigarriak ageri dira. Ikusten den moduan, sentimenduen sailkatzailearen asmatze-tasa orientazio positiboko iritzi-testuetan 1ekoa da. Emaitzak zeharo ezberdinak dira orientazio negatiboko iritzi-testuen kasuan. Mota horretako iritzi-testuekin, tresnaren asmatze-tasa 0,33koa da; izan ere, 24 orientazio negatiboko iritzi-testu egon arren, 8 iritzi-testu bakarrik identifikatu ditu negatibo moduan. Beraz, esan liteke, sentimenduen sailkatzaileak joera bat duela iritzi-testuei orientazio semantiko positiboa esleitzeko. Joera hori bat dator Alistair eta Dianak (2005) dioenarekin. Diotenez, lexikoian oinarritutako sentimenduen sailkatzaileek orientazio semantikoa esleitzeko joera erakusten du eta hori lotuta dago unibertsalki gizakiek hizkuntza positiboa erabiltzeko duten ohiturarekin Boucher eta Osgood (1969).

Testuen orientazio semantikoa	**Asmatutako testu-kopurua**	%
Positiboak	24/24	1
Negatiboak	24/8	0,33

4.19 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak iritzi-testuen orientazio semantikoa aintzat hartuta.

Azkenik, emaitzak domeinuen arabera aztertuta, domeinu guztietan asmatze-tasa antzekoa dela ikus daiteke, literaturan izan ezik.

4.20 taulak erakusten duenez, domeinuetan asmatze-tasa txikiena eguraldian, musikan eta politikan dago; asmatze-tasa 0,63koa da. Kirola eta zinema tartean kokatzen dira 0,75eko asmatze-tasarekin eta, azkenik, emaitza onenak literaturak lortzen ditu, testu baten orientazio semantikoa izan ezik, beste guztiena asmatu baitu tresnak bertan (0,88ko asmatze-tasa). Domeinutik domeinura ez dago ezberdintasun handirik, domeinu bakoitzean orientazio semantiko positibo eta negatiboko iritzi-testu kopuru bera dagoelako.

Domeinua	**Testu-kopurua**	**%**
Eguraldia	8/5	0,63
Musika	8/5	0,63
Politika	8/5	0,63
Kirola	8/6	0,75
Zinema	8/6	0,75
Literatura	8/7	0,88

4.20 taula: Sentimenduen sailkatzailearen ebaluazioaren emaitzak domeinuen arabera.

4.4 Laburpena

Kapitulu honetan, tesi-lanean garatutako baliabideak zein diren azaldu dugu. Lehenik eta behin, Euskarazko Iritzi Corpusa izeneko sei domeinutako 240 iritzi-testu biltzen dituen corpusa (Alkorta *et al.*, 2016) eratu dugula aipatu behar da.

Bigarrenik, sortu dugun *Sentitegi* izeneko euskarazko sentimenduen lexikoiaren (Alkorta *et al.*, 2018) berri ere eman dugu. Ikusi den moduan, sentimenduen lexikoiaren bi bertsio sortu ditugu, bata domeinuari lotu gabea (8.140 sarrerakoa) eta bestea, corpuseko domeinuei lotutakoa (1.237) eta izaera murriztaileagoa duena, kolokazioak eta adierazpenak ez baititu kontuan hartzen. Sorkuntzak emandakoa esplikatu ondoren, ebaluatu egin dugu. Horretarako, Euskarazko Iritzi Corpuseko maiztasun handieneko lehen 100 hitzak hartu ditugu, bi anotatzaileek hitzei sentimendu-balentzia esleitu die eta ondoren, bertatik urre-patroia sortu da. Jarraian, urre-patroi hori euskarazko sentimenduen lexikoiak zerrendako hitzei eman dien sentimendu-balentziarekin alderatu dugu. Emaitzek erakusten dutenez, sentimenduen lexikoiko hitzek duten sentimendu-balentzia egokia da (Pearson korrelazioa 0,76koa delako), baina berez sentimendu-balentzia eduki beharko luketen hitz batzuei lexikoiak ez die balentzia esleitu eta, ondorioz, Pearson korrelazioa 0,54koa da.

Azkenik, tesi-lan honen hirugarren emaitza euskarazko sentimenduen sailkatzailea da. Euskarazkoa sortzeko ingelesezko SO-CAL sentimenduen sailkatzailean oinarritu gara eta zenbait aldaketa egin behar izan ditugu. Esaterako, *Eustagger* testu lematizatzailea (Aduriz *et al.*, 2003) inplementatu dugu, baita *Sentitegi* euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018) ere. Sentimenduen sailkatzaile hori ebaluatu egin dugu eta, emaitzek erakusten dutenez, tresnak kasuen % 71etan iritzi-testuaren orientazio semantikoa ondo sailkatzen du.

BALENTZIA ALDATZAILEAK HIZKUNTZA MAILA EZBERDINETAN

5. KAPITULUA

Balentzia-aldatzaileak

Kapitulu honetan, hizkuntzako maila ezberdinetan dauden euskarazko balentzia-aldatzaileak identifikatzeko lanaren emaitzak azalduko ditugu. Guztira lau hizkuntza maila landu ditugu: fonologiko eta morfologikoa (5.1 atala), sintaktikoa (5.2 atala) eta diskurtsokoa (5.3 atala).

5.1 Fonologia eta morfologia maila

5.1.1 atalean, balentzia-aldatzaile fonologiko eta morfologikoen sailkapena hiru modutan aurkeztuko dugu. 5.1.2 atalean, hainbat balentzia-aldatzaile morfologiko dituzten hitzei buruz arituko gara, 5.1.3 atalean, eragin ezberdinetako balentzia-aldatzaileen eta orientazio semantiko ezberdinetako hitzen arteko konbinazioak azalduko ditugu. Azkenik, 5.1.4 atalean, maila fonologikoak eta morfologikoak euskaran duten garrantzia nabarmenduko dugu.

5.1.1. Balentzia-aldatzaileen sailkapena

5.1 taulan, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako balentzia-aldatzaile fonologiko eta morfologikoen zerrenda ageri da maiztasunean oinarrituz sailkatuta. Guztira 32 balentzia-aldatzaile aurkitu ditugu eta horiek guztiak 1.623 aldiz agertzen dira corpusean. Gehienak atzizkiak dira eta hiru bustidura adierazkor eta aurrizki mota ageri dira.

Balentzia-aldatzaileek sentimendu-balentzian duten eraginean oinarritutako sailkapena 5.2 taulan dago. Bertan, balentzia-aldatzaileak lau multzotan

Hizkia	**Maiztasuna**	**Adibidea**
-tasun	307	zailtasun
-garri	250	barregarri
-tsu	170	gatazkatsu
-en	163	handien
-gabe	71	akatsgabe
des-	62	desorekatu
-egi	61	bikainegi
-dun	60	berezidun
-gile	60	inbertsiogile
-keria	57	handikeria
-s-/-z- >-x-	39	goxo
-ez/ez-	39/0	ezegonkor
-kide	38	garaikide
-ezin/ezin-	36/3	garaiezin
-xe/-txe	30	oraintxe
-txo	27	erasotxo
-ezia	27	dotorezia
-zale	27	bizizale
-gintza	16	osasungintza
-t- >-tt-	12	punttu
-gai	12	osagai
-gaitz	10	ulergaitz
-gura	8	logura
-gailu	8	neurgailu
-min	7	ikusmin
-ska/-xka	6	herrixka
-z- >-tx-	6	txoratuta
-tza	4	burujabetza
-tzar	3	krisitzar
-nahi	2	handinahi
-txa	1	neskatxa
a-	1	atenporalitatea
Guztira	**1.623**	

5.1 taula: Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako balentzia-aldatzaile morfologiko eta fonologikoak.

banatu ditugu: i) balentzia indartzen dutenak, ii) ahultzen dutenak eta iii) balentzian eraginik ez dutenak.

Indartu	**Ahuldu**	**Eraginik ez**
-garri	-gabe	-ezia
-tsu	des-	-tasun
-en	-egi	
-s-/-z- >-x-	-keria	
-zale	-ez/ez-	
-t- >-tt-	-ezin/ezin-	
-gura	-xe/-txe	
-min	-txo	
-z- >-tx-	-gaitz	
-tza	-ska/-xka	
-tzar	-txa	
-nahi	a-	

5.2 taula: Balentzia-aldatzaile morfologiko eta fonologikoak sentimendu-balentzian duten eraginaren arabera sailkatuta.

Bukatzeko, 5.3 taulan, Euskarazko Iritzi Corpusetik (Alkorta *et al.*, 2016) lortutako hizkiak duten balio semantikoa irizpidetzat hartuta sailkatuta ageri dira. Bertan, hizkiak eta bustidura adierazkorra 12 multzo semantikotan banatu ditugu. Lehenik eta behin, ezeztapenarekin lotutakoak daude. Ezeztapenezko hizki horiek ezeztapen-marka moduan antzeko modu batean eragiten diote sentimendu-balentziari, baina hitzari eragin beharrean, esaldi edo sintagmari eragiten diote. Multzo horretan aurrizkiak eta atzizkiak aurki daitezke.

Datorren multzoan konparazioarekin lotutako hizkiak ageri dira: *-egi* hizkia, sentimendu-balentzia ahultzen duena eta *-en* hizkia, sentimendu-balentzia indartzen duena.

Hurrengo bi multzoak tamainarekin lotutakoa dira. Semantikoki handigarri funtzioa betetzen duen balentzia-aldatzaile bakarra aurkitu dugu: *-tzar*. Txikigarri funtzioa duten balentzia-aldatzaileak, aldiz, asko dira eta denak atzizkiak dira. Multzoa esanahiarekin lotura duten bi balentzia-aldatzaile aurkitu ditugu: *-tsu* eta *-tza* eta biek sentimendu-balentzia indartzen dute.

Ezeztapena	**Konparazioa**	**Handigarria**	**Txikigarria**
des-	-egi	-tzar	-xe/-txe
-ez/ez-	-en		-txo
-ezin/ezin-			-ska/-xka
-gabe*			-txa
(a-)			
Zailtasuna	**Nolakotasuna**	**Lanbidea**	**Jabetza**
-gaitz	-tasun	-gile	-dun
-garri	-keria	-gintza	-gabe*
	-ezia		
Nahia	**Hurbiltasuna/samurtasuna**	**Multzoa**	**Bestelakoak**
-gura	-s-/-z- >-x-	-tza	-zale
-nahi	-t- >-tt-	-tsu	
-min	-z- >-tx-		

5.3 taula: Euskarazko Iritzi Corpusetik lortutako hizkien sailkapen semantikoa.

Zailtasunarekin lotura duten beste bi balentzia-aldatzaile ere badaude: *-gaitz* eta *-garri*. Lehenak sentimendu-balentzia ahultzen du eta bigarrenak, berriz, sentimendu-balentzia indartzen du.

Bestalde, nolakotasuna adierazten duten hiru balentzia-aldatzaile daude, baina, *-keria* hizkiak sentimendu-balentzia ahultzen du eta *-tasun* eta *-ezia* atzizkiek sentimendu-balentzia indartzen dute.

Jabetza adierazten duten bi atzizki daude: *-dun* eta *-gabe*. Bigarren atzizkiak semantikoki jabetza edo ezeztapena adieraz dezake eta, ondorioz, bi multzoetan sailkatu dugu. Azken honek sentimendu-balentzia ahultzen du. Bustidura adierazkorrei dagokienez, hirurek semantikoki hurbilpena edota samurtasuna adierazten dute eta sentimendu-balentzia indartzen dute.

Azkenik, aipaturiko multzoekin zerikusirik ez duten balentzia-aldatzaile bakarra dago. Hizkia *-zale* da. Zaletasuna adierazten du eta sentimendu-balentzia indartzen du.

5.1.2. Hainbat balentzia-aldatzaile morfologiko dituzten hitzak

Hitz batean hainbat hitz agertzen direnean, hitz horrek egitura jakin bat du. Bi ezaugarri dituzte mota horretako hitzek:

- Hainbat balentzia-aldatzaile morfologiko eta balentzia-aldatzaile fonologikoa den bustidura adierazkorra batera agertzen diren hitzik ez dago. Hau da, hitz batean hainbat balentzia-aldatzaile morfologiko badaude, bustidura adierazkorrik ez da agertzen.
- Hitz batean gehienez hiru balentzia-aldatzaile morfologiko ager daitezke batera bata bestearen atzetik. Eta bitartean, hitzaren sentimendu-balentzian eta orientazio semantikoak aldaketak gertatzen dira.

5.4 taulan, balentzia-aldatzaile morfologiko bat baino gehiago dituzten hitzen egituraketa ageri da. Guztira, mota horretako 36 hitz eta 11 hizki aurkitu ditugu. Gehienez hiru hizki ager daitezke elkarren segidan eta hizkiak erantsi ahala hitzaren sentimendu-balentziak aldaketa ezberdinak jasaten ditu. Hirugarren mailara iristeko ezinbestekoa den bigarren mailan *-en* hizkia egotea. Eta bigarren mailatik hirugarrenera balentziaren ahultzea gertatzen da, nahiz eta hasieratik egoeratik balentzian indartze bat dagoen.

Hizkiak konbinatzeko aukerak eta beren ondorioak askotarikoak dira. Posible da hizkiak gehitu ahala hitzaren sentimendu-balentzia indartzea. Hori gertatzen da *prestigiotsuen* hitzean, *-tsu* eta *-en* hizkiek hitzaren balentziaren positibotasuna indartzen baitute. Hizkiak gehitu ahala hitzaren sentimendu-balentzia negatiboago bilakatzen ere joan daiteke. *Lotsagabekeria* hitzean, *-gabe* eta *-keria* hizkiek *lotsa* hitzaren sentimendu-balentzia ahuldu eta negatibo bilakatzen dute. Tarteko zerbait ere gerta daiteke, hots, hitzaren balentzia hasieran indartzea eta gero ahultzea. *Berritsukeria* hitzaren kasuan, *-tsu* hizkiak hitza positiboago bilakatzen du eta ondoren, *-keria* hizkiak positibotasun hori negatibo bilakatzen du.

5.4 taulan eta azalpenean adibide moduan jarritako hitzak (adibidez, *berritsu* eta *berritsukeria*) *Elhuyar* (Elhuyar, 2013) eta *Zehazki* (Sarasola, 2005) hiztegietan egon daitezke. Hiztegietako sarrerak direnez, sentimenduen lexikoian

Maila	Balentzia	Hizkia
0	handi gehi harritu malenkonia aspertu berri prestigio lotsa garrantzi axola garbiz zehatz	
1	handinahi gehiegi harrigarri malenkoniatsu aspergarri berritsu prestigiotsu lotsagabe garrantzitsu axolagabe garbizale zehazgabe	-nahi (+) -egi (-) -garri (0) -tsu (+) -gabe (-) -zale (+)
2	handinahikeria gehiegikeria harrigarrien malenkoniatsuago aspergarriago berritsukeria prestigiotsuen lotsagabekeria garrantzitsuen axolagabeen garbizaletasun zehazgabetasun	-en (+) -tasun (0) -keria (-) -ago (+)
3	harrigarrienetako prestigiotsuenetariko garrantzitsuenetako	-eneta(ri)ko (+)

5.4 taula: Hitz batean hainbat hizki agertzeko moduak eta horien eragina hitzaren balentzian.

ager daitezke. Sentimenduen balentzian koherentzia mantentzeko, beharrezkoa da *berritsu* hitzak duen sentimenduen balentziari *-keria* balentzia-aldatzailearen eragina aplikatuta, eragiketa horrek eta sentimenduen lexikoian legokeen *berritsukeria* hitzak sentimendu-balentzia bera edukitzea.

5.1.3. Balentzia-aldatzaileak orientazio semantiko ezberdinetako hitzetan

5.1.1 atalean, balentzia-aldatzaile fonologiko eta morfologikoek balentzia indartu, ahuldu edo balentzian eraginik ez dutela adierazi dugu. Eragin ezberdinetako balentzia-aldatzaile horiek orientazio semantiko ezberdinetako hitzekin batu daitezke eta, ondorioz, hitz horien sentimendu-balentzian gertatzen diren aldaketak askotariakoak dira.

5.5 taulan konbinazio posible guztiak ikus daitezke. Aintzat hartuta hitz batek ++ edo −− orientazio semantikoa duela, balentzia horiek aldaketa ezberdinak jasan ditzakete balentzia-aldatzaileen ezaugarrien arabera. Are positiboago edo negatiboago bilaka daitezke edo, bestela, negatibo edo positibo izaterantz jotzen dute, kasu batzuetan, zeinu aldaketa jasateraino.

Hizkia / hitza	**Hizkia (↑)**	**Hizkia (↓)**	**Hizkia (0)**
Hitza (+)	+ + +	+/−	++
Hitza (−)	− − −	−/+	−−
Hitza (0)			

5.5 taula: Balentzia aldatzen duten hizkiek sentimendu-balentzia ezberdineko hitzetan duten eragina aztertzeko taula.

Ondoren, indartzaileak, ahultzaileak edota eraginik ez duten balentzia-aldatzaileak orientazio semaniko ezberdinetako hitzekin batzean gertatzen diren aldaketak aurkeztuko ditugu.

- Sentimendu-balentzia positibodun hitz bat eta balentzia indartzailea den hizki bat.
  Batura honetan, hitzaren balentzia positiboa are positiboago bilakatzen da. Esaterako, *eroso* hitzak +2 balentzia du eta *erosoen* hitzak balentzia are positiboagoa du.

- Sentimendu-balentzia positibodun hitz bat eta balentzia ahultzailea den hizki bat.
  Baturaren ondorioz, hitzak balentzia positiboa manten dezake, nahiz eta ahulagoa izan, edota hitzaren balentzia positiboa negatibo bilaka daiteke. *Ondu* aditzak +2 balentzia du eta *des-* ezeztapenezko balentzia ahutzailearekin batuz gero, balentzia negatibo bilakatzen da. Emaitza *desondu* −2 da[1]. *Estilistiko* (+3) hitzari *-egi* hizkia gehituz gero, aldiz, *estilistikoegi* hitzaren balentzia jaisten da.

- Balentzia positibodun hitz bat eta balentzia-aldatzailea ez den hizki bat.
  Konbinaketa honetan, hitzak bere balentzia mantentzen du. *Lasai* adjektiboak +2 balentzia du eta *lasaitasun* hitzak ere sentimendu-balentzia bera mantentzen du.

- Hitzak sentimendu-balentzia negatiboa eta balentzia indartzaile bat.
  Horren ondorioa hitzaren sentimendu-balentzia are negatiboago bilakatzea da. *Kalte* izenak −2 balentzia du *-tzar* hizki handigarria erabilita, *kaltetzar* hitzak are balentzia negatiboagoa du.

- Balentzia negatibodun hitz bat eta balentzia ahultzailea den hizki bat.
  Kasu horretan, hitzaren balentziak negatibo izaten jarraitzen du, baina intentsitatea ahulago batez edo, bestela, hitzaren balentzia negatiboa positibo ere bilakatu daiteke. *Sentitegi* lexikoian (Alkorta *et al.*, 2018), *labur* hitzak −1 balentzia du eta *laburregi* hitzak balentzia are negatiboa izango luke. Baina, *nahasi* hitzari, −2 balentziarekin, *-ezin* atzizkia erantsiz gero, *nahastezin* sortzen da eta *-ezin* ezeztapena denez eta ezeztapenak ±4 balentzia duenez, hitza ahuldu eta +2 sentimendu-balentzia izatera igarotzen da.

- Balentzia negatibodun hitz bat eta balentzian eragiten ez duen hizki bat.
  Baturaren ondorioz, hitzaren sentimendu-balentzian ez da aldaketarik

[1]Lan honetan, ezeztapena lantzeko hautatu dugun hurbilpena desplazamendu-ezeztapena da eta, horregatik, ezeztapen-markei nahiz ezeztapenezko hizkiei ±4 balentzia esleitu diegu.

gertatzen. *Baldar* nahiz *baldartasun* hitzek −2 sentimendu-balentzia dute, *-tasun* hizkiak ez duelako balentzian eraginik.

- Balentziarik gabeko hitz bat eta balentzia indartzailea den hizki bat. Hitzak balentziari ez duenez, balentzia-aldatzaileak ez du eraginik. Esaterako, *kolore* eta *mendi* hitzak neutralak dira *Sentitegi* sentimenduen lexikoiaren (Alkorta *et al.*, 2018) arabera, hau da, ez dute balentziarik, eta *koloretsu* eta *menditzar* hitzek ere ez dute.

- Balentziarik gabeko hitz bat eta balentzia ahultzailea den hizki bat. Kasu honetan ere, hitzak sentimendu-balentziarik ez duenez, ez dago balentziaren aldaketarik. Adibidez, *mendi* hitza eta *-txo* balentzia ahultzailea den hizkia batzean (*menditxo*), hitzaren sentimendu-balentzian ez da aldaketarik gertatzen.

- Balentziarik gabeko hitz bat eta balentzia-aldatzailea ez den hizki bat. Konbinazio honetan ere, hitzaren sentimendu-balentzian ez da aldaketarik gertatzen. Horren adibide den *koloretasun* hitzak (*kolore* + *-tasun*) ez du sentimendu-balentziarik eta balentzia-aldaketarik ere jasaten.

5.1.4. Maila fonologikoa eta morfologikoaren garrantzia euskaran

Polanyi eta Zaenenek (2006) ez dute balentzia-aldatzaile morfologiko eta fonologikoen inguruko aipamenik egiten testuinguruko balentzia-aldatzaileen azalpenean.

Baina, lan horrek balentzia-aldatzaileez ari denean, oro har, ingelesa hartzen du ikergaitzat. Horregatik, bertan zerrendatzen diren balentzia-aldatzaile mota guztiak beste hizkuntzetan ezin daitezke baliagarriak izan edota beste hizkuntzetan ingelesean ez dauden balentzia-aldatzaileak egon litezke.

Gure ustez, hori gertatzen da fonologian eta, batez ere, morfologian egon daitezkeen balentzia-aldatzaileekin. Euskara eta ingelesa morfologia tipologiaren aldetik ezberdinak dira, Ingelesa hizkuntza analitikoa da (Szmrecsanyi,

2012), hots, esaldietan hitzen arteko loturak hitz laguntzaileen bidez (preposizioak eta partikulak, esaterako) edo hitz hurrenkeraren bidez egiten dira flexioarekin egin beharrean. Euskara, aldiz, hizkuntza sintetiko eranskaria da (Euskaltzaindia, 1985). Morfologiak garrantzi handia du eta eranskaritasuna erabiltzen du morfemak erabiltzeko. Morfemok hitz erroaren esanahian aldaketak sor ditzakete. Har ditzagun bi adibide:

(47) Hori ikusi zuenean, irribartxoa atera zitzaion aurpegian.

(48) When she saw that, a giggle came on his face.

Bi adibide horietan, euskarazko eta ingelesezko adibide bana dugu eta esanahi bera dute, baina esanahi bera adierazteko moduak ezberdinak dira. (47) adibidean, irribarrea txikia dela adierazteko, *-txo* hizkia erabiltzen da. Ingelesezko (48) adibidean, ordea, beste estrategia bat erabiltzen da eta irribarre txikia adierazteko *giggle* hitza erabiltzen da. Arrazoi horrengatik eta euskaran morfologia oso aberatsa delako, morfologian egon daitezkeen balentzia-aldatzaileak aztertzea erabaki dugu.

Bestalde, fonologiako balentzia-aldatzaileak aztertzearen arrazoia beste bat da. Balentzia-aldatzaileen azterketari ikuspegi orokor bat eman nahi diogu eta, horregatik, fonologia ere lanean sartu nahi dugu. Harluxet hiztegiak (Fundazioa, 1995) dionez, bustidura fonema ez-sabaikari bat testuinguru jakin batean sabaikari bilakatzean datzan fenomenoa. Bustidura adierazkorra bereiztu egin behar da asimilazio bustiduratik. Oñederrak (1990) horrela egiten du bi bustiduren arteko bereizkuntza:

> Euskarazko fonologiaz ari garela, bustikuntza bi motatakoa izan daitekeela izan behar da kontuan: katea fonikoan ingurune jakin batean asimilazioz sortua edo, labur adierazteagatik, bustidura adierazgarria esaten zaiona. (Oñederra, 1990, 13 orr.)

Gaineratzen duenez, bi bustidurak hots klase berari jartzen zaizkio eta aldaketaren ondorioa ere bietan berdina da fonetikan gertatzen den aldaketa berdina delako. Berdintasunak aipatzeaz gain, desberdintasunak zertan diren ere aipatzen du:

> Bata ala bestea gertatzeko baldintzak dira aldatzen direnak, asimilazioa katea fonikoko inguruneak sortutako bilakabide sintagmatikoa den bitartean, bustidura adierazgarria hizkuntzaren eremu semantikoari (*lato sensu*) bait dagokio. (Oñederra, 1990, 14 orr.)

Hots, asimilazio bustiduran, asimilazioa gertatzearen baldintza fonetikoa da eta bustidura adierazkorrean, semantikoa. Beste modu batean esanda, hitz batean aldaketa semantikoa gauzatzeko egiten da bustidura adierazkorra eta nahita eragindako aldaketa fonetikoa da. Bustidura adierazkorraz txikitasuna, maitasuna edo goxotasuna adierazi nahi da.

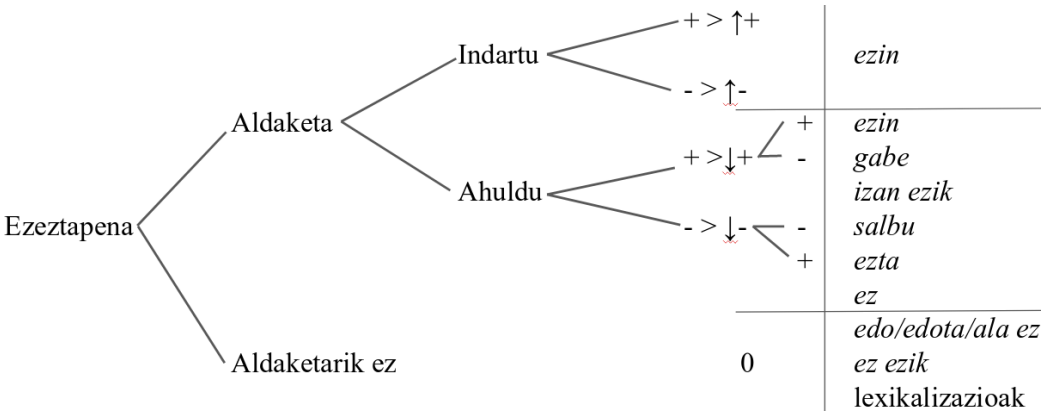
Bustidura adierazkorrak hitzen aldaketa semantikoa eragiten duenez, sentimenduen analisiaren ikuspegitik, balentzia-aldatzaile bat da eta horretan datza fenomeno fonologiko horren interesak.

5.2 Sintaxi maila

Sintaxi maileko 5.2.1 azpiatalean, ezeztapenezko balentzia-aldatzaileak aurkeztuko ditugu. 5.2.2 azpiatalean, trukaketa- eta desplazamendu-ezeztapenaren arteko ezberdintasunak azalduko ditugu, eta desplazamendu-ezeztapena metodologian zergatik erabili dugun ere adieraziko dugu. Azkenik, 5.2.3 azpiatalean, ezeztapen-markak eta beren irismena identifikatzeko erregelak *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) nola sortu eta ondoren, nola ebaluatu ditugun azalduko dugu.

5.2.1. Ezeztapenezko balentzia-aldatzaileak

5.1 irudian, ezeztapenezko balentzia-aldatzileak sentimendu-balentzian nola eragiten dute ikus daiteke.



5.1 irudia: Ezeztapen-marken eragin ezberdinak sentimendu-balentzian.

Ezin sentimendu-balentzia indartu dezakeen ezeztapen-marka bakarra da. Bestalde, sentimendu-balentzia ahultzen duten ezeztapen-markak gehiago dira: *ezin*, *gabe*, *izan ezik*, *salbu*, *ezta* eta *ez*. Azkenik, badaude zenbait egitura non ezeztapen-marka agertzen den eta ez duen sentimendu-balentzian eragiten. Egitura horiek ezeztapena juntagailuarekin, kontrastezko ezeztapena eta lexikalizatutako egiturak dira. Jarraian, 5.1 irudian azaldutako sailkapena xehetuago emango dugu.

Ezeztapen-markaren instantzia guztietatik (359), zazpi kasu antzeman ditugu ezeztapen-markak irismenaren sentimendu-balentzia indartu duena. Balentzia-aldaketa hori *ezin* ezeztapen-markak adjektiboari edo aditzondoari lotuta dagoenean bakarrik gertatzen da (corpusean agertu diren kasu guztien % 1,96).

(49) Dena nahasten da maisulan **ezin** ederragoa$_{+4}$ osatzeko. (MUS21).[2]

(49) adibidean, sentimendu-balentzia indartzen den adibide bat dago. Kasu horretan, ezeztapen-markak (*ezin*) konparaziozko atzizkia duen adjektiboari eragin eta bere sentimendu-balentzia indartu du. Hots, positiboa zen sentimendu-balentzia are positiboagoa bihurtzen du. Izan ere, esaldi horretan *ederragorik* ez dagoela esatea *ederrena* dela esatea da; beraz, sentimendu-balentziaren indartze bat dago.

Kasu gehienetan, aztertu dugun corpuseko instantziek sentimendu-balentzia ahultzen dute. Hainbat ezeztapen-marka motek ahultzen dute sentimendu-balentzia: i) *ez*, ii) *gabe*, iii) *ezin*, iv) *izan ezik*, v) *salbu* eta azkenik, vi) *ezta*. Kasuen % 89,98etan (323 instantzietan) gertatu da sentimendu-balentziaren ahultzea.

(50) Horrek **ez** die eragotzi$_{-2}$ ordea, 57 milioi euro ematea San Mames klub pribatuari! (POL30)

(51) Bruch-en kontzertua, munduko interesgarriena izan **gabe**, lan erakargarria da, oso, erraza entzuteko, eta bakarlari honen bertsioak dotoreziaren *plus*a izan zuen dudarik gabe. (MUS22)

(52) Irabazi$_{+2}$ **ezin**ik jarraitzen du Eibarrek, baina oso puntu ona eskuratu du Getaferen zelaian. (KIR17)

(53) "Ez baitute ezertarako balio izan, egun bateko soldata galtzeko$_{-2}$ **izan ezik**". (POL17)

[2]Esaldian dauden anotazio ezberdinen esanahiak hauek dira: **beltz koloreak** ezeztapen-marka adierazten du, azpimarkatutakoak ezeztapenaren irismen-eremua seinalatzen du. Berez, ezeztapen-markek ez lukete egon beharko azpimarkatuta, baina hori ez *ezin* ezeztapen-markaren kasuan. Izan ere, ezeztapen-marka horrek bi esanahi ditu: i) ezeztapena bera eta ii) posibilitatea. Horregatik, azpimarkatuta dago.

(54) E.T. eta Indiana Jones filmak **salbu**, noski. (ZIN18)

(55) **Ezta** Euskal Herria, Espainiako Estatua, Portugal edo Italia moduko beste herrialdeei ere. (POL08)

(50) adibidean, ezeztapen-markaren eragin-eremua esaldi osoa da eta *eragotzi* aditzaren −2 sentimendu-balentzia ahultzen du. Beraz, zeinu positiboa (+4) hartzen du ezeztapen-markaren eraginez. (51) adibidean, *gabe* ezeztapen-markak sintagma bati eragiten dio (*munduko interesgarriena izan*). (52) adibidean, *ezin* ezeztapen-markak bere ezkerretara dagoen *irabazi* aditzaren +2 sentimendu-balentzia ahultzen du. (53) adibidean, ezeztapen-marka *ezik* da eta *izan* aditzarekin batera *egun bateko soldata galtzeko* sintagma ezeztatzen du. (54) adibidean, *salbu* ezeztapen-markak izen-sintagma bati (*E.T. eta Indiana Jones filmak*) eragiten dio. Azkenik, (55) adibidean, *ezta* ezeztapen-markak sintagma osoari eragiten dio.

Azkenik, aurretik aipaturiko ezeztapen-markek sentimendu-balentzian eraginik ez duten kasuak ere badaude. Kasu hori sentimendu-balentzia indartzea baino ohikoagoa da, corpusean instantzien % 8,08an, 29 instantzietan, gertatu baita. Ezeztapen-markak berak ez du sentimendu-balentzian eraginik horrelako kasuetan: i) ezeztapen-marka juntagailuarekin agertzen denean, ii) ezeztapen-marka kontrastezko ezeztapenaren parte denean, eta iii) ezeztapen-marka egitura lexikalizatu batean agertzen denean (hots, egitura horiek beren berezko esanahia dutenean eta horietako batzuk hiztegietako sarrerak direnean).

(56) Cate Le Bon bai edo ez, hemen ez dago erdibidekorik. (MUS33)

(57) Ikuspuntu politikotik$_{-1}$ ez ezik, ekonomikotik$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)

(58) Sei puntu baino ez dituela, hamaseigarren postuan da Reala sailkapenean. (KIR27)

(56) adibidean, idazleak aurkako bi ideia adierazten ditu (bai eta ez). Horrenbestez, ez da balentzia-aldaketarik sortzen. (57) adibidean, kontrastezko

ezeztapen bat dago eta informazioa gehitzeko funtzioa betetzen du (Silvennoinen, 2017). Hots, ezeztapen-marka emendiozko juntadura batean agertzen. Horrelako egituretan, *ez ezik* ezeztapen-markak bere aurretik dagoen elementua (*Ikuspuntu politikotik* adibidearen kasuan) ezeztatzen du. Baina, sentimendu-balentziaren ikuspegitik, ez da balentziarik ezeztatu eta ondorioz, ahuldu. Esaldian, ezeztapen-marka daraman lehen zatiak bere atzetik informazioa emendatzen edo gehitzen du. Azkenik, (58) egitura lexikalizatuaren adibidea da.

5.2.2. Trukaketa- eta desplazamendu-ezeztapena

Sentimenduen analisian eta zehazki, lexikoian eta hizkuntza-ezagutzan oinarritzen den ikuspegian, bi hurbilpen daude ezeztapena lantzeko. Bi hurbilpen horiek trukaketa-ezeztapena (*switch negation*, ingelesez) (Sauri, 2008) eta desplazamendu-ezeztapena (*shift negation*, ingelesez) dira. Azter ditzagun bi hurbilpenak beheko bi adibideaetan[3].

(59) Edaritegia [**ez** da ona$_{+3}$]$^{-3}$, baina bertako musika ona$_{+3}$ da.

(60) Edaritegia [**ez**$_{-4}$da ona$_{+3}$]$^{-1}$, baina bertako musika ona$_{+3}$ da.

Lehen adibideak (59) trukaketa-ezeztapenari dagokio eta bigarrenak (60), berriz, desplazamendu-ezeztapenari. Lehen adibidean, *ez* ezeztapen-markak ez du balentziarik eta ezeztapen-markaren eragina da bere irismeneko parte den *ona* hitzaren sentimendu-balentzia (+3) alderantzikatzea. Beraz, trukaketa-ezeztapenean, irismeneko sentimendu-balentziari zeinua aldatzen zaio, positibotik negatibora edo alderantziz. Bigarren adibidean, desplazamendu-ezeztapenaren adibidea ageri da. Kasu horretan, *ez* ezeztapen-markak balentzia jakin bat du (−4) eta ezeztapen-markaren eragina da irismeneko hitzen sentimendu-balentziaren baturari (adibidean, −4) sentimendu-balentzia gehitzea. Irismeneko sentimendu-balentzia negatiboa bada, ezeztapen-markaren balentzia +4 izango dela.

Bi hurbilpen horien artean, guk desplazamendu-ezeztapena erabili dugu. Hurbilpen hori sintaxi mailako ezeztapenean eta ezeztapenarekin loturiko

[3](59) eta (60) adibideak (Taboada *et al.*, 2011) lanetik hartu eta itzuli dira.

hizkiekin erabili dugu. Balentzia-aldatzaile horri ±4 balentzia esleitu diogu eta, ondoren, balentzia duten hitzetan zer aldaketa eragiten duten aztertu dugu. Taboada *et al.*ek (2011) aipatzen duenez, azken hurbilpen hori hobeto dabil eta ez dago kontraesanik trukaketa-ezeztapenean bezala. Azter dezagun trukaketa-ezeztapenak duen akatsa.

(61) Bikaina$_{+5}$ → Ez bikaina$_{-5}$. Anker$_{-5}$.

(62) Ez bikaina$_{-5}$. Ez ona$_{-3}$.

(61) adibidean, trukaketa-ezeztapena erabiliz, *bikaina* adjektiboa +5 sentimendu-balentzia izatetik, ezeztapen-markaren eraginez, −5 sentimendu-balentzia izatea igarotzen da. Baina, hemen lehen kontraesana sortzen da; izan ere, intuitiboki *ez bikaina* eta *anker* elkarrengandik urrun daude, baina hurbilpen horren arabera, biek sentimendu-balentzia bera dute.

Bigarren kontraesana (62) adibidean beha daiteke. Bertan *bikaina* (+5) eta *ona* (+3) ezeztatuta daude. Ezeztapenaren eraginez, *ez bikaina* egiturak −5 sentimendu-balentzia du eta *ez ona*k aldiz −3. Baina biak alderatzen baditugu, *ez bikaina ez ona* baino positiboagoa dela ohartuko gara. Hala ere, trukaketa-ezeztapena jarraituz, *ez bikaina* egiturak sentimendu-balentzia negatiboagoa du *ez ona*k baino. Horrengatik guztiarengatik, mugimendu ezeztapena hurbilpena hautatu dugu eta ezeztapen-markek ±4 balentzia izango dute.

5.2.3. Ezeztapen-markak eta beren irismena identifikatzeko erregelak

Metodologian, 3.2.2 atalean, ezeztapen-markek aditzen eta sintagmen sentimendu-balentzia nola eragiten duten identifikatu ondoren; ezeztapen-markak eta beren inguruneko hitzen gramatika-kategoriekin 5.6 taulako erregelak osatu ditugu. Erregela horiek *Murriztapen Gramatika*n, MGn, (Karlsson *et al.*, 1995) erregelak sortzeko baliatu dira.

Murriztapen Gramatika (Karlsson *et al.*, 1995) testuingurua iruditu zaigu egokiena ebaluazioa egiteko, gramatikaren ezaugarriak ondo uztartzen baitira

Zenb.	Ezeztapen-marka	Erregelen egitura
1	ezin	PM *ezin* + [adjektiboa/adberbioa] (+ atzizki konp.) PM
2	ez	PM [(IS +) aditza +] *ez* PM PM [(IS +)] *ez* [+ ad. lag. (+ IS) + aditza (+ IS)] PM PM [IS +] *ez* PM
3	ez	PM *ez* [+ IS +] *ez* [+ IS] (...) PM
4	gabe	PM [IS/AS/sintagma +] *gabe* PM
5	ezin	PM [(IS) + aditza + *ezin*] PM PM [*ezin* (+ ad. lag.) (+ IS) + aditza (+ IS)] PM
6	izan ezik	PM [IS/sintagma] + *izan ezik*
7	salbu	PM [IS] + *salbu* PM
8	ezta	PM *ezta* + [IS/sintagma] PM
9	ez	PM [aditza/*bai*] *edo/ala/edota ez* PM
10	ez	PM [IS] *ez ezik* PM
11	ez, gabe, ezin, ezean	Egitura lexikalizatuak

5.6 taula: Sentimendu-balentzian eragin ezberdinak dituzten ezeztapenak identifikatzeko proposatu diren erregeletako batzuk.

ebaluatu nahi ditugun alderdiekin. Hiru ezaugarri nagusi eta abantaila ditu gramatika horrek:

- *Murriztapen Gramatika* (Karlsson *et al.*, 1995) hizkuntzarekiko independentea da. *Murriztapen Gramatika* analisi morfologikoan oinarritzen da, edozein testu analizatzeko helburuarekin. Gainera, egoera finituko mekanismoetan oinarritzen da, eta hori hizkuntzalaritzaren ikuspegitik gramatikaren formalismoa oso intuitiboa eta erabilerraza da.

- Desanbiguazio-erregelak. Gramatikako erregelek analisi morfologikoaren eta funtzio sintaktikoen desanbiguazioa gauzatzen dute. Erregelek testuinguru jakin batean zuzenak edo egokiak ez diren interpretazioak kentzen dituzte.

- Islapen-erregelak. Erregela horiek etiketatze morfologikoan eta sintaktikoan egon daitezkeen hutsuneak osatzen dituzte, testuingurua kontuan hartuz. Guk ezeztapen-markaren testuingurua, hau da, irismena identifikatu nahi dugu eta, horregatik, islapen-erregelak egokiak dira guretzat.

5.2.3.1. Murriztapen Gramatikako erregelak sortzeko eta ebaluatzeko prozedura

Erregelen egituraketa

Jarraian urratsez urrats, 5.6 taulako erregelak ingurunera nola egokitu ditugun esplikatuko dugu. *Murriztapen Gramatika* (Karlsson *et al.*, 1995) testuinguruan sartu behar ditugun elementuak hauek dira[4]:

- Puntuazio-markak, ezeztapen-markak eta egitura lexikalizatuetako hitzak elementuen zerrendetan (LIST) sartzea erabaki dugu. Ia erregela denetan agertuko diren hitzak dira eta hitz jakinak dira, hots, beti forma bera dute.

  (63) LIST EZ = ”ez”;

  (64) LIST BESTERIK = ”beste”;

  (63) eta (64) adibideetan, berriz, *ez* ezeztapen-marka eta egitura lexikalizatuetako *besterik* hitza, hurrenez hurren, elementuen zerrendetako elementu bakarrak dira, modu horretan, hitz horiek ezeztapen-marken erregela ezberdinekin konbinatzeko aukera edukitzeko.

- Irismeneko hitzen gramatika-kategoriak eta haien posizioa islapen-erregeletan sartu ditugu. Irismenean, aurreko kasuetan ez bezala, ez dakigu zein hitz ager daitezkeen bertan. Horregatik, ezin ditugu irismeneko hitzekin elementuen zerrendak osatu. Dakigun bakarra irismenean ager daitezkeen hitzen gramatika-kategoria eta ezeztapen-marketatik haien distantzia dira, 5.6 taularen bidez lortu ditugunak. Horrenbestez, informazio hori islapen erregeletan adierazi dugu:

  (65) MAP (!salbuBUK) TARGET (ADB) IF (0C SALBU);
  MAP (!salbuHAS1) TARGET (IZE) IF (1C SALBU);

  (65) adibidean, bestalde, bi erregela ageri dira *larunbata salbu* bezalako egiturak identifikatzeko. *!salbuBUK* etiketa duen erregelan, 0 posizioan SALBU elementuen zerrendak egon behar duela adierazten du

[4]Elementu horiek ikusgai daude A eranskinean.

eta, gainera, elementu zerrendakoak adberbioa izan behar du. Bestalde, *!salbuHAS1* etiketan duen erregelan, *salbu* ezeztapen-markaren aurretik izen batek egon behar duela adierazten da. Kasu horretan, hitza edozein izan daiteke, baina badakigu izenaren gramatika-kategoriakoak izan behar duela, 5.6 taularen bidez. Aipaturiko etiketa horiek izango dira sortu ditugun erregelak corpusetik pasatzean utziko duten aztarna.

Erregelak ebaluatzeko prozedura

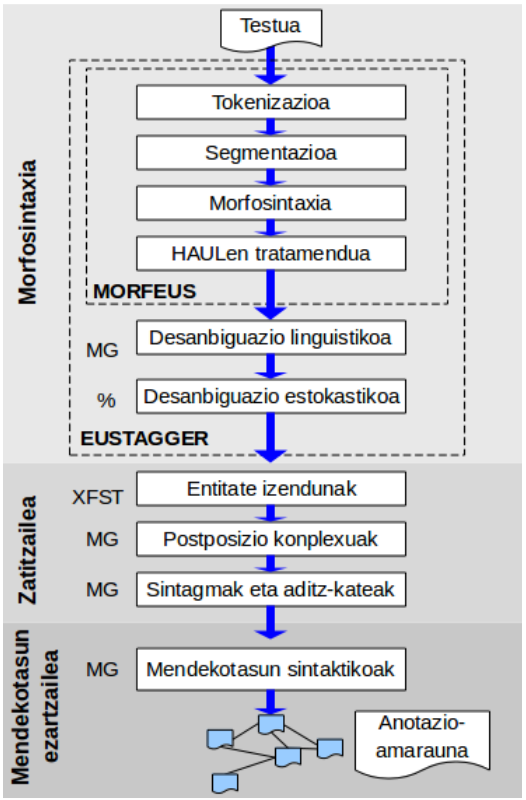
Erregelen ebaluazioa etiketatzaile batek gauzatu du, baina, aldez aurretik, etiketatzaile horren eta beste etiketatzaile baten adostasun maila neurtu dugu, ebaluazioaren kalitatea ebaluatu ahal izateko. Bi etiketatzailek Euskarazko Iritzi Corpusaren (Alkorta *et al.*, 2016) % 10a (2.706 hitz) etiketatu dute. Bi etiketatzaileen arteko adostasuna zein hitz etiketatzeari dagokionez, kappa 0,91koa da. Bi etiketatzaileak etiketatutako beharreko hitzak nola etiketaturi (ETIK_ONDO, ETIK_FALTA eta ETIK_GAIZKI) dagokionez, kappa 0,69koa da, zeina Landis eta Kochen (1977) arabera *sendoa* den.

Bi etiketatzaileen arteko adostasuna neurtu ostean, Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 144 iritzi-testu erabili ditugu *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelak ebaluatzeko. 144 iritzi-testuko corpusari Ixa taldearen analisi-katea (Aduriz *et al.*, 2004) pasa diogu. Analisi-katearen arkitektura 5.2 irudian ikus daiteke.

Ixarko analisi-katearen oinarrian EDBL-LBDBL baliabide lexikal konputazionala (Aldezabal *et al.*, 2006) dago eta analisi-kateak hiru atal ditu: morfosintaxia, zatitzailea eta mendekotasun ezartzailea. Corpusari analisi-katea pasa ondoren, emaitza hitz bakoitzeko analisi bakarra izan da 5.3 irudian agertzen den moduan.

Bertan, analisi-katearen arabera, *azaroaren* hitza izen (IZE) arrunta (ARRU) da eta bere kasua genitiboa (GEN>) da.

Euskarazko Iritzi Corpuseko (Alkorta *et al.*, 2016) 144 iritzi-testuei analisi-katea pasa ostean, *Murriztapen Gramatika*n (Karlsson *et al.*, 1995) sortutako erregelak aplikatu ditugu corpusean eta erregela horiek corpuseko analisietan beren etiketa utzi dute. 3.2.2 atalean aipatu dugun moduan, etiketa erregelak ebaluatzeko erabili ditugu.



5.2 irudia: Ixa taldearen analisi-katea (Oronoz, 2009).

```
"<azaroaren>"
        <Correct!> "azaro" IZE ARR GEN NUMS MUGM ZERO w67,L-A-IZE-ARR-26,
            lsfi6 @IZLG>
```

5.3 irudia: Ixa taldeko analisi-katearen emaitza.

Murriztapen Gramatika (Karlsson *et al.*, 1995) esleitutako etiketak zuzenak edo okerrak diren, edo etiketaren bat falta den ebaluatzeko hiru kasuistika bereiztu ditugu[5]:

- ETIK_ONDO. Sorturiko erregelak dagokion hitzari etiketa jarri badio ontzat joko dugu. Kasu horretan, beraz, erregelak ondo identifikatu du

[5]Erregelak ebaluatzeko etiketak corpusaren zati batean aplikatuta 3.9 irudian daude ikusgai.

ezeztapenarekin loturiko elementuren bat (ezeztapen-marka, irismena edo egitura lexikalizatua) corpusean.

- ETIK_FALTA. Sorturiko erregelak dagokion hitzari etiketa ez badio jartzen okertzat joko dugu. Beste kasu batzuetan, gerta liteke erregelak ezeztapenarekin loturiko elementuren bat ez identifikatzea corpusean eta kasu horretan erabiliko dugu etiketa hori.
- ETIK_GAIZKI. Sorturiko erregelek ez dagokion hitzari etiketa jartzen badio okertzat joko dugu. Azken kasuan, gerta liteke erregelek ezeztapenarekin loturarik ez duen hitz bat ezeztapenarekin lotura duen elementutzat jotzea eta, kasu horretan ere, kasuistika hori okertzat joko dugu.

5.2.3.2. Erregelen ebaluazioaren emaitzak

Lehenik eta behin, emaitzak ezeztapen-markek sentimendua balentzian eragiten dutena (indartu, ahuldu edo eraginik ez) kontuan hartuta emango ditugu. Ondoren, emaitzak ezeztapenari lotutako elementuetan (ezeztapen-markak, irismena eta egitura lexikalizatuak) arabera emango ditugu.

5.7 taulako datuen arabera, erregelek ezeztapen-markekin lotutako osagaien hitzak ondo identifikatzen ditu, F1 puntuazioa 0,86 baita. Zehatzago esanda, erregelek ezeztapenarekin lotutako hitzak identifikatzerakoan 0,91ko doitasuna eta 0,80ko estaldura izan dute.

Eragina	**Doitasuna**	**Estaldura**	**F1**	**Hitzak (instantziak) ✓/✗/✢**
Indartu	1,00	1,00	1,00	4 (2): 4/0/0
Ahuldu	0,93	0,80	0,86	1.050 (192): 784/63/203
Eraginik ez	0,97	1,00	0,98	36 (19): 35/1/0
Guztira	0,93	0,80	0,86	1.090: 823/64/203

5.7 taula: Emaitzak ezeztapen-markek sentimendu-balentzian eragiten dutenaren ikuspegitik.

Ezberdintasun batzuk daude F1 puntuazioan, ezeztapen-markek balentzian eragiten dutenaren arabera. Esaterako, ezeztapen-markak indartzen duenean edota eraginik ez duenean, F1 puntuazioa altua da. Baina, eurei loturiko

hitz eta instantzia[6] kopurua baxua da. Ezeztapenak balentzia ahultzearen kasuan, aldiz, F1 puntuazioa 0,86 da, baxuagoa. Indartzearen kasuan, ezeztapenari lotutako hitz guztiak (4) ondo identifikatu dira. Ahultzean, 784 ondo identifikatu dira, 63 gaizki eta 203 ez dira identifikatu. Eraginik ez dagoen kasuetan, 35 hitz ondo identifikatu dira eta bakarra gaizki. Ezeztapena ahultze eraginaren kasuan, F1 puntuazioa 0,86 da. Emaitzak ezeztapenari loturiko elementu-motan oinarrituz ageri da 5.8 taulan.

	Doitasuna	**Estaldura**	**F1**	**Hitzak (instantziak): ✓/×/✢**
Ezeztapen-markak	1,00	0,96	0,98	195 (195): 188/0/7
Egitura lexikalizatuak	0,96	1,00	0,98	28 (16): 27/1/0
Irismena	0,91	0,75	0,82	867 (195): 601/63/196
Guztira	0,93	0,80	0,86	1.090: 823/64/203

5.8 taula: Emaitzak ezeztapenari loturiko elementu-motaren ikuspegitik.

Bertan beha daitekeenez, ezeztapen-markak eta egitura lexikalizatuak oso ondo identifikatu dira, F1 puntuazioa 0,98 baita bietan. Zazpi ezeztapen-marka ez dira identifikatu eta corpuseko hitz bat egitura lexikalizatutzat hartu da, nahiz eta ez den horrela. Irismenean, ordea, F1 puntuazioa baxuagoa da, 0,82koa, hain zuzen ere. Irismenaren parte diren 196 hitz ez dira identifikatu eta irismenaren parte ez diren 64 hitz irismeneko zatitzat hartu dira. Horretaz gain, 867 hitzetatik 196 ez dira irismeneko parte bezala jo, nahiz eta hala diren.

Errore-analisiak erregelen osaketari buruzko informazioa eman digu. Irismenari dagokionez, erregelek ez dituzte 196 hitz identifikatu (identifikatu gabeko zazpi ezeztapen-markak kontuan hartu gabe). Hitz horiek zer dela eta identifikatu gabe gelditu diren aztertu ondoren, kasu gehienetan, puntuaziomarkaren murriztapenarekin lotura dutela ohartu gara. Hau da, puntuaziomarka murriztapena irismena esaldi barrura mugatzeko sortua izan da, baina

[6] 5.7 taulan, eskuineko zutabean, hitzei eta instantziei loturiko datuak jarri ditugu. Lehenik eta behin, hitz-kopurua agertzen da, ondoren, parentesi artean, instantzia-kopurua eta ondoren, ezeztapenari loturiko hitzak ondo, gaizki edo ez diren asmatu. Ikusten den moduan, hitz eta instantzia-kopuruak ezberdinak dira. Ezeztapen-marken kasuan, hitz bat instantzia bat da. Irismenaren kasuan, instantziako hitz-kopurua ezberdina da. Guztira 192 irismen daude corpusean eta horiek 1.050 hitz dituzte. Azkenik, egitura lexikalizatuetan, hitz batetik hiru hitzerainokoa izan daiteke instantzia bat, betiere egitura lexikalizatu horren ezaugarrien arabera.

horrek zeharkako kalte batzuk sortzen ditu. Esaldiren batean zerrendaketa bat dagoenean, eta zerrendaketa hortako elementuak koma bidez bereizita daudenean, puntuazio-markaren murriztapena ez da ondo ibiltzen, zerrendaketako elementu bakarra identifikatzen baitu, esaldia hor bukatzen dela interpretatzen duelako.

5.3 Diskurtso maila

5.3.1 azpiatalean, diskurtso mailan aztertu ditugun balentzia-aldatzaileak aurkeztu ditugu. 5.3.2 atalean, Egitura Erretorikoaren Teoria (RST) (Mann eta Thompson, 1988) deskribatuko dugu.

5.3.1. Balentzia-aldatzaileak: nukleartasuna eta unitate zentrala

Balentzia-aldatzaileen azterketa diskurtso mailan hiru ikuspegitatik egin dugu. Alde batetik, nukleartasunak balentzia-aldaketan eraginik duen aztertu dugu (5.3.1.1 atala). Beste aldetik, unitate zentralak balentzia-aldaketan eraginik ere baduen aztertu dugu (5.3.1.2 atala). Azkenik, unitate zentralak Azkenik, erlaziozko diskurtso-egiturak unitate zentraletik duen distantzia bakoitzean zenbateko maiztasunez agertzen diren aztertu dugu (5.3.1.3 atala).

5.3.1.1. Orientazio semantikoa eta nukleartasuna

Balentzia-aldatzaileen eta nuklearitatearen arteko lotura aztertzeko diskurtso-erlazioak aztertu ditugu. Zehazki, zer diskurtso-unitatek (nukleoak edo sateliteak edota lehen testu-zatia edo azken-testu zatia) duen erlazio osoarekin orientazio semantikoaren adostasun gehien neurtu dugu. 5.9 taulan, orientazio semantikoaren adostasunari buruzko emaitzak azter daitezke, nuklearitatea eta osagaien lekua aintzat hartuz[7]. Emaitza horietan, adostasunik handiena +1 balioa da eta adostasunik baxuena 0 balioa.

5.9 taulako emaitzen arabera, kasu gehienetan, nukleo-unitateak erlaziozko diskurtso-egitura osoaren orientazio semantikoarekin adostasun handiagoa erakusten du satelite-unitateak baino. Orientazio semantikoaren adostasunean nukleo-unitateak eta satelite-unitateak duten ezberdintasuna ez da oso handia. 0,73 da nukleo-unitatearen adostasuna eta 0,64, berriz, satelitearena. Hala ere, joera hori ez da gertatzen erlaziozko diskurtso-egitura mota guztietan. Ebaluazioa taldearen[8] kasuan, satelite-unitateak erlaziozko

[7]Letra xehez dauden erlaziozko diskurtso-egiturak multzoak dira. Letra larriz daudenak erlaziozko diskurtso-egitura bakarra dira.

[8]Ebaluazioa taldea EBALUAZIOA eta INTERPRETAZIOA erlaziozko diskurtso-

Adostasuna Xrekin	Instantziak	Nukleoa	Satelitea
Kausa taldea	71	0,83	0,66
Aurkaritza taldea	70	0,70	0,33
BALDINTZA	18	0,61	0,56
Ahalbideratzea taldea	6	0,60	0,5
Ebaluazioa taldea	140	0,69	0,82
Ebidentzia taldea	53	0,83	0,64
KONTRASTEA*	26	0,73	0,31
Guztira	384	0,73	0,65

5.9 taula: Orientazio semantikoaren adostasuna erlaziozko diskurtso-egituren eta nukleoaren/satelitearen artean.

diskurtso-egitura osoaren orientazio semantikoarekin adostasun handiagoa erakusten du satelite-unitateak baino. Kasu horretan, satelite-unitatearen adostasuna 0,82koa da nukleo-unitatearena 0,69 den bitartean. Bestalde, Aurkaritza taldeak[9] du nukleo-unitatearen eta satelite-unitatearen arteko ezberdintasunik handiena. Nukleo-unitateak 0,70ko adostasuna du eta satelite-unitateak, aldiz, 0,33. Beraz, puntuazioan 0,44ko ezberdintasuna dago. Bukatzeko, nukleo-unitateak Kausa taldean[10] eta Ebidentzia taldean[11] erlaziozko diskurtso-egitura osoaren orientazio semantikoaren adostasunik handiena dute, puntuazioa 0,83koa baita bietan.

Erlaziozko diskurtso-egitura osoaren orientazio semantikoaren adostasunik handiena lehen eta azken unitateak duen aztertzen badugu, emaitzak argigarriagoak dira. 5.10 taulak erakusten duen moduan, erlaziozko diskurtso-egituraren azken unitateak erlazio osoaren orientazio semantikoarekin adostasun handiagoa du. Azken unitatearen kasuan, adostasunaren puntuazioa 0,78koa da eta lehen unitatearen kasuan, berriz, 0,58koa. Beren artean dagoen aldea esanguratsua da. Kasu horretan ere, erlaziozko diskurtso-egitura mota guztiek ez dute joera erakusten. Ahalbideratzea[12] taldean, erlazioa-

egiturek osatzen dute.

[9]Aurkaritza taldea KONTZETSIOA eta ANTITESIA erlaziozko diskurtso-egiturek osatzen dute.

[10]Kausa taldea KAUSA, ONDORIOA eta HELBURUA erlaziozko diskurtso-egiturek osatzen dute.

[11]Ebidentzia taldea EBIDENTZIA eta JUSTIFIKAZIOA erlaziozko diskurtso-egiturek osatzen dute.

[12]Ahalbideratzea taldea AHALBIDERATU eta MOTIBAZIOA erlaziozko diskurtso-

Adostasuna Xrekin	Instantziak	Lehen unitatea	Azken unitatea
Kausa taldea	71	0,66	0,83
Aurkaritza taldea	70	0,30	0,73
BALDINTZA	18	0,28	0,89
Ahalbideratzea taldea	6	0,67	0,50
Ebaluazioa taldea	140	0,68	0,83
Ebidentzia taldea	53	0,79	0,68
KONTRASTEA*	26	0,35	0,69
Guztira	384	0,58	0,78

5.10 taula: Orientazio semantikoaren adostasuna erlaziozko diskurtso-egituren eta lehen osagaiaren/azken osagaiaren artean.

ren lehen unitateak azkenak baino orientazio semantikoaren adostasun handiagoa dute. Ahalbideratzearen kasuan, lehen unitatearen adostasun puntuazioa 0,67koa da eta azken unitatearena 0,50ekoa. Bestalde, Ebidentzia taldean puntuazioak 0,79 eta 0,68 dira, hurrenez hurren. Lehen unitatearen eta azken unitatearen artean, puntuazioaren ezberdintasunik handiena BALDINTZA erlazioan gertatu da. Orobat, lehen unitarearen puntuazioa 0,28koa da eta azken unitatearena 0,89. Hots, 0,71ko puntuazioaren tartea dago. Azkenik, BALDINTZA erlazioko, Ebaluazioa taldeko eta Kausa taldeko azken unitateek erlaziozko diskurtso-egitura osoaren orientazio semantikoarekin adostasun puntuaziorik handiena dute, puntuazioa 0,89koa baita BALDINTZAren kasuan eta 0,83koa Ebaluazioarenean eta Kausarenean.

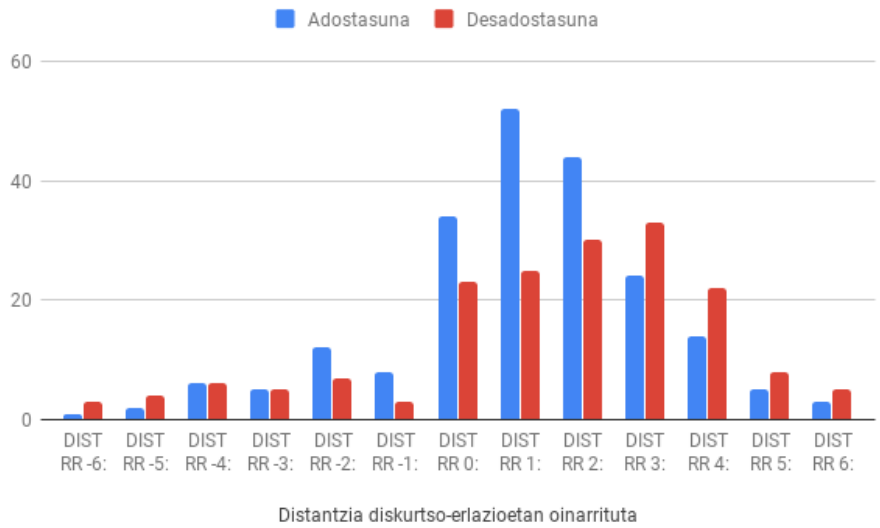
Laburbilduz, erlaziozko diskurtso-egituraren orientazio semantikoa baldintzatzen dute bai nukleartasunak, bai lehen edo azken unitatea izateak ere.

5.3.1.2. Orientazio semantikoa eta unitate zentrala

5.4 irudian, erlaziozko diskurtso-egitura guztiek beren testu osoarekin duten orientazio semantikoaren adostasun eta desadostasun instantziak agertzen dira. Unitate zentrala (DIST RR 0) da eta erlaziozko diskurtso-egiturak bere aurretik eta atzetik daude.

Erlaziozko diskurtso-egitura unitate zentraletik zenbat eta gertuago egon, erlaziozko diskurtso-egituren eta beren iritzi-testuen arteko orientazio semanti-

egiturek osatzen dute.



5.4 irudia: Testuek eta erlaziozko diskurtso-egiturek beren artean dituzten orientazio semantikoaren adostasun eta desadostasunak.

koaren adostasuna handiagoa da. −2, −1, 0, +1 eta +2 distantzietan orientazio semantikoaren adostasunaren instantzia gehiago daude desadostasunak[13] baino. Beste distantzietan, ordea, orientazio semantikoaren desadostasunaren instantziak gehiago dira edo bestela adostasun eta desadostasunen artean ez dago ezberdintasun nabarmenik.

Emaitzak erlaziozko diskurtso-egitura moten arabera aztertzen badira, zenbait ezberdintasun antzeman daitezke. Kausa taldearen kasuan, orientazioaren adostasunaren instantziak gehiago dira 0 eta +1 distantzietan bakarrik. Aurkaritza taldeak, aldiz, ez du erlaziozko diskurtso-egiturek distantzietan erakusten duten joera jarraitzen. Izan ere, Aurkaritza taldean, adostasun instantziak bi gune bereizitan dira nagusi. Hau da, adostasuna gune batean nagusi da (−2 distantzia); ondoren, desadostasun instantziak gehiago dira (0 distantzia) eta, azkenik, adostasun instantziak berriz nagusitzen dira (+1 eta +2 distantziak).

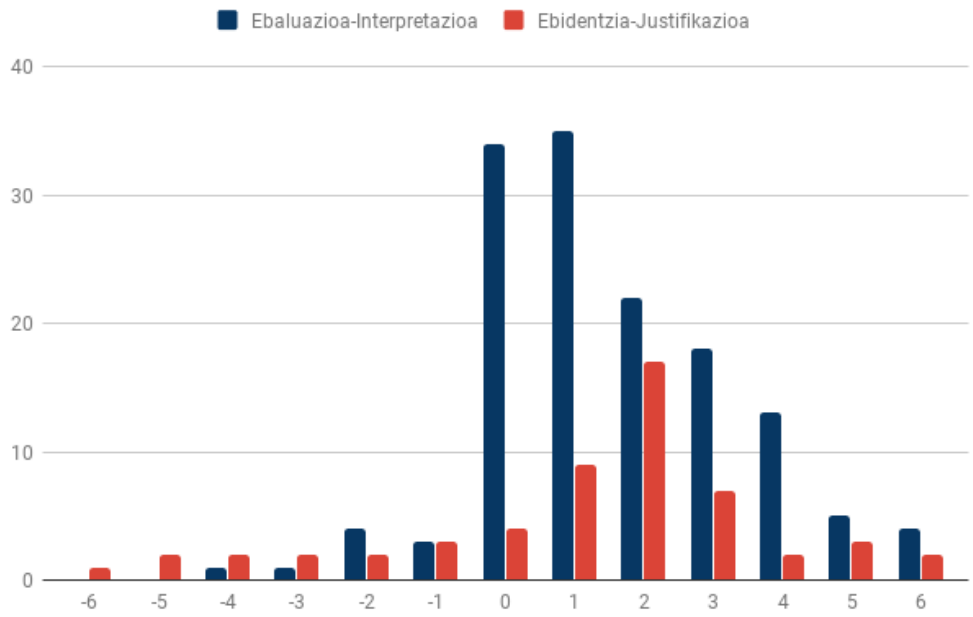
[13]Desadostasuna izendapenez, erlaziozko diskurtso-egiturak eta iritzi-testuak aurkako orientazio semantikoa duten instantziak adierazi nahi ditugu.

Laburbilduz, unitate zentralak erlaziozko diskurtso-egiturak eta iritzi-testuak duten orientazio semantikoaren adostasunean eragiten du, izan ere, erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, erlazioak duen orientazio semantikoak iritzi-testuak duenarekin gehiagotan bat egiten du. Gure ustez, unitate zentraletik gertuen dauden erlaziozko diskurtso-egituren gaiak iritzi-testuaren gai nagusitik hurbilago daudelako dago bien arteko orientazio semantikoaren adostasun handiagoa. Unitate zentraletik urrun dauden erlaziozko diskurtso-egituren, ordea, gaia iritzi-testuarengandik ezberdinak dira eta, ondorioz, orientazio semantikoan adostasuna ez da hainbestekoa.

5.3.1.3. Erlaziozko diskurtso-egituren agerpena distantzia bakoitzean

Iritzi-testuetan, erlaziozko diskurtso-egiturak distantzia bakoitzean zenbateko maiztasunez agertzen diren ere neurtu dugu. Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeen emaitzak aurkeztuko ditugu. Distantzia unitate zentraletik neurtu dugu. Distantzia negatiboa da unitate zentralaren ezkerretara eta positiboa unitate zentralaren eskuinetara. 5.5 irudian, Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeek instantzia gehien zer distantzietan duten beha daiteke.

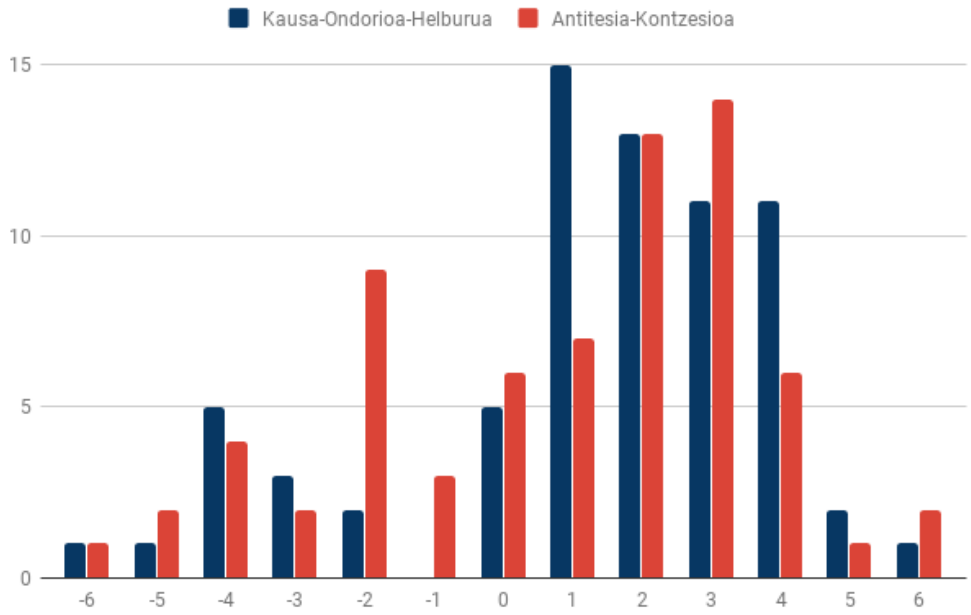
Ebaluazioa-Justifikazio erlaziozko diskurtso-egituraren taldearen instantzia gutxi daude unitate zentralaren aurretik, baina, 0 eta +1 distantzietan, bere instantzia kopurua asko igotzen da. Ondoren, berriz, instantzia kopurua jaitsi egiten da, unitate zentralaren aurreko egoerara itzuliz. KONTRASTEA erlaziozko diskurtso-egitura multinuklearrak ere modu berean jokatzen du. Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldean, ordea, distantzia guztietan dago erlaziozko diskurtso-egituraren talde horren instantziaren bat eta +1 eta +3 distantzien artean, bere instatzien kopurua nabarmen igotzen da. Beraz, Ebaluazioa-Interpretazioa erlaziozko diskurtso-egituraren taldeak distantzietan agerpen irregularra du eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldeak, aldiz, agerpen erregularra. Bestalde, 5.6 irudiak Kausa-Ondorioa-Helburua eta



5.5 irudia: Ebaluazioa-Interpretazioa eta Ebidentzia-Justifikazioa erlaziozko diskurtso-egituren instantziak.

Antitesia-Kontzetsioa erlaziozko diskurtso-egituraren taldeen emaitzak erakusten ditu.

Bi erlaziozko diskurtso-egitura taldeek ezaugarri batzuk partekatzen dituzte, unitate zentraletik duten instantzia handieneko guneei dagokienez. Ebidentzia-Justifikazioa erlaziozko diskurtso-egituraren taldean bezala, erlazio horiek unitate zentralarekiko distantzia guztietan agertzen dira. Baina erlazio taldeek instantzien kantitatean bi goreneko gune dituzte. Bi multzo horietan, goreneko uneetako bat unitate zentralaren aurretik dago eta bestea unitate zentralaren ondoren. Antitesia-Kontzetsioa erlazioaren kasuan, bi goreneko uneak -2 eta $+3$ distantzietan daude. Instantzia maiztasunen bi goreneko une -4 eta $+1$ distantzietan agertzen dira Kausa-Ondorioa-Helburua erlaziozko diskurtso-egituraren taldean.



5.6 irudia: Kausa-Ondorioa-Helburua eta Antitesia-Kontzetsioa erlaziozko diskurtso-egitura taldeen instantziak.

2 Ikerketa hipotesia

1.2 ataleko ikerketa hipotesiari erantzunez, euskaran hizkuntza maila ezberdinetan testuinguruko balentzia-aldatzaileak daude.
Ezeztapena eta diskurtso-egiturako balentzia-aldatzaileak beste hizkuntzetan ere badaude, baina euskararen berezitasuna da morfologiazko balentzia-aldatzaileetan duen aberastasuna.

3 Ikerketa hipotesia

1.2 ataleko hipotesiari erantzunez, erlaziozko diskurtso-egituran nukleartasuna eta iritzi-testuetan unitate zentrala daude erlaziozko diskurtso-egituraren edota bertako osagaien orientazio semantikoa zeinu jakin batekoa izatea eragiten dutenak.
Nuklearitateak eta erlaziozko diskurtso-egituretakoko azken testuzatiak orientazio semantikoaren adostasun handiagoa dute. Beraz, elementu horiek diskurtso-unitateek erlaziozko diskurtso-egituraren orientazio semantikoa duten adostasuna indartu egiten dute.
Unitate zentralak testuko erlaziozko diskurtso-egiturek iritzi-testuaren orientazio semantikoarekin adostasun handiagoa izatea eragiten du. Erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, bere eta iritzi-testuaren orientazio semantikoaren adostasuna handiagoa da, hots, adostasuna indartu egiten du.

5.3.1.4. Unitate zentralaren azterketa

Unitate zentralaren azterketan, i) ezaugarri sintaktikoak aintzat hartuta domeinu bakoitzeko unitate zentralaren karakterizazioa eta ii) sentimendubalentziadun hitzen banaketa domeinu ezberdinetako unitate zentraletan aurkeztuko ditugu.

Unitate zentralen ezaugarri sintaktikoak

5.11 taulan, domeinu bakoitzeko unitate zentralean maiztasun handienez agertzen diren gramatika-kategoriak ikus daitezke[14].

Domeinuka iritzi-testuen unitate zentralek dituzten ezaugarriak identifikatu ostean, domeinuen artean zenbait berdintasun antzeman ditugu. Eguraldiaren domeinuko unitate zentralak beste domeinukoetatik ezberdinak dira hiru alderditan. Batetik, eguraldiaren domeinuak, denbora aipatzea ohikoa da (esaterako: *gaur*, *bihar* edo *asteburua*). Bestetik, beste domeinuetan ez bezala, ez da entitaterik agertzen bertan. Azkenik, aditza ez agertzea ere

[14]Gramatika-kategoriez gain, entitateak, denbora eta determinatzailea ere kontuan hartu ditugu unitate zentraleko hitzak maiztasunaren arabera aztertzean, mota horretako hitzak asko agertu direlako.

	Ize.	Adj.	Adi.	Adb.	Det.	Entitateak
Eguraldia	X	X		X		
Kirola			X		X	X
Musika	X		X			X
Politika		X				X
Filmak	X	X	X			X
Liburuak	X	X	X		X	X

5.11 taula: Unitate zentraletan bereizgarri diren ezaugarriak domeinuen eta gramatika-kategorien arabera.

ezaugarri bereizle bat da. Kirolaren domeinuari dagokionez, bertan determinatzaile ugari (partidetako emaitzak) agertzen dira. Aditzean zer entitatek irabazi edo galdu egin duen zehazten da.

Musika, film eta liburuetako domeinuek zenbait antzekotasun partekatzen dituzte. Izenak, adjektiboak, aditzak eta entitateak azaltzen dira unitate zentraletan. Azkenik, politikan, entitateak eta adjektiboak beste gramatika-kategoriak baino dezentez gehiago agertzen dira eta beste domeinuetatik, kasu horretan, ezberdindu egiten da.

Datu horiek aintzat hartuz, iritzi-testuen domeinuak unitate zentraletan gramatika-kategoria mota bat bestea baino gehiagotan agertzea eragiten du. Eguraldia eta kirolak ezaugarri bereizleak dituzte unitate zentralean eta beste domeinuek tarteko ezaugarriak dituzte. Unitate zentralen azterketa hori baliagarria izan daiteke etorkizunean iritzi-testuetako unitate zentralak antzemateko tresna edo baliabideak sartzeko.

Sentimendu-balentziadun hitzen banaketa unitate zentraletan

Corpuseko 192 unitate zentraletako hitzei sentimendu-balentzia euskarazko sentimenduen sailkatzailea erabiliz esleitu zaie. 5.12 taulak esleipen horren emaitzak erakusten ditu.

Corpuseko unitate zentraletan, 2.081 hitzetatik 258 hitzek (% 12,40ak) dute sentimendu-balentzia. Gramatika-kategorien ikuspegitik, 99 adjektiboek dute sentimendu-balentzia, 77na izen eta aditzek eta azkenik, bost adberbik. Datuak ehunekoetan emanda, adjektiboek sentimendu-balentziadun hitzen % 38 hartzen dute; izen eta aditzek % 30 eta azkenik, adberbioek % 2.

		%
Hitzak guztira	2.081	
Balentziadun hitzak guztira	258	12,40
Adjektiboak	*99*	*38*
Aditzak	*77*	*30*
Izenak	*77*	*30*
Adberbioak	*5*	*2*

5.12 taula: Euskarazko sentimenduen sailkatzaileak sentimendu-balentzia esleitutako unitate zentraleko hitzak gramatika-kategoriaren arabera sailkatuta.

Beraz, adjektiboetan sentimendu-balentziadun hitz gehiago agertzen diren arren, banaketa nahiko orekatua da. Domeinutik domeinura dauden ezberdintasunak ikus daitezke 5.13 taulan.

	Eguraldia		**Kirola**		**Literatura**		**Musika**		**Filmak**		**Politika**		**Guztira**	
Izena	7	0,16	11	0,33	17	0,36	10	0,26	12	0,33	20	0,33	77	0,30
Aditza	12	0,27	16	**0,48**	12	0,25	13	0,34	10	0,28	14	0,23	77	0,30
Adjektiboa	22	**0,50**	5	0,15	18	**0,38**	15	**0,39**	14	**0,39**	25	**0,42**	99	**0,38**
Adberbioa	3	0,07	1	0,03	-		-		-		1	0,02	5	0,02
Guztira	44		33		47		38		36		60		258	

5.13 taula: Sentimendu-balentziadun hitzen agerpena.

Sentimendu-balentziadun hitzen kopuruan, politikaren domeinuan 60 sentimendu-balentziadun hitz agertzen dira eta kirolean, sentimendu-balentziadun hitzak 33 dira. Eguraldiaren domeinuan, adjektiboaren gramatika-kategoriak nabarmenki balentziadun hitzen agerpen gehiago ditu (sentimendu-balentzia duten hitzen erdiak adjektiboak dira). Kirolaren domeinuan, aditza da sentimendu-balentziadun hitz gehien dituen gramatika-kategoria (aditzak % 48 dira). Literaturan, musikan, filmeetan eta politikan adjektiboak dira sentimendu-balentziadun hitz gehien dituztenak. Adjektiboak balentziadun hitzen % 38-42 hartzen dute, baina beste gramatika-kategoriak datu horietatik gertu daude. Laburbilduz, balentziadun hitzak dituen gramatika-kategoria ohikoena adjektiboena da, indar handiagoz eguraldian eta oso indar gutxiz kirolean non bertan aditza nagusitzen den modu nabarmen batean.

Atal honetako emaitzak eta 5.3.1.4 atalean lortutakoak, sentimenduen analisira begira unitate zentralak identifikatzeko, beren sentimendu-balentzia

eta orientazioa semantikoa kalkulatzeko eta domeinu bakoitzeko unitate zentralean zer gramatika-kategoriari eman behar zaion garrantzia antzemateko balio dezake. Gerora, bildutako informazio hori baliagarria izan daiteke euskarazko sentimenduen sailkatzailean unitate zentrala lantzeko modulu bat sortzeko.

4 Ikerketa hipotesia

1.2 ataleko hipotesiari erantzunez, domeinutik domeinura unitate zentralak ezaugarri bereziak dituela egiaztatu dugu.
Kirolaren domeinuan aditzak garrantzi handia du eta eguraldiaren domeinuan, aldiz, adjektiboak. Entitateak ere garrantzitsuak dira hainbat domeinutan.
Sentimendu-balentzia duten gramatika-kategorien kasuan, kirolaren domeinuan, aditzak ohikoenak dira eta beste gramatika-kategorietan adjektiboak.

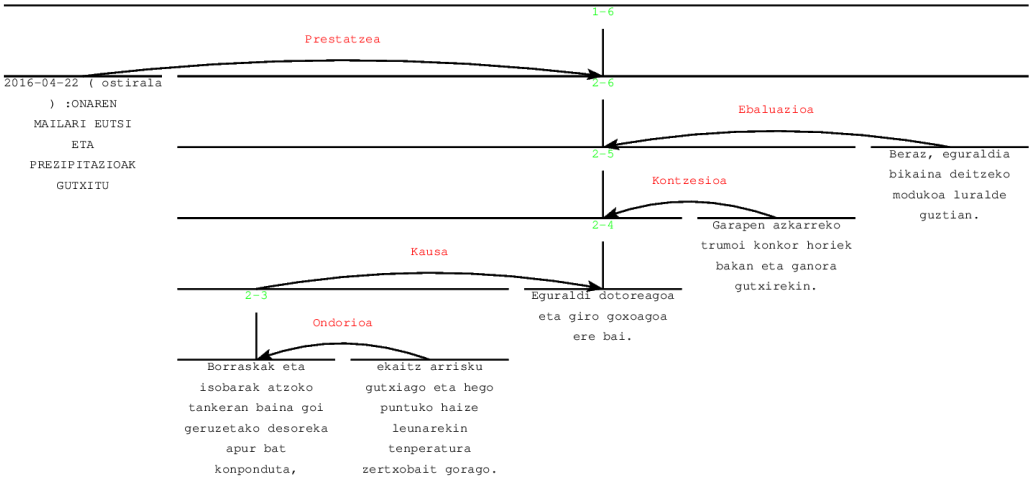
5.3.2. Egitura Erretorikoaren Teoria (RST)

Lan honetan, diskurtso-egitura aztertzeko Egitura Erretorikaren Teoria (*Rhetorical Structure Theory* edo RST) (Mann eta Thompson, 1988) hurbilpena hautatu dugu. Teoria honek testuaren egitura edo koherentzia deskribatzen du. Hizkuntzalaritza Konputazionalaren diziplinan eta erlaziozko diskurtso-egituran koherentzia aztertzeko gehien erabiltzen den teoria da eta euskaraz dagoen bakarrenetakoa da.

Hizkuntzalaritza Konputazionalean diskurtso-egitura lantzean bi fenomeno landu ohi dira: i) erreferentziazko fenomenoak: koerreferentziaren ebazpena lantzen da eta ii) erlaziozko fenomenoak: koherentzia erlazioen esleipena lantzen da. Egitura Erretorikoaren Teorian koherentzia erlazioak dira aztergai, baita gure lanean ere.

Mann eta Thompsonek (1988) RST konputagailu bidez testu-sorkuntza (*text generation*) egiteko asmoz sortu zuten. Testu-sorkuntza egin ahal izateko, testuaren ezaugarriak zehaztu behar dira eta testuen ezaugarririk garran-

tzitsuena koherentzia da. RSTk testuaren koherentzia deskribatzen duenez, konputagailua gauza izango litzateke testuen erlazio-egitura sortzeko edota prozesatzeko.



5.7 irudia: EGU06 iritzi-testua RSTz etiketatuta.

Testuaren egitura deskribatzeko, zenbait urrats bete behar dira. Hasteko, testuaren osagaiak identifikatu behar dira eta, ondoren, osagai horien arteko lotura edo erlazioa zehaztu. Testuan osagaia testu-zatia da. Testua aztertzeko testua zatitu egin behar da eta testua sortzeko testuko zatiak lotu egin behar dira. Testua aztertzean, testu-zatiak egiteari *segmentazioa* deritzo. Testu bat zatitzean, testu-zati bakoitzak proposizio bat adierazi behar du. 5.7 irudian, testua sei zatitan banatuta dago eta zati horietako bakoitzak proposizio bat adierazten du.

Ondoren, testu-zati horiek lotu egin behar dira eta hori egiteko koherentziazko erlazioak (*coherence relation*) ezartzen dira eta, testu-zatietan, koherentziazko erlazio horiek erlazio erretoriko (*rhetorical relation*) izena hartzen dute. RSTko zuhaitz-egituretan, erlazio erretorikoak errekurtsiboak dira. Hau da, erlazio bat beste erlazio baten unitatea da eta, modu horretan, testuko unitate guztiak zuhaitz-egituran aurkezten dira. 5.7 irudian, testu-zatiak erlazio erretorikoz (ONDORIOA, EBALUAZIOA, etab.) lotuta daude eta erlazioen artean errekurtsibitatea dago. Esaterako, ONDORIOA erlazio

erretorikoa KAUSA erlazio erretorikoaren unitatea da.

5.3.2.1. Nukleartasuna

RSTn nukleartasunaren bidez erlazio-egituretan unitate bakoitzak duen garrantzia zehazten da. RSTko zuhaitz-egituretan hierarkia bat osatzen da; izan ere, diskurtso-unitate guztiek ez dute garrantzi bera. Idazleak irakurlearri testuko mezua adierazi nahi dionean, unitate batzuek ideia garrantzitsuak dituzte eta beste batzuek, aldiz, bigarren mailako ideiak dira eta ideia nagusia ulertzeko baliagarriak dira. Beraz, diskurtso-unitate batzuk besteak baino garrantzitsuagoak dira. Diskurtso-unitate garrantzitsuenei nukleo-unitate (*nuclear unit*, N) deritze eta bigarren mailakoei satelite-unitate (*satellite unit*, S). Diskurtso-unitate bat nukleo- edo satelite-unitatea den ebazteari nukleartasuna (*nuclearity*) deritzo. 5.7 irudiko ONDORIOA erlazio erretorikoan, eskuineko diskurtso-unitatea ezkerrekoari lotzen zaio. Hori dela eta, ezkerreko unitatea nukleo-unitatea da eta eskuinekoa satelite-unitatea.

Nukleartasunak lotura du koherentziarekin, izan ere, idazleak testu batean adierazi nahi duena diskurtso-egiturako nukleo-unitateekin ulertzen da. Satelite-unitateekin, aldiz, testuaren ulermena zaila da, bertan dauden ideiak bigarren mailakoak direlako eta koherentziarik ez dagoelako. Erlazio nukleo-bakardunetan, satelitea nukleoari erlazio erretorikoz batzen zaio (N-S) eta erlazio horri *erlazio hipotaktiko* deritzo. Bi nukleoz (N-N) osatuta dauden erlazioei, aldiz, *erlazio parataktiko* deritze.

5.3.2.2. Erlazio erretorikoak

RSTn bi motatako erlazioak bereizten dira:

- Erlazio nukleoaniztunak (N-N). Mota honetako erlazioak nukleoz bakarrik osatuta daude eta, horrenbestez, erlazioa osatzen duten unitate guztiak maila berean daude eta unitateen artean ez da hierarkiarik sortzen. Erlazio nukleoaniztasunak hauek dira: KONJUNTZIOA, KONTRASTEA, DISJUNTZIOA, BATERATZEA, LISTA, BIRFORMULAZIO NUKLEOANIZTUNA eta SEKUENTZIA.

- Erlazio nukleobakardunak (N-S/S-N). Mota honetako erlazioetan, erlazioa osatzen duten unitateak ez daude maila berean eta hierarkia bat sortzen da. Nukleo-unitateak (N) adierazi nahi denaren zati garrantzitsuena biltzen du eta satelite-unitateak (S) nukleo-unitatean aipaturiko gaia hobeto ulertzen informazioa ematen du. Erlazio nukleobakardunak esaldien eta paragrafoen barruan ager daitezke. Mann eta Thompsonek (1988) bi motatako erlazio nukleobakardunak bereizten dituzte.

  – Edukizko erlazioak (*subject matter*). Unitateen arteko loturaren efektuak izaera semantikoa duenean gertatzen da. Hots, idazleak irakurleari bi unitateen artean erlazio bera dagoela adierazi nahi dio. ZIRKUNSTANTZIA, BALDINTZA, ELABORAZIOA, EBALUAZIOA, INTERPRETAZIOA, METODOA, KAUSA, ONDORIOA, AUKERA, HELBURUA, ARAZO-SOLUZIOA, EZ-BALDINTZATZAILEA eta ALDERANTZIZKO BALDINTZA mota honetako erlazioak dira.

  – Aurkezpenezko erlazioak (*presentational*). Unitateen arteko loturak izaera erretorikoa duenean gertatzen da. Kasu honetan, lotura horrek helburutzat irakurlearengan efektu jakin bat sortzea du. ANTITESIA, TESTUINGURUA, KONTZETSIOA, AHALBIDERATZEA, EBIDENTZIA, JUSTIFIKAZIOA, MOTIBAZIOA, PRESTATZEA, BIRFORMULAZIOA eta LABURPENA aurkezpenezko erlazioak dira.

5.3.2.3. Unitate zentrala

Iruskietak (2014) aipatzen duen moduan, pertsona batek testu baten laburpena egiten duenean, testuaren koherentzia globala edo makroegitura esplizitu egiteko gaitasuna erakusten du. Laburpen zientifiko eta akademikoetan ohikoak dira “The principal aim of this paper is to investigate ...” moduko egitura edo formulak. Horiek kasurik gehienetan adierazgailuak izan ohi dituzte eta adierazgailu horiek gramatika-kategoria ezberdinetakoak izan daitezke. Lan akademikoetan, ohikoak dira gramatika-kategoria ezberdinetako adierazgailu hauek: izenak (*paper*, *article*, *method*, *result* eta abar), aditzak (*discuss*, *introduce*, *analy-* eta *stud-*, besteak beste), erakusleak (*this*, *some*

eta abar) eta azkenik, izenordainak (*we*, *I*). Hala ere, adierazgailu guztiek ez dute beti makroegiturako gai nagusia adierazten; izan ere, mikroegitura adierazteko ere erabil baitaitezke.

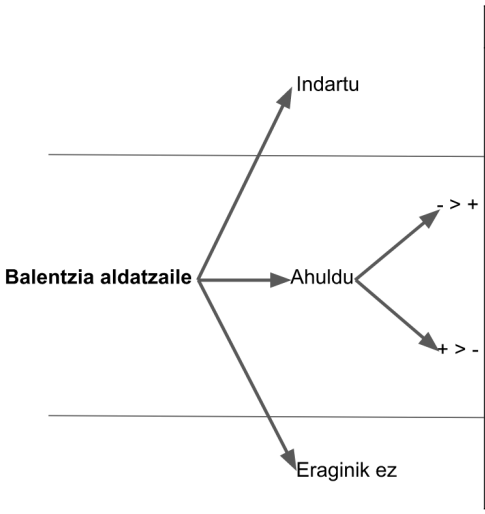
RST jarraituta eginiko zuhaitz-egitura batean, gai nagusia da erlazio-egiturako oinarrizko diskurtso-unitaterik (EDU) garrantzitsuena eta adierazgailuek hori identifikatzen lagun dezakete. Hau da, unitate zentrala da testuko EDU guztien artean garrantzitsuena eta zuhaitz-egitura eraiki nahi denean, bera erdigunetzat hartuta hasi behar da eraikitzen. Unitate zentralaren beste ezaugarri bat testuko diskurtso-unitate ez-errekurtsibo bakarra izatea da. Hots, unitate zentrala ez da inoiz beste erlazio baten satelite-unitatea izango.

RSTn oinarrituta egindako zuhaitz-egituran, beti egongo da unitate zentrala, nahiz eta tesi-adierazpen edo esaldi tematikorik ez egon. Izan ere, beti egon behar du zentralitatea duen nukleo-unitate bat eta ez-errekurtsiboa dena.

5.7 irudian, unitate zentrala *Eguraldia dotoreagoa eta giro goxoagoa ere bai* testu-zatia da. *Eguraldia* hitza makroegiturako gai nagusia izateak testu-zati hori unitate zentrala dela identifikatzen laguntzen digu.

5.4 Laburpena

Kapitulu honetan, euskaran hizkuntzako maila ezberdinetan dauden testuinguruko balentzia-aldatzaileak identifikatu eta horiek hitz eta sintagmen orientazio semantikoan eta sentimendu-balentzian duten eragina neurtu dugu. Emaitza 5.8 irudian ageri da.



Fonologia	Morfologia	Sintaxia	Diskurtsoa
<s>/<z> -> <x> <t> -> <tt> <z> -> <tx>	-garri -min -tsu -tza -en -tzar -zale -nahi -gura	ezin	· unitate zentrala · nukleoa · erlazioetako azken zatia
	-gabe des- -egi -keria ez-/-ez ezin-/-ezin -xe/-txe -txo -gaitz -ska/-xka -txa a-	ezin gabe izan ezik salbu ezta ez	
	-ezia -tasun	· edo/edota/ala + ez · ez ezik · egitura lexikalizatuak	

5.8 irudia: Euskarazko balentzia-aldatzaileak hizkuntza maila ezberdinetan.

Euskaran testuinguruko balentzia-aldatzaileak hizkuntza maila ezberdinetan daude eta beren eragina mota ezberdinetakoa da: hitzaren, sintagmaren edo esaldiaren sentimendu-balentzia indartu edo ahuldu egin dezakete. Halaber, badaude sentimendu-balentzian eraginik ez duten elementuak ere.

5.8 irudiko taulako lehen zutabean, euskarazko balentzia-aldatzaile fonologikoak zerrendatuta ageri dira. Bustidura adierazkorra balentzia-aldatzaile fonologikoa da eta beti hitzaren sentimendu-balentzia indartzen du. Izan ere, bustidura adierazkorra, samurtasuna edo gertutasuna adierazteko erabiltzen da eta mezuaren hartzaileari baliabide fonologiko hori erabiliz zuzentzean intentsitate indartu egiten da.

Bigarren zutabean, euskarazko hizkiak zerrendatuta eta sailkatuta ageri dira. Ahultzen duten balentzia-aldatzaile morfologikoak gehiago dira indar-

tzen dutenak baino. Kopuruaren ezberdintasunaren jatorrian ezeztapenezko hizkiak daude, horiek sentimendu-balentzia beti ahultzen baitute. Kontuan hartu behar da, ezeztapenezko hizkiek ±4 sentimendu-balentzia dutela, horrenbestez, hitzaren balentziaren zeinua alda dezakete baldin eta hitzaren sentimendu-balentzia ±3 edo baxuagoa bada.

Hirugarren zutabean, sintaxiko eta zehazki ezeztapeneko balentzia-aldatzaileak ageri dira. Ezeztapen-marka gehienek sintagmaren edo esaldiaren sentimendu-balentzia ahultzen dute. Kasu honetan ere, aintzat hartu behar da ezeztapen-marka horiek ±4 balentzia dutela eta sintagma edo esaldiaren sentimendu-balentzia ±3 edo txikiagoa bada, horien sentimendu-balentziaren zeinuan alderantzikatzea gertatzen da. Bestalde, *ezin* ezeztapen-markaren eskuinean adjektibo edo adberbioren bat badago, ezeztapen-marka horrek sentimendu-balentziaren indartzea eragina du. Bukatzeko, ezeztapen-markak ez du eraginik hautakariak diren lokailuekin agertzean, emendiozko lokailutzat edo ezeztapen kontrastibotzat jotzen den *ez ezik* egiturarekin agertzean eta ezeztapen-marka egitura lexikalizatu baten parte denean.

Azken zutabean, berriz, diskurtso-egiturari lotutako balentzia-aldatzaileak daude. Nukleartasunak, erlaziozko diskurtso-egituretako azken unitateak eta unitate zentralak eragina dute diskurtso-unitateek edota erlaziozko diskurtso-egiturek orientazio semantiko positiboa edo negatiboa izatean.

ONDORIOAK

6. KAPITULUA

Ekarpenak, mugak eta etorkizuneko lanak

Tesi-lan honen hasieran aipatu dugun moduan, gaur egun sarean iritzi-testu ugari egonda, testu horien informazio subjektiboa erauztea eta, zehazki, testu horiek balorazio positiboa edo negatiboa duten automatikoki jakitea abantaila bat da bai norbanakoentzat, bai gizartearentzat. Norabide horretan, euskarazko iritzi-testuetako informazio subjektiboa erauzteko tresna eta baliabideak garatu ditugu eta, halaber, euskarazko testuinguruko balentzia-aldatzaileak landu ditugu.

Jarraian, tesi-lan honetan euskarari dagokionez sentimenduen analisiaren inguruan egindako ekarpenak eta tesi-lanak izan dituen mugak aurkeztuko ditugu. Etorkizunera begira ditugun erronkak ere aipatuko ditugu.

6.1 Ekarpenak

Tesi hau euskarazko testu idatzien sentimenduen analisia lantzen duen lana da hizkuntzalaritza aplikatuaren ikuspegitik. Iritzi-testuak biltzen dituen corpus bat osatu eta sentimenduen lexikoi bat garatu ondoren, hitzen balentzietan eragiten duten fenomenoak aztertu ditugu. Jarduera horien ondorio dira garatu ditugun baliabideak eta tresnak.

6.1.1. Sentimenduen analisirako baliabideak eta tresnak

Euskarazko iritzi-testuen corpus etiketatua, euskarazko sentimenduen lexikoia eta dokumentu mailako sentimenduen sailkatzailea garatu ditugu.

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016). Corpus horrek sei domeinutako 240 iritzi-testu biltzen ditu. Diskurtso-egituraren nahiz subjektibitateaz etiketatuta dago. Diskurtso-egitura etiketatzeko RST hurbilpena (Mann eta Thompson, 1988) erabili dugu eta 70 iritzi-testu etiketatu ditugu. 192 iritzi-testuren unitate zentralak ere identifikatu ditugu. Halaber, testuek orientazio semantiko positiboa edo negatiboa duten ere zehaztu dugu. Gainera, 28 iritzi-testutan, erlaziozko diskurtso-egitura mota batzuei eta beren osagaiei (nukleoa eta satelitea) orientazio semantikoa (positiboa edo negatiboa den) nahiz sentimendu-balentzia (zenbakizko sentimenduen balioa) esleitu diegu.

- *Sentitegi* izeneko euskarazko sentimenduen lexikoia (Alkorta *et al.*, 2018). Sentimenduen lexikoi horrek 5 gramatika-kategoritako 1.237 hitz biltzen ditu. Bere oinarrian SO-CAL tresnaren gaztelaniazko lexikoia (Brooke *et al.*, 2009) dago eta ingelesezko lexikoiaz (Taboada *et al.*, 2011) ere aberastuta dago. Sentimenduen lexikoia sei domeinuetarako egokitu dugu (eguraldia, kirola, politika, musika, zinema eta literatura) eta, horretarako, Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) baliatu dugu.

- Tesi-lan honen beste ekarpen bat dokumentu mailako eta lexikoian oinarritutako sentimenduen sailkatzailea da. Euskarazko sentimenduen sailkatzailea garatzeko SO-CAL tresnan (Taboada *et al.*, 2011) oinarritu gara.

6.1.2. Euskarazko balentzia-aldatzaileen azterketa

Sentimenduen analisia lantzeko, hitzen sentimendu-balentzian eragiten duten hizkuntza maila ezberdinetako testuinguruko balentzia-aldatzaileak aztertu

ditugu. Balentzia-aldatzaileak aztertzerakoan, Polanyi eta Zaenenen (2006) lanean oinarritu gara.

- Fonologia eta morfologia mailako balentzia-aldatzaileak aztertu ditugu. Bertan ikusi dugunez, fonologian bustidura adierazkorra izeneko fenomenoa dago eta beren eragina da hitzaren sentimendu-balentzia indartzea. Morfologian, berriz, atzizkiak (adibidez, *-txo*/*-tzar*) aurkitu ditugu hitzaren balentzia indartu edo ahuldu dezaketenak. Gainera, ezeztapenarekin loturiko zenbait aurrizki (adibidez, *des-*) eta atzizki (adibidez, *-ezin*) ere aurkitu ditugu.

- Maila sintaktikoa ere aztertu dugu eta, zehazki, ezeztapen-markak eta irismena. Azterketak erakusten duenez, ezeztapen-markek hitzaren edo sintagmaren balentzia indartu (adibidez, *ezin*) edo ahuldu (esaterako, *gabe*) dezakete eta badaude aldaketarik eragiten ez duten kasuak (esaterako, *baino ez*) ere.

  Halaber, aipatzekoa da ezeztapen-marka batek (*ezin*) jokamolde bikoitza duela inguruan duen hitzaren gramatika-kategoriaren arabera. Ezeztapen-marka batzuk beste hitz jakin batzuekin maiztasun handiz agertzen direla ere erakutsi digute emaitzek. Egitura lexikalizatuak deitu diegu hauei. Irismenari dagokionez, *Murriztapen Gramatika* (Karlsson *et al.*, 1995) hurbilpena erabiliz, berau ezeztapen-markak eta egitura lexikalizatuak identifikatzea baino zailagoa dela behatu dugu, egitura eta luzera irregularrekoa baita.

- Azkenik, diskurtso maila ere landu dugu eta, horretarako, RST hurbilpena (Mann eta Thompson, 1988) erabili dugu. Lanean, erlaziozko diskurtso-egiturei eta unitate zentralari eman diogu garrantzia. Erlaziozko diskurtso-egituretan, ikusi dugu erlazioetako nukleoak edo azken EDUak osagai horien beraren eta erlaziozko diskurtso-egituraren arteko orientazio semantikoaren adostasuna indartzen dutela. Horretaz gain, emaitzek erakusten dute erlaziozko diskurtso-egitura bat unitate zentraletik gertuago egon ahala, erlazioaren beraren eta testu osoaren orientazio semantikoaren adostasuna indartu egiten dela. Horrenbes-

tez, emaitzek erakusten dute erlazioetan nukleoa eta azken EDUa eta testuan unitate zentrala balentzia indartzaile direla.

Unitate zentralean, balentziadun hitzak nola agertzen diren ere aztertu dugu. Emaitzen arabera, unitate zentraletako hitzen % 12,40k balentzia dute, batez ere, adjektibo, izen eta aditzek. Unitate zentralean, balentziadun hitzetan domeinuen arabera ezberdintasunak daudela behatu dugu. Esaterako, kirolaren domeinuan aditz-kategoriako hitzak dira balentziadunak, batik bat. Eguraldiaren domeinuan, ordea, adjektibo-kategoriako hitzak dira balentziadunak, batez ere.

6.2 Lanaren mugak

Atal honetan, tesi-lan honek izan dituen mugak aipatuko ditugu:

- Euskaran dauden testuinguruko balentzia-aldatzaile gehiago identifikatzen eta aztertzen jarraituko dugu. Tesi-lan honetan, hizkuntza maila bakoitzeko balentzia-aldatzaile bat aztertu dugu, baina oraindik badira aztertzeko geratu direnak. Hori da lan honek izan duen mugetariko bat. Aztertu gabe gelditu diren balentzia-aldatzaileen artean, baldintza, galde-perpausak edota konplexuagoa den ironia aipa ditzakegu. Balentzia-aldatzaileetako batzuek euskaran berezitasunak eduki ditzakete eta beste batzuk, aldiz, hizkuntza ezberdinetan antzekoak izan.

- Diskurtso mailako azterketa gehiago sakondu nahi dugu. Lan honetan, erlaziozko diskurtso-egiturak eta unitate zentrala aztertu ditugu. Baina, sentimenduen analisian eragin dezaketen diskurtso-egituran elementu gehiago egon daitezke. Horregatik, etorkizunean, haratago joan nahi dugu eta diskurtso-egiturako beste osagai batzuk ere aztertu nahi ditugu. Beste modu batean esanda, diskurtso-egituran dauden beste elementu batzuk aztertu nahi ditugu, azpiosagai zentrala esaterako. Halaber, erlaziozko diskurtso-egitura motetan ere sakondu nahi dugu eta beren artean egon daitezkeen ezberdintasunak aurkitu nahi ditugu.

6.3 Etorkizuneko lanak

Sentimenduen analisiaren azterketan eta beraren tratamendu konputazionalean hurrengo lan-ildoak aurreikusten dugu etorkizunerako:

- Euskarazko Iritzi Corpusa (Alkorta *et al.*, 2016) are anitzagoa izatea. Gaur-gaurkoz corpusak egunkari eta webgune espezializatutako 6 domeinutako iritzi-testuak biltzen ditu. Etorkizuneko asmoa jendeak idazten dituen iritzi-testuak biltzea da. Metodologiaren garapenean aipatu bezala, jende arruntak idatzitako iritzi-testu gutxi aurkitu ditugu edota aurkitu ditugunen kalitatea ez da egokia izan diskurtsoa lantzeko. Asmoa da mota horretako testuak biltzea, oraindik aztertu gabe dauden bestelako fenomenoak ager baitaitezke. Adibidez, jende arruntak idatzita testuetan posible da letra bat errepikatuz hitzak luzatzea, baita emotikonoak bezalako elementuak agertzea ere, eta horiek orientazio semantikoan ere eragina izan dezakete.

- Corpusaren diskurtso-egituraren etiketatzea hobetu eta zabaldu. Mementoz, corpuseko 240 testuetatik 70 daude etiketatuta eta kopuru hori handitu egin nahi dugu.

- *Sentitegi* sentimenduen lexikoia (Alkorta *et al.*, 2018) handitzea eta metodologian izandako gertaerei aurre egitea. Ikusi dugun moduan, sortu dugun lexikoiaren kalitatea ona da, baina oraindik ez da gauza subjektibitatea adierazten duten hitz edota esamolde guztiak identifikatzeko. Ondorioz, bere tamaina handitu nahi dugu eta domeinu gehiagotarako erabilgarria izatea lortu. Bukatzeko, hitz polisemikoen arazoa ere ebatzi nahi dugu, hau da, testuinguru ezberdinetan, aurkako bi orientazio dituzten hitzekin zer egin erabaki. Izan ere, *ikaragarria* bezalako hitzak aurkako orientazio semantikoa du, esaterako, pelikula bati buruz edo istripu bati buruz ari garenean.

- Dokumentu mailako sentimenduen sailkatzailea gehiago garatu nahi dugu. Tresna, mementoz euskarazko lexikoiaz, lematizatzaileaz eta erregela orokorrez osatuta dago, baina bertan aspektu gehiago integratu nahi ditugu. Esaterako, tesi-lan honetan landutako balentzia-

aldatzaileei lotutako modulua eratu nahiko genuke eta, modu horretan, tresnaren kalitatea igotzen den aztertu beharko litzateke. Modulu horretan, fonologiatik hasi eta diskurtso-egiturarainoko balentzia-aldatzaileei buruzko erregelak garatu nahiko genituzke.

Bibliografia

Aduriz I., Aldezabal I., Alegria I., Arriola J., Díaz de Ilarraza A., Ezeiza N., eta Gojenola K. Finite state applications for Basque. *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing*, 3–11, 2003.

Aduriz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M., eta Uria L. A Cascaded Syntactic Analyser for Basque. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 124–134, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24630-5.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., eta Maritxalar M. Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism. *Proceedings of Recent Advances in NLP (RANLP97), Tzigov Chark (Bulgary)*, 282–288, 1997.

Agerri R., Bermudez J., eta Rigau G. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In Chair) N.C.C., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., eta Lersundi M. EDBL: A general lexical basis for the automatic processing of

Basque. *Proceedings of the IRCS Workshop on linguistic databases*. IRCS Workshop on linguistic databases., 2006.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., eta Urizar R. Robustness and customisation in an analyser/lemmatiser for Basque. *LREC-2002 Customizing knowledge in NLP applications workshop*, 1–6, 2002.

Alistair K. eta Diana I. Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*, 2005.

Alkorta J., Gojenola K., eta Iruskieta M. Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela, September*, 29 lib., 2016.

Alkorta J., Gojenola K., eta Iruskieta M. SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas*, 22(4), 2018.

Altuna B., Aranzabe M.J., eta Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria eta Patxi Salaberri (arg.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea, 127-134, Udako Euskal Unibertsitatea (UEU), Bilbo*, 2017.

Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B.n., eta Rotaetxe K. Euskal Gramatika Lehen urratsak (EGLU) II. *Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo*, 1985.

Banea C., Mihalcea R., Wiebe J., eta Hassan S. Multilingual Subjectivity Analysis Using Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 127–135, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613715.1613734`.

Barnes J., Badia T., eta Lambert P. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018a. European Languages Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L18-1104`.

Barnes J., Klinger R., eta Schulte im Walde S. Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages. *CoRR*, abs/1805.09016, 2018b. URL `http://arxiv.org/abs/1805.09016`.

Bautin M., Vijayarenu L., eta Skiena S. International sentiment analysis for news and blogs. *ICWSM*, 19–26, 2008.

Beigman Klebanov B., Burstein J., Madnani N., Faulkner A., eta Tetreault J. Building subjectivity lexicon(s) from scratch for essay data. *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'12, 591–602, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28603-2. URL `http://dx.doi.org/10.1007/978-3-642-28604-9_48`.

Benesty J., Chen J., Huang Y., eta Cohen I. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4. Springer, 2009.

Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G., eta Reynar J. Building a Sentiment Summarizer for Local Service Reviews. *WWW Workshop on NLP Challenges in the Information Explosion Era (NLPIX)*, 2008.

Blitzer J., Dredze M., eta Pereira F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P07-1056`.

Boldrini E., Balahur A., Martínez-Barco P., eta Montoyo A. EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. *Proceedings of the Fourth Linguistic Annotation Workshop*, 1–10. Association for Computational Linguistics, 2010.

Bond F., Ohkuma T., Da Costa L.M., Miura Y., Chen R., Kuribayashi T., eta Wang W. A Multilingual Sentiment Corpus for Chinese, English and Japanese. *Emotion and Sentiment Analysis*, page 59, 2016.

Boucher J. eta Osgood C.E. The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, 8(1):1–8, 1969.

Brooke J., Tofiloski M., eta Taboada M. Cross-linguistic sentiment analysis: From English to Spanish. *Proceedings of the International Conference RANLP-2009*, 50–54, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/R09-1010`.

Cambria E., Das D., Bandyopadhyay S., eta Feraco A. *A Practical Guide to Sentiment Analysis*, 5 lib. 01 2017. ISBN 978-3-319-55392-4.

Carenini G., Ng R., eta Pauls A. Multi-Document Summarization of Evaluative Text. *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E06-1039`.

Chen Y. eta Skiena S. Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 383–389, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P14-2063`.

Clematide S., Gindl S., Klenner M., Petrakis S., Remus R., Ruppenhofer J., Waltinger U., eta Wiegand M. MLSA - A Multi-layered Reference Corpus for German Sentiment Analysis. In Calzolari N., Choukri K., Declerck T., Dogan M.U., Maegaard B., Mariani J., Odijk J., eta Piperidis S., editors,

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, 3551–3556, 2012.

Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Cruz F.L., Troyano J.A., Pontes B., eta Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994, 2014.

Dadvar M., Hauff C., eta de Jong F. Scope of negation detection in sentiment analysis. *Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011*, 16–20. University of Amsterdam, 2 2011. ISBN not assigned.

Das D. eta Taboada M. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184, March 2018. ISSN 1574-020X. URL `https://doi.org/10.1007/s10579-017-9383-x`.

Das D. eta Martins A.F. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4(192-195): 57, 2007.

Ding X., Liu B., eta Yu P.S. A Holistic Lexicon-based Approach to Opinion Mining. *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, 231–240, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-2. URL `http://doi.acm.org/10.1145/1341531.1341561`.

Ding X., Liu B., eta Zhang L. Entity discovery and assignment for opinion mining applications. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 1125–1134, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. URL `http://doi.acm.org/10.1145/1557019.1557141`.

Du W., Tan S., Cheng X., eta Yun X. Adapting Information Bottleneck Method for Automatic Construction of Domain-oriented Sentiment Lexicon. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 111–120, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-889-6. URL http://doi.acm.org/10.1145/1718487.1718502.

Edmonds P. eta Hirst G. Near-synonymy and Lexical Choice. *Comput. Linguist.*, 28(2):105–144, June 2002. ISSN 0891-2017. URL http://dx.doi.org/10.1162/089120102760173625.

Egaña I. *Kritikarako hurbilketa literaturaren soziologiatik. Egunkari eta aldizkarietako euskal literatur kritikaren analisia (1975-2005).* Doktoretzatesia, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2013.

Egg M. eta Redeker G. How Complex is Discourse Structure? *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/796_Paper.pdf.

Eguchi K. eta Lavrenko V. Sentiment Retrieval Using Generative Models. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 345–354, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610124.

Elhuyar H.Z. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. Usurbil: Elhuyar, 2013.

Euskaltzaindia. Euskal Gramatika. Lehen urratsak I, Iruñea: Euskaltzaindia, 1985.

Euskara Institutua E. Sareko euskal gramatika, 2011. URL www.ehu.eus/seg.

Fernández J., Boldrini E., Gómez J.M., eta Martínez-Barco P. Análisis de sentimientos y minería de opiniones: el corpus emotiblog. *Procesamiento del lenguaje natural*, 47:179–187, 2011.

Fundazioa K.H. *Euskal hiztegi entziklopedikoa*. Klaudio Harluxet Fundazioa, 1995.

Gamon M. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `https://doi.org/10.3115/1220355.1220476`.

Gamon M., Aue A., Corston-Oliver S., eta Ringger E. Pulse: Mining Customer Opinions from Free Text. *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, 121–132, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-28795-7, 978-3-540-28795-7. URL `http://dx.doi.org/10.1007/11552253_12`.

Ghosh A., Li G., Veale T., Rosso P., Shutova E., Barnden J., eta Reyes A. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 470–478, Denver, Colorado, jun 2015. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/S15-2080`.

Guo H., Zhu H., Guo Z., Zhang X., eta Su Z. Opinionit: A text mining system for cross-lingual opinion analysis. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 1199–1208, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. URL `http://doi.acm.org/10.1145/1871437.1871589`.

Halliday M.A.K., Matthiessen C., eta Halliday M. *An introduction to functional grammar*. Routledge, 2014.

Hassan A., Abu-Jbara A., Lu W., eta Radev D. A random walk: Based model for identifying semantic orientation. *Comput. Linguist.*, 40(3):539–

562, September 2014. ISSN 0891-2017. URL http://dx.doi.org/10.1162/COLI_a_00192.

Hatzivassiloglou V. eta McKeown K.R. Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. URL https://doi.org/10.3115/976909.979640.

Hu M. eta Liu B. Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. URL http://doi.acm.org/10.1145/1014052.1014073.

Hu Y.H., Chen Y.L., eta Chou H.L. Opinion Mining from Online Hotel Reviews A Text Summarization Approach. *Inf. Process. Manage.*, 53(2): 436–449, March 2017. ISSN 0306-4573. URL https://doi.org/10.1016/j.ipm.2016.12.002.

Iruskieta M. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, Informatika Fakultatea*, 2014.

Iruskieta M., Aranzabe M.J., Díaz de Ilarraza A., Gonzalez I., Lersundi M., eta Lopez de Lacalle O. The RST Basque TreeBank: an online search interface to check rhetorical relations. *4th workshop RST and discourse studies*, 40–49, 2013.

Iruskieta M., da Cunha I., eta Taboada M. Principles of a qualitative method for rhetorical analysis evaluation: A contrastive analysis English-Spanish-Basque. *Language Resources and Evaluation*, 49(2):263–309, 2015.

Jia L., Yu C., eta Meng W. The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. *Proceedings of the 18th ACM Confe-*

rence on Information and Knowledge Management, CIKM '09, 1827–1830, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. URL http://doi.acm.org/10.1145/1645953.1646241.

Jindal N. eta Liu B. Mining Comparative Sentences and Relations. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, 1331–1336. AAAI Press, 2006. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597348.1597400.

Kanayama H. eta Nasukawa T. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 355–363, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610125.

Karlsson F., Voutilainen A., Heikkila J., eta Anttila A., editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter & Co., Hawthorne, NJ, USA, 1995. ISBN 3110141795.

Karoui J., Farah B., Moriceau V., Patti V., Bosco C., eta Aussenac-Gilles N. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 262–272. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/E17-1025.

Kennedy A. eta Inkpen D. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 22:110–125, 05 2006.

Kim J., Li J.J., eta Lee J.H. Evaluating Multilanguage-comparability of Subjectivity Analysis Systems. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 595–603, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858681.1858742.

Kim S.M. eta Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, 1–8, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-75-2. URL http://dl.acm.org/citation.cfm?id=1654641.1654642.

Kim S.M., Pantel P., Chklovski T., eta Pennacchiotti M. Automatically assessing review helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 423–430, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610135.

Kouloumpis E., Wilson T., eta Moore J. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.

Landis J.R. eta Koch G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 1977.

Leturia I., Arregi X., eta Sarasola K. Weba euskarazko corpus gisa. *Ekaia*, 27:281, 12 2014.

Li F., Huang M., Yang Y., eta Zhu X. Learning to Identify Review Spam. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, 2488–2493. AAAI Press, 2011. ISBN 978-1-57735-515-1. URL http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-414.

Li H., Cheng X., Adson K., Kirshboim T., eta Xu F. Annotating opinions in German political news. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 1183–1188, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/640_Paper.pdf.

Lim E.P., Nguyen V.A., Jindal N., Liu B., eta Lauw H.W. Detecting Product Review Spammers Using Rating Behaviors. *Proceedings of the 19th*

ACM International Conference on Information and Knowledge Management, CIKM '10, 939–948, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. URL http://doi.acm.org/10.1145/1871437.1871557.

Liu B. Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

Liu B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Liu B., Hu M., eta Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, 342–351, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9. URL http://doi.acm.org/10.1145/1060745.1060797.

Liu F., Li B., eta Liu Y. Finding opinionated blogs using statistical classifiers and lexical features. *Third International AAAI Conference on Weblogs and Social Media*, 2009.

Liu F., Wang D., Li B., eta Liu Y. Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 309–312, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N10-1042.

Long C., Zhang J., eta Zhut X. A Review Selection Approach for Accurate Feature Rating Estimation. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 766–774, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944654.

Lu Y., Tsaparas P., Ntoulas A., eta Polanyi L. Exploiting Social Context for Review Quality Prediction. *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 691–700, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. URL http://doi.acm.org/10.1145/1772690.1772761.

Mann W.C. eta Thompson S.A. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

McDonald R., Hannan K., Neylon T., Wells M., eta Reynar J. Structured Models for Fine-to-Coarse Sentiment Analysis. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 432–439. Association for Computational Linguistics, 2007. URL `http://aclweb.org/anthology/P07-1055`.

Meng X., Wei F., Xu G., Zhang L., Liu X., Zhou M., eta Wang H. Lost in Translations? Building Sentiment Lexicons using Context Based Machine Translation. *Proceedings of COLING 2012: Posters*, 829–838, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `https://www.aclweb.org/anthology/C12-2081`.

Miao Y., Su J., Liu S., eta Wu K. *SO-CAL Based Method for Chinese Sentiment Analysis*, 345–351. 12 2013.

Miller G.A. WordNet: A Lexical Database for English. *Commun. ACM*, 38 (11):39–41, November 1995. ISSN 0001-0782. URL `http://doi.acm.org/10.1145/219717.219748`.

Mohammad S.M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, 201–237. Elsevier, 2016.

Mohammad S.M., Salameh M., eta Kiritchenko S. How Translation Alters Sentiment. *J. Artif. Int. Res.*, 55(1):95–130, January 2016. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=3013558.3013562`.

Morsy S.A. eta Rafea A. Improving Document-level Sentiment Classification Using Contextual Valence Shifters. *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, NLDB'12, 253–258, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-31177-2. URL `http://dx.doi.org/10.1007/978-3-642-31178-9_30`.

Mujika L.M. Morfología de la composición lexical euskérica. *Fontes linguae vasconum: Studia et documenta*, 14(39):233–272, 1982.

Mullen T. eta Collier N. Sentiment analysis using support vector machines with diverse information sources. *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2004.

Na S.H., Lee Y., Nam S.H., eta Lee J.H. Improving Opinion Retrieval Based on Query-Specific Sentiment Lexicon. In Boughanem M., Berrut C., Mothe J., eta Soule-Dupuy C., editors, *Advances in Information Retrieval*, 734–738, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-00958-7.

Nakagawa T., Inui K., eta Kurohashi S. Dependency Tree-based Sentiment Classification Using CRFs with Hidden Variables. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 786–794, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858119`.

Ngoc Phu V. eta Thi Tuoi P. Sentiment classification using Enhanced Contextual Valence Shifters. 224–229, 10 2014.

O'Donnell M. RSTTool 2.4: A Markup Tool for Rhetorical Structure Theory. *Proceedings of the First International Conference on Natural Language Generation - Volume 14*, INLG '00, 253–256, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. ISBN 965-90296-0-8. URL `https://doi.org/10.3115/1118253.1118290`.

Oñederra L. *Euskal fonologia: palatalizazioa: asimilazioa eta hots sinbolismoa*. Servicio Editorial de la Universidad del País Vasco= Euskal Herriko Unibertsitatea, 1990.

Oronoz M. *Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura*. Doktoretzatesia, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2009.

Otegi A., Imaz O., Díaz de Ilarraza A., Iruskieta M., eta Uria L. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84, 2017.

Pang B. eta Lee L. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 115–124, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. URL `https://doi.org/10.3115/1219840.1219855`.

Pang B. eta Lee L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008. ISSN 1554-0669. URL `http://dx.doi.org/10.1561/1500000011`.

Pang B., Lee L., eta Vaithyanathan S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL `https://doi.org/10.3115/1118693.1118704`.

Polanyi L. eta Zaenen A. Contextual valence shifters. 20 lib., 1–10. 01 2006.

Qu L., Ifrim G., eta Weikum G. The Bag-of-opinions Method for Review Rating Prediction from Sparse Text Patterns. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 913–921, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1873781.1873884`.

Quan C. eta Ren F. Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, 1446–1454, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-63-3. URL `http://dl.acm.org/citation.cfm?id=1699648.1699691`.

Rao D. eta Ravichandran D. Semi-supervised Polarity Lexicon Induction. *Proceedings of the 12th Conference of the European Chapter of the As-*

sociation for Computational Linguistics, EACL '09, 675–682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609142.

Refaee E. eta Rieser V. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In Chair) N.C.C., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Riloff E., Patwardhan S., eta Wiebe J. Feature subsumption for opinion analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 440–448, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL http://dl.acm.org/citation.cfm?id=1610075.1610137.

Rodriguez I., Martínez-Otzeta J.M., Lazkano E., eta Ruiz T. Adaptive Emotional Chatting Behavior to Increase the Sociability of Robots. *International Conference on Social Robotics*, 666–675. Springer, 2017.

Rushdi-Saleh M., Martínez-Valdivia M.T., Ureña;a López L.A., eta Perea-Ortega J.M. OCA: Opinion Corpus for Arabic. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):2045–2054, October 2011. ISSN 1532-2882. URL http://dx.doi.org/10.1002/asi.21598.

San Vicente I., Saralegi X., eta Agerri R. EliXa: A modular and flexible ABSA platform. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 748–752, Denver, Colorado, June 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S15-2127.

Saralegi X., San Vicente I., eta Ugarteburu I. Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-resourced Languages. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, 96–108,

Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 978-3-642-37255-1. URL http://dx.doi.org/10.1007/978-3-642-37256-8_9.

Sarasola I. *Zehazki: gaztelania-euskara hiztegia*. Alberdania, 2005.

Sauri R. *A Factuality Profiler for Eventualities in Text*. Doktoretza-tesia, Waltham, MA, USA, 2008. AAI3304029.

Schulz J.M., Womser-Hacker C., eta Mandl T. Multilingual Corpus Development for Opinion Mining. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., eta Tapias D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Shin H., Kim M., Jang H., eta Cattle A. Annotation Scheme for Constructing Sentiment Corpus in Korean. *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 181–190. Faculty of Computer Science, Universitas Indonesia, 2012. URL http://aclweb.org/anthology/Y12-1019.

Silvennoinen O.O. Not only apples but also oranges: Contrastive negation and register. *Studies in Variation, Contacts and Change in English*, 19, 2017.

Snyder B. eta Barzilay R. Multiple Aspect Ranking Using the Good Grief Algorithm. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 300–307, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N07-1038.

Sorby S. Translating news from English to Chinese: Complimentary and derogatory language usage. *Babel*, 54:19–35, 01 2008.

Stone P.J., Dunphy D.C., Smith M.S., eta Ogilvie D.M. The General Inquirer: A Computer Approach to Content Analysis. 1966.

Stone P.J. eta Hunt E.B. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), 241–256, New York, NY, USA, 1963. ACM. URL http://doi.acm.org/10.1145/1461551.1461583.

Szmrecsanyi B. *Analyticity and syntheticity in the history of English*. Oxford University Press, 2012.

Taboada M. SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University, 2008.

Taboada M., Brooke J., Tofiloski M., Voll K., eta Stede M. Lexicon-based Methods for Sentiment Analysis. *Comput. Linguist.*, 37(2):267–307, June 2011. ISSN 0891-2017. URL http://dx.doi.org/10.1162/COLI_a_00049.

Tan P.N., Steinbach M., eta Kumar V. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005. ISBN 0321321367.

Tsur O. eta Rappoport A. RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews, 2009. URL https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/180.

Turney P.D. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL https://doi.org/10.3115/1073083.1073153.

Turney P.D. eta Littman M.L. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. *CoRR*, cs.LG/0212012, 2002. URL http://arxiv.org/abs/cs.LG/0212012.

Turney P.D. eta Littman M.L. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Trans. Inf. Syst.*, 21(4):

315–346, October 2003. ISSN 1046-8188. URL http://doi.acm.org/10.1145/944012.944013.

Vicente I.S. eta Saralegi X. Polarity lexicon building: to what extent is the manual effort worth? In Chair) N.C.C., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., eta Piperidis S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Vilares D., Garcia M., Alonso M.A., eta Gómez-Rodríguez C. Towards Syntactic Iberian Polarity Classification. *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 67–73. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/W17-5209.

Villarroel Ordenes F., Ludwig S., De Ruyter K., Grewal D., eta Wetzels M. Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research*, 43(6):875–894, 2017.

Vinodhini G. eta Chandrasekaran R. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

Wan X. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 553–561, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613783.

Wan X. Co-training for Cross-lingual Sentiment Classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, 235–243, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9. URL http://dl.acm.org/citation.cfm?id=1687878.1687913.

Wang G., Xie S., Liu B., eta Yu P.S. Identify Online Store Review Spammers via Social Review Graph. *ACM Trans. Intell. Syst. Technol.*, 3(4):61:1–61:21, September 2012. ISSN 2157-6904. URL http://doi.acm.org/10.1145/2337542.2337546.

Wei B. eta Pal C. Cross Lingual Adaptation: An Experiment on Sentiment Classifications. *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, 258–262, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858842.1858890.

Westerski A. Sentiment Analysis: Introduction and the State of the Art overview. *Universidad Politecnica de Madrid, Spain*, 211–218, 2007.

Wiebe J. Learning Subjective Adjectives from Corpora. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 735–740. AAAI Press, 2000. ISBN 0-262-51112-6. URL http://dl.acm.org/citation.cfm?id=647288.721121.

Wiebe J. eta Mihalcea R. Word Sense and Subjectivity. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 1065–1072, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL https://doi.org/10.3115/1220175.1220309.

Wiebe J., Wilson T., Bruce R., Bell M., eta Martin M. Learning Subjective Language. *Comput. Linguist.*, 30(3):277–308, September 2004. ISSN 0891-2017. URL http://dx.doi.org/10.1162/0891201041850885.

Wiebe J.M., Bruce R.F., eta O'Hara T.P. Development and Use of a Gold-standard Data Set for Subjectivity Classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. URL https://doi.org/10.3115/1034678.1034721.

Wiegand M., Balahur A., Roth B., Klakow D., eta Montoyo A. A Survey on the Role of Negation in Sentiment Analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, 60–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858959.1858970`.

Wilson T.A. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. Doktoretza-tesia, University of Pittsburgh, 2008.

Yu H. eta Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL `https://doi.org/10.3115/1119355.1119372`.

Zhang W., Yu C., eta Meng W. Opinion Retrieval from Blogs. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 831–840, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. URL `http://doi.acm.org/10.1145/1321440.1321555`.

Terminologia eta laburdurak

Sentimenduen analisi (*Sentiment analysis*)

Orientazio semantiko (*Semantic orientation*)

Sentimendu-balentzi (*Sentiment valence*)

(Testuinguruko) balentzia-aldatzaile (*(Contextual) valence shifter*)

Sentimenduen sailkatzaile (*Sentiment classifier*)

Sentimenduen lexikoi (*Sentiment lexicon*)

Murriztapen Gramatika, MG (*Constraint Grammar, CG*)

Ezeztapen-marka (*Negation mark*)

(Ezeztapenaren) irismena (*Scope*)

Egitura lexikalizatu (*Lexicalized structure*)

Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory, RST*)

Unitate zentrala, UZ (*Central Unit, CU*)

Erlaziozko diskurtso-egitura (*Relational discourse structure*)

ERANSKINA

A. KAPITULUA

Murriztapen Gramatikako erregelak

Eranskin honetan ezeztapen-markak eta beren irismena identifikatzeko erregelak aurkezten ditugu. Erregelak Murriztapen Gramatika (MG) hurbilpenean idatzirik daude.

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT
   PUNT_GALD PUNT_ESKL PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = ``ez'';
LIST EZA = ``ez'';

LIST BAINO = ``baino'';
LIST BESTERIK = ``beste'';
LIST ZALANTZARIK = ``zalantza'';
LIST DUDARIK = ``duda'';
LIST BESTERIKGABE = ``besterik_gabe'';
LIST EZINEZKOA = ``ezinezko'';
LIST EZINEAN = ``ezinean'' "ezin'';
LIST NORA = ``nora'';
LIST EZEAN = ``ezean'';
```

```
LIST NORAEZEAN = ``nora_ezean'';

LIST EZINIZAN = ``ezin_izan'';
LIST EZIN = ``ezin" ``ezindu'';
LIST EZINHOBEA = ``ezin_hobe'';
LIST EZINIK = ``ezinik'';

LIST SALBU = ``salbu'';
LIST EZEZIK = ``ez_ezik'';
LIST IZANEZIK = ``izan_ezik'';

LIST GABE = ``gabe'';

LIST JUNTAGAILUA = ``edo'' ``ala'' ``edota'';
LIST BAI = "bai'';

LIST EZTA = ``ezta'';

#MAPPINGS

MAP (!gabe) TARGET (ADB) IF (0C GABE);
MAP (!ez) TARGET (PRT) IF (0C EZ);
MAP (!eza) TARGET (IZE) IF (0C EZA);

#LEXIKALIZAZIOAK

# (1)
#``baino ez'' egitura
MAP (!bainoezHAS) TARGET (LOT) IF (0C BAINO)    (1C EZ);

# (2)
# ``besterik ez'' egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)    (1
   C EZ);
```

```
#MAP (!besterikezBUK) TARGET (PRT) IF (-1C BESTERIK)
    (0C EZ);

# (3)
# ``zalantzarik gabe''
MAP (!zalantzarikgabeHAS) TARGET (IZE) IF (0C
    ZALANTZARIK) (1C GABE);
#MAP (!zalantzarikgabeBUK) TARGET (ADB) IF (-1C
    ZALANTZARIK) (0C GABE);

# (4)
# ``dudarik gabe''
MAP (!dudarikgabeHAS) TARGET (IZE) IF (0C DUDARIK) (1C
    GABE);
#MAP (!dudarikgabeBUK) TARGET (ADB) IF (-1C DUDARIK) (0
    C GABE);

# (5)
# ``besterik gabe'' antzemateko
MAP (!besterikgabe) TARGET (LOT) IF (0C BESTERIKGABE);

# (6)
# ``ezinezkoa'' antzemateko
MAP (!ezinezkoa) TARGET (ADJ) IF (0 EZINEZKOA);

#(7)
# ezinean antzemateko
MAP (!ezinean) TARGET (ADB) IF (0C EZINEAN);

#(8)
# ``nora ezean'' egitura.
MAP (!noraezeanHAS) TARGET (ADB) IF (0C NORA) (1C EZEAN
    );
MAP (!noraezeanBUK) TARGET (ADB) IF (-1C NORA) (0C
```

```
    EZEAN);
MAP (!noraezean) TARGET (ADB) IF (0C NORAEZEAN);

##ERREGELA ZEHATZAK

#(1)
# EZIN: ``ezin izan''
MAP (!ezinizanHAS) TARGET (ADI) IF (0C EZINIZAN);
MAP (!ezinizanTAR1) TARGET (ADL) OR (ADI) OR (ADT) OR (
    IZE) OR (ADJ) OR (DET) OR (IOR) OR (LOT) IF (-1
    EZINIZAN);
MAP (!ezinizanTAR2) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-2 EZINIZAN);
#MAP (!ezinizanBUK1) TARGET (ADI) OR (IZE) IF (-2
    EZINIZAN);
MAP (!ezinizanBUK2) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-3 EZINIZAN) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK3) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-4 EZINEAN) (NOT
     -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK4) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-5 EZINIZAN) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK5) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-6 EZINIZAN) (
    NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3
    PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!ezinizanBUK6) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-7 EZINIZAN) (
    NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA) (NOT -4
    PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA)
    (NOT -1 PUNTUAZIOA);
```

```
MAP (!ezinizanBUK7) TARGET (ADI) OR (ADL) OR (IZE) OR (
    ADJ) OR (DET) OR (IOR) OR (LOT) IF (-8 EZINIZAN) (
    NOT -7 PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);


## EZIN orokorra
MAP (!EZIN) TARGET (IZE) IF (0C EZIN);
MAP (!ezinik) TARGET (ADB) IF (0C EZINIK);

# EZIN: ``ezinik''
MAP (!ezinik2) TARGET (ADI) IF (1C EZINIK);
MAP (!ezinik1) TARGET (IZE) IF (2C EZINIK);
##KOPONduta

# EZIN: ``ezin'' + beste aditzak
MAP (!ezinTAR) TARGET (ADT) OR (ADI) OR (ADL) OR (
    ADI_IZEELI) IF (-1C EZIN);
MAP (!ezinTAR2) TARGET (ADI) OR (ADL) OR (ADI_IZEELI)
    IF (-2C EZIN);
MAP (!ezinBUK) TARGET (ADI) OR (ADL) OR (ADI_IZEELI) IF
     (-3C EZIN);



# EZIN: ``ezin hobea''
MAP (!ezinhobea) TARGET (ADJ) IF (0C EZINHOBEA);
#MAP (!ezinadjHAS) TARGET (IZE) OR (ADB) IF (0C EZIN);
MAP (!ezinadjBUK) TARGET (ADJ) OR (ADB) IF (-1C EZIN);

# SALBU
MAP (!salbuBUK) TARGET (ADB) IF (0C SALBU);
MAP (!salbuHAS1) TARGET (IZE) OR (ADT) IF (1C SALBU);
MAP (!salbuHAS3) TARGET (IZE) OR (DET) IF (2C SALBU);
```

```
MAP (!salbuHAS3) TARGET (IZE) IF (3C SALBU);

# IZAN EZIK
MAP (!izanezik) TARGET (LOT) IF (0C IZANEZIK);
MAP (!izanezikHAS2) TARGET (IZE) IF (4C IZANEZIK);
MAP (!izanezikHAS1) TARGET (DET) IF (3C IZANEZIK);
MAP (!izanezikTAR1) TARGET (IZE) OR (DET) IF (2C
   IZANEZIK);
MAP (!izanezikTAR2) TARGET (ADI) OR (IZE) (1C IZANEZIK)
   ;

# EZ EZIK
MAP (!ezezik) TARGET (LOT) IF (0C EZEZIK);

#(2)
# EDO/EDOTA/ALA EZ
MAP (!juntagailuaHAS1) TARGET (ADI) OR (ADL) OR (ADT)
   OR (PRT) IF (1C JUNTAGAILUA) (2C EZ);
MAP (!juntagailuaHAS2) TARGET (ADI) OR (ADL) OR (IOR)
   IF (2C JUNTAGAILUA) (3C EZ);
MAP (!juntagailuaHAS3) TARGET (ADI) OR (IZE) IF (3C
   JUNTAGAILUA) (4C EZ);
MAP (!juntagailuaTAR) TARGET (LOT) IF (0C JUNTAGAILUA)
   (1C EZ);
#MAP (!juntagailuaBUK) TARGET (PRT) IF (-1C JUNTAGAILUA
   ) (0C EZ);

# EZ... EZ...
MAP (!ezez0) TARGET (PRT) IF (0C EZ) (2C EZ);
MAP (!ezez1) TARGET (DET) OR (IZE) OR (ADJ) IF (-1C EZ)
    (1C EZ);
MAP (!ezez2) TARGET (PRT) IF (-2C EZ) (0C EZ);
MAP (!ezez3) TARGET (DET) OR (IZE) OR (ADJ) IF (-1C EZ)
    (-3C EZ);
```

```
# EZ egiturak
MAP (!EZ-3) TARGET (IZE) IF (3C EZ) (NOT 2 BAINO) (NOT
    2 BESTERIK) (NOT 1 PUNTUAZIOA) (NOT 2 PUNTUAZIOA);
MAP (!EZ-2) TARGET (IZE) OR (ADI) OR (ADB) IF (2C EZ) (
    NOT 1C BAINO) (NOT 1C BESTERIK) (NOT 1C PUNTUAZIOA);
MAP (!EZ-1) TARGET (ADI) OR (ADJ) OR (ADL) OR (ADB) IF
    (1C EZ);
#MAP (!EZ0) TARGET (PRT) IF (0C EZ);
MAP (!EZ1) TARGET (ADT) OR (ADL) OR (ADI) OR (ADB) OR (
    IZE) OR (DET) OR (IOR) IF (-1C EZ) (NOT -2 BAINO) (
    NOT -2 BESTERIK);
MAP (!EZ2) TARGET (ADI) OR (ADL) OR (ADT) OR (DET) OR (
    IZE) OR (ADJ) OR (ADB) OR (IOR) IF (-2C EZ) (NOT -3C
     BAINO) (NOT -3C BESTERIK) (NOT -2 PUNTUAZIOA) (NOT
    -1 PUNTUAZIOA);
MAP (!EZ3) TARGET (ADI) OR (ADL) OR (ADT) OR (IZE) OR (
    ADJ) OR (IOR) IF (-3C EZ) (NOT -4C BAINO) (NOT -4
    BESTERIK) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (
    NOT -1 PUNTUAZIOA);
MAP (!EZ4) TARGET (ADI) OR (ADL) OR (IZE) OR (DET) OR (
    IOR) IF (-4C EZ) (NOT -5C BAINO) (NOT -5 BESTERIK) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ5) TARGET (ADI) OR (IZE) OR (IOR) OR (ADT) OR (
    ADL) IF (-5C EZ) (NOT -6C BAINO) (NOT -6 BESTERIK) (
    NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ6) TARGET (ADI) OR (IZE) OR (IOR) IF (-6C EZ) (
    NOT -7C BAINO) (NOT -7 BESTERIK) (NOT -5 PUNTUAZIOA)
     (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ7) TARGET (ADI) OR (IZE) OR (IOR) IF (-7C EZ) (
    NOT -8C BAINO) (NOT -8 BESTERIK) (NOT -6 PUNTUAZIOA)
```

```
    (NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3
   PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ8) TARGET (ADI) OR (IZE) OR (ADL) OR (IOR) IF
   (-8C EZ) (NOT -9C BAINO) (NOT -9 BESTERIK) (NOT -7
   PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
   (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
   PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZ9) TARGET (ADI) IF (-9C EZ) (NOT -10C BAINO) (
   NOT -10 BESTERIK) (NOT -8 PUNTUAZIOA) (NOT -7
   PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
   (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
   PUNTUAZIOA) (NOT -1 PUNTUAZIOA);

MAP (!izeezaHAS) TARGET (IZE) IF (1C EZA);
MAP (!izenezaBUK) TARGET (IZE) IF (0C EZA);



# GABE egiturak

MAP (!gabeAUR) TARGET (IZE) OR (ADI) OR (ADJ) OR (IOR)
   IF (1C GABE);
MAP (!gabeAUR2) TARGET (IZE) OR (ADI) IF (2C GABE);
MAP (!gabeAUR3) TARGET (IZE) IF (3C GABE);

# EZTA
#MAP (!EZTA0) TARGET (LOT) IF (0C EZTA);
MAP (!EZTA1) TARGET (ADT) OR (ADL) OR (ADI) OR (ADB) OR
    (IZE) OR (DET) OR (IOR) IF (-1C EZTA) (NOT -2 BAINO
   ) (NOT -2 BESTERIK);
MAP (!EZTA2) TARGET (ADI) OR (ADL) OR (ADT) OR (DET) OR
    (IZE) OR (ADJ) OR (ADB) OR (IOR) IF (-2C EZTA) (NOT
    -3C BAINO) (NOT -3C BESTERIK) (NOT -2 PUNTUAZIOA) (
   NOT -1 PUNTUAZIOA);
MAP (!EZTA3) TARGET (ADI) OR (ADL) OR (ADT) OR (IZE) OR
```

```
    (ADJ) OR (IOR) IF (-3C EZTA) (NOT -4C BAINO) (NOT
    -4 BESTERIK) (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA)
    (NOT -1 PUNTUAZIOA);
MAP (!EZTA4) TARGET (ADI) OR (ADL) OR (IZE) OR (DET) OR
    (IOR) IF (-4C EZTA) (NOT -5C BAINO) (NOT -5
    BESTERIK) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA5) TARGET (ADI) OR (IZE) OR (IOR) OR (ADT) OR
    (ADL) IF (-5C EZTA) (NOT -6C BAINO) (NOT -6
    BESTERIK) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (
    NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA6) TARGET (ADI) OR (IZE) OR (IOR) IF (-6C
    EZTA) (NOT -7C BAINO) (NOT -7 BESTERIK) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA7) TARGET (ADI) OR (IZE) OR (IOR) IF (-7C
    EZTA) (NOT -8C BAINO) (NOT -8 BESTERIK) (NOT -6
    PUNTUAZIOA) (NOT -5 PUNTUAZIOA) (NOT -4 PUNTUAZIOA)
    (NOT -3 PUNTUAZIOA) (NOT -2 PUNTUAZIOA) (NOT -1
    PUNTUAZIOA);
MAP (!EZTA8) TARGET (ADI) OR (IZE) OR (ADL) OR (IOR) IF
    (-8C EZTA) (NOT -9C BAINO) (NOT -9 BESTERIK) (NOT
    -7 PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5
    PUNTUAZIOA) (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA)
    (NOT -2 PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
MAP (!EZTA9) TARGET (ADI) IF (-9C EZTA) (NOT -10C BAINO
    ) (NOT -10 BESTERIK) (NOT -8 PUNTUAZIOA) (NOT -7
    PUNTUAZIOA) (NOT -6 PUNTUAZIOA) (NOT -5 PUNTUAZIOA)
    (NOT -4 PUNTUAZIOA) (NOT -3 PUNTUAZIOA) (NOT -2
    PUNTUAZIOA) (NOT -1 PUNTUAZIOA);
```

Tesi honen idazketa
2019ko urriaren 15ean
bukatu zen.

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Didactics of Language and Literature

PhD dissertation

Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels

---

Jon Alkorta Agirrezabala

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Didactics of Language and Literature

Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels

This summary is a shortened and translated version of the dissertation entitled "Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa", written by Jon Alkorta under the supervision of Dr. Mikel Iruskieta and Dr. Koldo Gojenola. It also includes the papers which the candidate has published in English on the research presented here.

Donostia, 15th of October 2019

Aitari, Amari eta Josuri

Acknowledgements

Lehenik eta behin, nire zuzendariak eskertu nahiko nituzke: Koldo eta Mikel, eman didaten laguntzagatik. Eskerrik asko beti nire atzean egoteagatik eta baita tesian zehar sortu diren zalantzetan laguntza eskaintzeagatik ere. Azkenik, eskerrik asko egin dizkidazuen iruzkinengatik (asko lagundu baitidate) eta ikerkuntzaren mundua nolakoa den erakusteagatik ere. Halaber, eskerrak Maxuxi eta Rodriri tesi-txostena hobetzen laguntzeagatik.

Eskerrik asko tesian lagundu didaten beste pertsonei ere. Eskerrik asko Arantxari *Eustaggerr*ekin izan ditudan une zailetan laguntzeagatik eta Intxari *include*k Latexen duen garrantzia erakusteagatik. Eskerrak Estherri Kanadatik Ixarako konexioa ahalbidetzeagatik, Kikeri emandako laguntzagatik eta Amaiari administrazioaren munduan laguntzeagatik. Eskerrik asko, halaber, Itziar Gonzalez-Diosi *Murriztapen Gramatika*rekin eta tesi-txostena vs. Latex borrokan laguntzeagatik. Eskerrik asko zerrenda honetan aipatu gabe gelditu diren baina lagundu didaten beste guztiei ere.

Eskertzekoa da tesia hasi eta amaitu arte inguruan mugitu denari ere: 318 bulegoari giro ona sortzeagatik, tupper txokoari askotariko gaiak (batzuetan, frikiak) aipatzeagatik, pintxo-poteetan eta egindako beste planetan atera zareten guztioi eta, azkenik, estatistika klaseetan, Oierri izandako pazientziagatik eta *nukleo duro*ri estatistika ikasteko laguntza morala emateagatik.

También tengo que agradecer a Maite Taboada por su ayuda en la estancia de Vancouver y en los trabajos relacionados con el discurso y análisis de sentimientos.

Familia ere ezin da aipatu gabe utzi. Eskerrik asko aitari, amari eta Josuri beti hor egoteagatik, eta laguntza behar izan dudanean laguntzeagatik.

Eskerrik asko kuadrillari ere, asteburuetan, musa eta *Comunio*rekin une

biziki ziragarriak emateagatik. Eskerrak baita Urkizu eta Gazteluko bazkariengatik eta Kanpezura eta Jakatik mendira egindako planengatik ere.

Bukatzeko eskerrik asko Amets eta Lierni pisukideei eta aipatzea ahaztu zaizkidan beste guztiei.

Mila-mila esker denoi!

Official acknowledgments

This dissertation was funded by the PRE_2014_1_190, PRE_2015_1_0121, PRE_2016_2_0153, PRE_2017_2_0041, EP_2017_1_0003 and PRE_-2018_2_0033 grants awarded by the Basque Government's Education, Universities and Research Department as well as by IXA group, Research Group of type A (2010-2015) (Basque Government, ref.: IT344-10) and READERS (CHIST-ERA, MINECO, ref.: PCIN-2013-003-C02-02).

Abstract

In the research work, we have taken the first steps on sentiment analysis from point of view of applied linguistics. The work developed consists of two aspects. On the one hand, based on the contextual valence shifter approach to sentiment analysis, we have identified valence shifters of different language levels in Basque, from phonology to discourse through morphology and syntax. Moreover, we have measured their effect on the sentiment valence of different linguistic elements.

The second aspect of this work focuses on the creation and development of tools and resources for sentiment analysis in Basque. Firstly, a corpus with 240 opinion texts has been built and it has been annotated: from the point of view of semantic orientation and discourse information. Secondly, a sentiment lexicon with 1,237 entries has been created. Finally, a document level and lexicon based sentiment classifier has been created based on the SO-CAL tool.

Contents

List of Figures

List of Tables

INTRODUCTION

1 Introduction

1.1 Motivation

Opinions are the basis of human activity and they have a significant impact on our behavior (Liu, 2012). Even our choices and beliefs about reality are largely conditioned by what the world sees and evaluates. As a result of this situation, we often seek the opinions of others when we make a decision. It is a situation that occurs in individuals and organizations.

In opinion texts, there are some concepts related to subjectivity. These include sentiment, evaluation, attitudes, and emotions, among others. All of them, in addition to opinion, are objects of study in sentiment analysis or opinion mining.

As Liu (2012) reminds us, the beginning and growth of this area is entirely associated with the creation of social networks. The social networks of the web are varied: comments, forum discussions, blogs, microblogs, and Twitter. All of these applications have made available a large volume of opinions in digital resources for the first time in human history. Therefore, it is one of the largest and most active growth areas in natural language processing since 2000. Sentiment analysis is also used in data mining, web mining and text mining.

In general, sentiment analysis has become important in business and society, which has expanded the field from computational science to management science and social science. Around it, the industry has developed and companies have also created services to work the area. So, it can be said that sentiment analysis can be found in many business or social fields.

As we have seen, sentiment analysis internationally has undergone considerable development. English is the world's *lingua franca* today and the majority of the works in sentiment analysis have been done in that language. In contrast, little has been done in Basque.

Sentiment analysis is a very broad area. Some tasks are connected with computer science and others with linguistics. Besides, some task involves two approaches: based on language knowledge and based on statistical methods.

From a linguistic point of view, the first step is to find out the opinion of the word or, in other words, the subjective information. The next step is to look for language-related phenomena that affect the subjective information of these words. Let us look at the following examples.

(1) I $[\text{like}_{+}]_{+}$ that movie.

(2) I $[\text{like a lot}_{+}]_{++}$ that movie.

(3) I do not $[\text{like}_{+}]_{-}$ that movie.

(4) I $[\text{would like}_{+}]_{0}$ very much that movie if it were more vivacious.

In Example (1), it is mentioned that a person liked the movie, so the opinion of the sentence is positive. In the same way, in Example (2), the opinion of the sentence is positive but the intensifier[1] *a lot* makes this sentence more positive than the previous sentence. In contrast, in Example (3), although the opinion of the verb (*like*) is positive, the opinion of the sentence is negative. Finally, in Example (4) there is only one word (*like*) with positive opinion but the sentence does not express an opinion, because the opinion is in a conditional mood.

In these examples, although there is one word that expresses opinion in all sentences (*like*), the opinion expressed by the whole sentence can be different. Some linguistic phenomena modify the meaning of words and with opinion. The phenomena that change the value of a word or phrase are called contextual valence shifters.

This aspect mentioned motivates our work. On the one hand, we aim to create tools and resources related to sentiment analysis for Basque to extract subjective content from Basque texts. On the other hand, we seek linguistic phenomena that influence words or phrases with opinion, from phonology to discourse.

[1]Intensifiers are usually adverbs or adjectives. They have little inherent semantic content but they intensify the meaning of the words or phrases.

Our first aim is to create basic tools and resources for sentiment analysis in Basque. Specifically, we want to create *i)* a corpus of opinion texts, *ii)* a sentiment lexicon and *iii)* a document-level sentiment classifier.

There are few corpora of opinion texts and sentiment lexicons in Basque and the current tools are not useful for our goals, because they do not identify relevant language features. In the case of opinion texts, some corpora or databases collect reviews and tweets that appear on websites but they are not useful to us because texts are short and, as a result, it is difficult to analyze certain elements of discourse structure.

In terms of lexicons, the available lexicons indicate the semantic orientation of words (for instance, Stone and Hunt (1963) indicates if the word is positive or negative) but they do not indicate intensity (for example, *good* and *excellent* are positive but their difference in intensity is not indicated) and therefore, they are not useful to us.

As far as the sentiment classifier is concerned, there are few works in Basque language that perform sentiment classification. One of them is EliXa (San Vicente et al., 2017) and it performs an aspect-level sentiment classification[2]. Our objective is to provide Basque with another resource for language processing and, for this purpose, we want to create a document-level sentiment classifier based on lexicon. It must be said that to meet this objective, it is necessary to fulfill the aforementioned objectives: it is necessary to create a sentiment lexicon to integrate in sentiment classifier and to study valence shifters is also needed to measure the changes that affect words with opinion.

Our second objective is to study linguistic phenomena that affect the sentiment valence of the words of lexicons created in the previous step and to measure their influence on words.

As we have seen in Examples (1), (2), (3) and (4), for a document-level classifier based on the sentiment lexicon, the lexicon alone is not enough. More linguistic information is needed to make a good sentiment classification. Therefore, based on the work of Polanyi and Zaenen (2006), we will aim to identify contextual valence shifters[3] at different language levels in Basque.

In short, this thesis combines Basque and sentiment analysis. There has been little work done on Basque in sentiment analysis and most are limited to

[2]In aspect-level sentiment analysis, the aim is to identify aspects related to the entity under study (if the entity is London, the aspects are economics, tourism, etc.) and to determine whether it is positive or negative.

[3]Contextual valence shifters are phenomena that cause changes in the sentiment valence of words and/or sentences. This change can be a strengthening or weakening of the valence.

the creation of a sentiment lexicon. We want to provide the Basque language with resources and tools, as well as investigate the features of Basque that affect sentiment classification.

1.2 General hypotheses

In the previous section, we have mentioned sentiment analysis as a motivation for the thesis, as well as its utility in society. In this section, we will outline the scope of the thesis. This work has four general hypotheses:

- **Research Hypothesis 1:** *The quality of the sentiment lexicon in Basque obtained by translation is comparable to those that can be derived from the corpus or lexical databases.*
  Sentiment lexicons assign the semantic orientation and sentiment valence to words of a texts and this is the first step in calculating the semantic orientation of a text. There are two common approaches to creating lexicons: *i)* corpus and *ii)* lexical databases.

  Our approach is a mixture of both approaches. We will use two sentiment lexicons (Spanish and English lexicons) as a basis and we will translate them by enriching the information provided by the Basque Opinion Corpus.

  In our opinion, the sentiment lexicon that has already created is comparable in quality to corpus-based and lexical database-based lexicons. If the translation methodology is optimal, no information will be lost or transformed along and it would maintain its initial quality. Besides, the use of a corpus with opinion texts would prevent the incorrect assignments of semantic orientation and sentiment valence to words with opinion.

- **Research Hypothesis 2:** *At all levels of language (from phonology to discourse) certain linguistic phenomena influence influence the sentiment valence of words (and phrases).*
  We think that this effect can either strengthen or weaken their sentiment valence. In our opinion, these linguistic phenomena can appear at different levels of grammar: for example, through affixes in phonology, through different syntactic phenomena or even through discourse. We believe it is necessary to use this kind of information in the creation

of a document-based sentiment classifier, otherwise, the results of the classifier could be poorer.

- **Research Hypothesis 3**: *In the discourse structure, there are constituents that affect the semantic orientation of EDUs and discourse relations.*
  From our point of view, in the case of EDUs or discourse relations, having a positive or negative semantic orientation is not a random event. In other words, we believe that these elements are affected by their position in the rhetorical structure tree. Specifically, we hypothesize that some or all of the factors in the discourse structure affect the semantic orientation of EDUs and discourse relations.

- **Research Hypothesis 4:** *Due to the domain, the central units in texts are different regarding the grammatical category semantic orientation of words.*
  Although the central unit is present in all texts, we believe that its features differ in domains. We think that the domain could affect in the way the words with semantic orientation appear in central units.

1.3 Goals

In the previous sections, we discussed the context of this work and the general hypotheses to guide our research. In this section, we will outline our specific goals to verify if they support our assumptions.

- **Objective 1:** Basic tools for studying sentiment analysis and creation of tools and resources.
  - ***Objective 1.1:*** To create a corpus of opinion texts in Basque, enriched with the semantic orientation and discourse information using *Rhetorical Structure Theory* (RST).
  - ***Objective 1.2:*** To generate a sentiment lexicon in Basque, indicating the semantic orientation of words by a numerical value.
  - ***Objective 1.3:*** To create a document-based sentiment classifier for Basque, based on sentiment lexicon.

- ***Objective 2:*** Identify and measure the effect of contextual valence shifters in Basque that affect the valence and sentiment of words:

- ***Objective 2.1:*** To identify phonological valence shifters and measure their effect.
- ***Objective 2.2:*** To identify morphological valence shifters and measure their effect.
- ***Objective 2.3:*** Regarding syntax, to measure how negation marks affect the sentiment valence of words or phrases.
- ***Objective 2.4:*** Regarding discourse structure, to measure the effect of the central unit on opinion texts and the effect of nuclearity on discourse relations.
- ***Objective 2.5:*** To study the appearance of discourse relations in the RST-tree.
- ***Objective 2.6:*** Analysis of central units of opinion texts: to study the grammatical category of words and the distribution of words with semantic orientation in central units.

1.4 Organization of the thesis report

This thesis consists of six chapters. In the following lines, we will summarize the contents of each chapter:

- **Chapter 1: Introduction.**
  In chapter one, we first discuss work motivation, general assumptions, goals, and research questions.

- **Chapter 2: Methodology and resources.**
  In this chapter, we will discuss the methodology and resources used to accomplish the goals mentioned in the introduction and to answer the research questions. Following an overview of the methodology, we will outline the steps to create a corpus of opinion text in Basque and a sentiment lexicon. After that, we will define the methodology for identifying the valence shifters. We will explain individually what has been done in phonology, morphology, syntax, and discourse. We will then explain the steps to create the document-level sentiment classifier in Basque.

- **Chapter 3: Resources developed.**
  In the third chapter, we will discuss the resources developed and their

evaluation as a result of following the defined methodology. Some of the resources we have mentioned include the Basque Opinion Corpus, a sentiment lexicon called *Sentitegi*, and a document-level sentiment classifier.

- **Chapter 4: Valence shifters at different language levels.**
  In this fourth chapter, we will also report on the results obtained. Specifically, we will explain the contextual valence shifters we have found at the phonological and morphological, syntactic, and discourse levels. Moreover, we have also measured their influence on the semantic orientation of words and/or sentences, or the sentiment valence of them.

- **Chapter 5: Conclusions and future work.**
  This chapter will first discuss the implications of the thesis work. The initial objectives will be mentioned and research questions will be answered. The contributions of the work will also be mentioned. Finally, we will focus on future work.

- **Terminology and abbreviations.**
  In this appendix, we will list the terminology and abbreviations mentioned in this thesis report.

METHODOLOGY

2 Methodology

We followed a specific methodology to accomplish the goals set out and we also used some specific resources for that. In this chapter, we will deal with them. There are three main steps in the methodology:

1- Create basic resources for carrying out research. In other words, we explain the creation of opinion text corpus and sentiment lexicon in Basque.

2- Identify the valence shifters. We analyze and measure the linguistic phenomena that may affect words with opinions at different language levels and their effects.

3- Develop a document-level sentiment classifier in Basque. This sentiment classifier will be based on a lexicon we developed and a sentiment classifier of the SO-CAL tool (Taboada et al., 2011).

2.1 Creating resources for sentiment analysis

In the first step, we created resources for sentiment analysis. We created the Basque Opinion Corpus and the sentiment lexicon in Basque.

2.1.1 The Basque Opinion Corpus

To create the Basque Opinion Corpus, we first had to decide what type of corpus we wanted to create. We decided to use the features of the corpus'

texts as follows:

- A clear appraisal of the opinion texts. Since some of the texts did not show a clear opinion, we preferred texts with an opinion either for or against.
- To have resources to express a rich syntactic structure and a clear appraisal. We wanted the opinion texts to have different syntactic structures and different methods of appraisal.

After, we collected Basque opinion texts from websites. These websites are either specialized or belong to magazines and newspapers. In short, we used three different sources: *i)* newspapers, *ii)* specialist websites and, finally, *iii)* several blogs.

After finding websites to provide the Basque Opinion Corpus, we established the corpus' characteristics:

- The corpus will be large and will contain 240 opinion texts.
- The corpus will cover many fields. Opinion articles will belong to six fields: weather, politics, sport, movies, music and literature.
- The corpus' texts will be appraised in a balanced manner. Each field will be balanced concerning its appraisal.

In the next step, we created a database with the collected opinion texts. In that database, we specified the following criteria for each opinion text:

- Code: We gave each opinion text a code as follows: *THEME-number appraisal*.
- Title: A title was given to each opinion text.
- Address: We stated where the opinion text is located on the Internet, showing its source to allow anyone to access it.
- Appraisal: In the database, we specified whether each opinion text was either positive or negative.
- Size: We also specified the number of words in each opinion text, using the *Analhitza* tool (Otegi et al. 2017).

Furthermore, we also adapted the format of opinion texts so they could be processed using natural language tools. We connected articles to *txt* format and UTF-8 codification formats. The development phase set contains 144 opinion texts and the training set 48.

We also wanted to measure whether the corpus is suitable for study and therefore compared it with another subjective corpus in English (*SFU Review Corpus*), and two objective corpora in both English and Basque (same subject as in the subjective corpus). We studied the following aspects:

- Presence of the first person. In English with personal pronouns and in Basque with verbs, we were able to measure the extent to which the first person was present. We studied this because people usually talk about their experiences in the first person singular or plural.

- Presence of adjectives. In all the grammatical categories, we measured the weight of adjectives in both the objective and subjective corpora. We analyzed the presence of adjectives because they are the most widely used grammatical category when expressing feelings.

- Negation marks. We wanted to study negation marks in Basque because they can influence combinations of feelings in words or phrases. We measured how many negation marks appeared in the Basque Opinion Corpus. In this case too, we were able to do this using computer-based resources.

- Finally, we tagged the Basque Opinion Corpus for discourse and subjectivity information. First, out of 240 texts, 70 were tagged using *Rhetorical Structure Theory* (RST).

	A1	**A2**	**Annotated texts in total**	**Double annotation**
Movies	30	9	30	9
Weather	15	5	15	5
Literature	5	25	25	5
Total	**50**	**39**	**70**	**19**

Table 2.1 – The number of opinion texts annotated by two annotators (A1 and A2).

As we can see in Table 2.1, overall 70 texts (29.16% of the corpus) were tagged and of these 19 (27.14% of those tagged) were tagged by two annotators. To complete the annotation, two taggers had to follow the criteria of Das and Taboada (2018). In different fields, there were certain differences concerning the tagging process. For example, annotators had to spend approximately 20 minutes on weather-related texts, whilst literary texts required around one hour to tag. The results of correspondences between taggers can be seen in Table 2.2. More precisely, we measured whether they correctly tagged the type of discourse relations.

Domain	**Agreement (%)**
Weather	43.59 (17/39)
Literature	41.67 (70/168)
Movies	37.73 (83/220)
Total	**39.81 (170/427)**

Table 2.2 – Inter-annotator agreement in the annotation of texts using *RST* approach (Mann and Thompson, 1988).

As the results in Table 2.2 show, when tagging the type of discourse relation the correspondence between two annotators is 39.81%. There were variations from one field to another. For weather, there was 43.59% agreement, whereas, with regard to cinema, this agreement dropped to 37.73%. For literature, the agreement is 37.73%.

	Components		**Attachment**		**Nuclearity**		**Relation**	
Domain	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**
Weather	20/37	0.54	9/37	0.24	22/37	0.59	15/37	0.41
Literature	84/155	0.54	67/155	0,43	105/155	0.68	48/155	0.31
Movies	112/221	0.56	88/221	0.40	147/221	0.67	68/221	0.31
Total	**216/413**	**0.52**	**164/413**	**0.40**	**274/413**	**0.66**	**131/413**	**0.32**

Table 2.3 – Qualitative evaluation of automatic tools for inter-annotator agreement.

We also used a qualitative evaluation. As Iruskieta et al. (2015) refers, this type of evaluation is called qualitative because it allows comparisons of discourse structures developed in different languages and/or by different people. How they have been evaluated so far have been quantitative, but the difference is that EDUs take into account textual parts, nuclearity and discourse relations. tool to measure the level of agreement between annotators.

The results are provided in Table 2.3. Unlike manual measurements, in this case, in addition to the type of discourse relation, certain other aspects were also measured. Regarding the type of discourse relation, there was 0.31 of correspondence which is 0.08 lower in comparison with the manual evaluation. Behind this difference, unlike with manual evaluations, is the fact that the tool takes into account the position of central subconstituents. Amongst other aspects that were evaluated, there was greater correspondence. In the case of connectors, there was 0.40 correspondence and in component and nuclearity aspects, correspondence was above 0.50, 0.52 and 0.66, respectively.

The corpus itself was tagged based on subjectivity. More precisely, certain semantic orientation components in texts and discourse structures were tagged.

Two annotators had to tag three rhetorical relation components: *i)* the rhetorical relation itself, *ii)* its nucleus or nuclei (in the case of CONTRAST's relation) and *iii)* its satellite. There were three tags to be assigned: *i)* positive semantic orientation, *ii)* negative semantic orientation and *iii)* neutral semantic orientation. Overall, the two annotators annotated the rhetoric relation of one annotator and 40% of its components.

To achieve this, certain guiding principles agreed upon. These describe the tasks of each annotator, including the most difficult, such as metaphoric phrases and how they will be annotated. As can be seen from the difference in tagging results between two different annotators, the Cohen kappa coefficient was 0.58. Therefore, there is *moderate agreement*. As shown by Table 2.4, the largest non-correspondence between two annotators related to neutral semantic orientation.

		A2			
		NEG	**NEU**	**POS**	**Total**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Total	92	120	181	393

Table 2.4 – Contingency table of two annotators with respect to the semantic orientation of rhetorical relations.

Instead of considering agreement between two annotators in terms of correct semantic orientation, the annotators assigned a sentiment value for

each rhetorical relation and EDU.

2.1.2 Creation and evaluation of the sentiment lexicon in Basque

In sentiment lexicon creation in Basque, we first studied the existing conditions. We can distinguish three aspects:

- Time. Starting with sentiment lexicon, we aimed to work on certain features of sentiment analysis. This means that our time needed to be divided between certain parts of sentiment lexicon creation and sentiment analysis. Consequently, we had limited time to create the Basque lexicon.

- Tools and resources. Another limitation we faced for lexicon creation was related to the features available with Basque tools and resources. Certain sentiment lexicons were created using different approximations (Chen and Skiena 2014, Cruz et al. 2014, Barnes et al., 2018, Saralegi et al. 2013 and San Vicente and Saralegi, 2016) but the way sentiment was assigned to words is not. Sometimes, the value assigned to a word is not on a scale and, therefore, we cannot measure the intensity of change that can occur in a word due to valence shifters. Other times, however, in the lexicon, values for word sentiment are on two scales (positive and negative), and due to the polysemy of these words, finally, there are cases when the scale itself is not appropriate. This being the case, we decided to translate the Basque sentiment lexicon.

- Quality. We aim to create a high-quality sentiment lexicon. Moreover, we wanted to create an up-to-date lexicon which can be improved in the future. Consequently, we wanted to create a sentiment lexicon with features similar to those of the the SO-CAL tool.

In the next stage, we decided to create the sentiment lexicon. Following the SO-CAL tool's specific lexicon method for Spanish, we decided to translate it. We also decided to use SO-CAL's English lexicon in the translation process. We saw the following advantages when taking our decision:

- Features of SO-CAL lexicons. To take into account different linguistic phenomena, the values of lexicon words and values used to express

feeling sare between 5 and +5. In our opinion, both the values and value scales are appropriate for the aspects we want to study in our sentiment analysis, especially, since value changes can be measured on that scale.

- Translation resources. As previously mentioned, finding resources to create the Basque sentiment lexicon is difficult and those which do exist are not suitable. However, in Basque lexicography, there are many resources available, like the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* dictionaries (Sarasola, 2005), available both online.

- Choice to compare and evaluate. The sentiment lexicon we shall develop uses the same features as certain other sentiment lexicons and it can be used to compare the results. This also provides an opportunity to evaluate the lexicon.

Next, we collected resources and tools to develop the sentiment lexicon. Overall, we used five resources or tools:

- The Spanish version of the SO-CAL lexicon.

- Two bilingual dictionaries: *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005).

- The English version of the SO-CAL lexicon. The English version of the lexicon contains 6,610 words in five grammatical categories: nouns, adjectives, verbs, adverbs, and intensifiers. We looked at whether the entries of the first Basque version appeared in this dictionary and also what type of sentiment value they have. Then, we decided to assign the word a sentiment value or assign it the same value as the English version.

- The Basque Opinion Corpus.

- *Key Word In Context* (KWIC) technique. The Basque word translated from Spanish was used to search for the word in the corpus and identify the word's context. This way, we assigned a sentiment value attached to the field of the word translated from Spanish into Basque.

Instead of collecting resources and tools, we translated the sentiment lexicon and created them. There are three stages in the translation process: *i)* translate the lexicon from Spanish into Basque, *ii)* clean up the Basque lexicon and *iii)* evaluate the Basque lexicon. We will carry out the translation process based on the examples in Table 2.5.

Phenomenon	SPA	SPA in group	EUS	ENG	The last value
F1	desacreditar “discredit”	desacreditar −2 “discredit”	ospea_kendu −2 “discredit” izena_kendu −2 “discredit” sona_kendu −2 “discredit”	-	-
F2	atrofiar “atrophy”	atrofiar −1 “atrophy”	atrofiatu −1 “atrophy”	-	-
F3	amago ‘feint”	amago −1 ‘feint” cicatriz −2 “scar”	seinale −1 “signal”	-	-
F4	franquismo “Francoism”	franquismo −2 “Francoism”	frankismo −2 “Francoism”	-	−2
F5	correcto “correct”	acertado +3 “right” correcto +3 “correct” decente −2 “decent”	zuzen +3 “right”	right +1 correct +3	+3

Table 2.5 – Example of five phenomena related to the translation process

1- Translation. In this major stage, the SO-CAL tool’s Spanish lexicon was translated.

   i) Automatic translation from Spanish to Basque. First, we translated the Spanish lexicon automatically into Basque using the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries. If a Spanish word could be translated into Basque in several ways, we took into consideration all of the possibilities. This way, the value of each word in Spanish was inherited by each of the translation possibilities.

   In Table 2.5, the word in the first column was translated into Basque (third column).

   ii) Filtering and grouping. In this stage, we filtered and grouped the translations obtained in Basque, that is to say, when translating

the Spanish lexicon certain Basque translations appeared as repetitions and we grouped these. As a result of this grouping, we collected both the translation and the value of the original word.

In Table 2.5, in the third column, phenomenon F3, *amago* ("feint") and *cicatriz* ("scar") in Spanish can both be translated as *seinale* ("signal") in Basque. Likewise, for phenomenon F5, the original Spanish *acertado* ("correct"), *correcto* ("correct") and *decente* ("decent") are all translated as *zuzen* ("correct").

iii) Only choosing translated words which are *Elhuyar* (Zerbitzuak, 2013) or *Zehazki* (Sarasola, 2005) dictionary entries. We studied the translations one by one to see whether they were entries in the two dictionaries. When translating and creating the lexicon, we only took into consideration those translations which were dictionary entries.

In Table 2.5, phenomenon F1, *sona_kendu* ("discredit"), *izena_kendu* ("discredit") and *ospea_kendu* ("discredit") are not dictionary entries and therefore we did not take them into consideration. However, *atrofiatu* ("atrophy") is a dictionary entry and so was taken into account.

iv) Selecting sentiment value. After removing translations which are not dictionary entries, in the remaining translations, we chose the Basque translations which matched Spanish sentiment values and meanings, whenever the Spanish word could be translated into Basque in several different ways. When the Basque translation had a single meaning in Spanish, we chose that meaning and corresponding sentiment value.

When choosing the Spanish meaning and corresponding Basque translation, we followed the procedure below:

* If the translated Basque word had a single translation (and value) in the Spanish original, we use the corresponding translation and value. This is what happened in F2 and F4.
* If the Basque translation corresponds to several words and values in Spanish, the sentiment value assigned to the word (and Spanish meaning) is based on the Basque Opinion Corpus. This was the case with F3 and F5.
* There were certain cases where the translation obtained did

not appear in the corpus but there were several and original words with sentiment valence in Spanish. In this situation, priority was given to the most widely or commonly used translation of the word.

2- Cleaning up. In this second stage, we cleaned up the initial draft version of the sentiment lexicon. More precisely, we adapted the lexicon to certain specific fields and using the SO-CAL's English lexicon we enriched and corrected it.

   v) Adapting the field and corpus. Intending to create the second version, we used the frequency of lexicon words in the Basque Opinion Corpus.
   In Table 2.5, phenomenon F2, *atrofiatu* ("atrophy") is a concept in health and since that field is not present in our corpus, we removed the concept from the lexicon.

   vi) Re-examining and improving each of the lexicon's entries. We translated each of the lexicon's words from Basque into English using the *Elhuyar* dictionary (Zerbitzuak, 2013), then we looked whether these words appeared in SO-CAL's English version of the lexicon. If the Basque entry appeared in the English lexicon, we gave the Basque entry the same value as the English entry. This means we prioritize the value of SO-CAL's English version over the Spanish version. However, if the Basque entry does not appear in the English version, we either removed or maintained this entry according to how appropriate the word was.
   In Table 2.5, phenomenon F5, the Basque word *zuzen* ("correct") corresponds in SO-CAL's English lexicon to *correct* and *right* with different sentiment values.

3- Evaluating the lexicon. In the final phase, the Basque sentiment lexicon was evaluated. For this, firstly, based on two annotators' annotations a gold standard was created and the results obtained by the gold standard and lexicon were compared.

   vii) Annotation of the corpus' most frequently used words' gold standard. Using *Analhitza* (Otegi et al., 2017), the Basque Opinion Corpus' 400 most frequent words were removed (100 words for each grammatical category) and two annotators as-

signed a sentiment value between 5 and +5. Next, based on both annotations, the gold standard was created.

viii) In the second version of the lexicon created, the sentiment lexicon was assigned to the 400 words used in the gold standard.

ix) The values assigned by the gold standard and lexicon were compared using the Pearson correlation. Using this correlation, the correspondence was measured in two ways:

- Pearson 1. In this measurement, in the list of 400 words, only words tagged by two annotators were taken into account. If a word is only tagged by one person it was not taken into consideration.
- Pearson 2. In this measurement, 400 words from the list were used. If a word is not tagged by one or two annotators, it was assigned the value 0, so that the word can be taken into account.

2.2 Identification of valence shifters

After developing the basic guidelines and tools for carrying out the research, we began to identify contextual valence shifters in Basque. In a step-by-step study, we examined different types of valence shifters on different language levels and their impact on words with opinion.

2.2.1 Phonology and morphology level: expressive palatalization and affixes

Firstly, we started to identify the valence shifters of phonology and morphology-level. With this aim, in the first step, we searched for several bibliographic sources that work on the morphology and phonology of the Basque language and we collected a list of language affixes. In this list, we included an example of the grammatical category of the affixes and their semantic meaning. Among the bibliographic sources we used to complete Table 2.6, we can mention (Urdangarin, 1982), (Euskara Institutua, EHU, 2011) and (Oñederra 1990).

Mopheme	Noun	Adjective	Verb	Semantic classification	Example
-zale	X			Liking (of)	Ardozale "liking of wine"
ez-	X	X	X	Negation	Ezberdin ("different")
[z] → [x]				Closeness/ smallness	Gazte ("young") → gaxte ("young") Zahar ("old") → xahar ("old")

Table 2.6 – Morphemes with their characteristics.

Table 2.6 shows examples of expressive palatalization and affixes collected from bibliographic sources. The first letter is *-zale* ("fan of"), it appears with names and it means liking. The second is a prefix *ez-* ("no"); it may appear with nouns, adjectives or verbs and denotes negation.

Finally, in the last example there is an expressive palatalization: [z] becomes in [x]. Like all expressive palatalizations, there is no grammatical restriction and it is used to indicate closeness or smallness. We followed the same procedure with all other affixes.

Word	Number of instances
zati ("part")	11
zertxobait ("a little")	11
zerua ("sky")	9

Table 2.7 – Examples of words extracted based on number of instances.

In the next step, we took 192 opinion texts from the Basque Opinion Corpus and we analyzed them using the *Analhitza* tool (Otegi et al., 2017). After that, we took all the words from the text listed by frequency. Table 2.7 shows the word list provided by the *Analhitza* tool (Otegi et al., 2017), and is based on frequency. As can be seen, the word *zertxobait* ("slightly") that contains the suffix *-txo* ("little") appears in the corpus eleven times.

We then looked at what affixes (prefixes and suffixes) and what expressive palatalizations of the Basque language in the corpus. To do this, we used the list we created in the first step and, based on that, we collected a list of all how the affixes and expressive palatalizations appeared in the corpus. In the list, we individually described their characteristics as shown in Table 2.8.

In Table 2.8, in Example (1), there are no affixes or expressive palatalizations. In Example (2), the suffix *-txo* ("small") is in the word and it weakens the sentiment valence of the word "attack" (-2). Finally, in Example (3), because of the expressive palatalization, [t] becomes in [tt], and as this change

	Word	Morpheme	Valence	Effect on valence	Semantic	Noun	Adj.	Verb
(1)	istorio ("story")							
(2)	erasotxo ("a little attack")	-txo	−2	Weaken	Smallness	X	X	
(3)	pattal ("ill")	[t] → [tt]	−2	Strengthen	Proximity			

Table 2.8 – A sample of morphemes and expressive palatalization in the corpus.

has reinforced the expressivity, the sentiment valence (−2) becomes even stronger and negative. In total, we found for 59 suffixes, 7 prefixes and 13 expressive palatalization instances in the Basque Opinion Corpus, and with all of them, we followed the same procedure.

2.2.2 Syntactic level: negation marks

Another aspect of language which can influence the value of words is syntax. However, in this case, changing the value will not affect a word, but rather a sentence or phrase which consists of a group of words. Just as morphology or phonology valence shifters influence a word, syntactic valence shifters may influence, in addition to words, phrases or sentences.

With the aim of studying how negation marks change the sentiment value of phrases, firstly, we collected a corpus with sentences which have negative marks. However, before collecting the corpus, we created a list of Basque negation marks based on Altuna et al. (2017) and Altuna et al. (1985). These are the negation marks we used for our research: *ez* ("no"), *ezin* ("to be unable to"), *gabe* ("without"), *ezik* ("except"), *salbu* ("except"), *ezta* ("not only") and *ezean* ("if not").

In the next stage, in 96 corpus texts, we collected sentences with at least one negation mark. In this part of our research, we used 96 out of 192 texts which were not part of the test. We did not use all the texts because we obtained most negation marks which appeared in (Altuna et al., 2017) and because despite analyzing more texts there were no new negation marks. Overall, we obtained 359 instances of negation marks. According to data provided by *Analhitza* (Otegi et al., 2017), these 359 negation marks were distributed amongst 320 sentences. Consequently, there are sentences with more than one negation mark. The sub-corpus of negation marks we obtained 5,515 words.

After that, we assigned a sentiment value to both words and full sentences

in these 320 cases. To achieve this, we added words to the *Eustagger* tool (Alegria et al., 2002) used to assign grammatical categories and/or lexical marks from the *Sentitegi* sentiment lexicon (Alkorta et al., 2018); in the lexicon, the words are distributed according to grammatical category.

(5) Pogostkinak ezin hobeki$_{+2}$ atera zituen. (MUS20)
Pogostkina pulled them off perfectly$_{+2}$.

(6) Irabazi$_{+2}$ ezinik jarraitzen du Eibarrek. (KIR17)
Eibar continues without winning$_{+2}$.

(7) Ikuspuntu politikotik$_{-1}$ ez ezik, ekonomikotik$_{+3}$ ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)
Not only from a political$_{-1}$ point of view but also economically$_{+3}$, Greece has brought hope to other southern European countries, including the Basque Country.

In Examples (5), (6) and (7), the tagger and the sentiment lexicon added to it assigned a value to four words: *hobeki* "better" (adjective), *irabazi* "win" (verb), *politiko* "politic" (adjective) and *ekonomiko* "economic" (adjective). The first two words were assigned value +2 and the adjectives 1 and 3, respectively. This way, with the negation marks and words with a sentiment value, we prepared the context for a linguistic analysis at the next stage.

When carrying out the linguistic analysis of the negation marks, what we took into consideration was the shift in orientation caused by the negation marks in the semantics and, more precisely, in the sentiment value. In other words, we wanted to study the consequences of a negation mark in the case of words or groups of words with values, its range and, consequently, in phrase values too. The analysis was done manually and subsequently, we followed the described procedure.

In the analysis, we looked at the corpus' sentences manually one by one. After identifying the sentences' negation marks, we stated their range. Finally, taking into consideration the consequence on both the field of scope's sentiment value and the full sentence, we grouped the sentence itself and its negation marks.

(8) Pogostkinak [ezin hobeki$_{+2}$] atera zituen. (MUS20)
Pogostkina pulled them off [perfectly$_{+2}$].

(9) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)
Eibar continues [without winning$_{+2}$].

(10) [Ikuspuntu politikotik$_{-1}$ ez ezik], [ekonomikotik$_{+3}$] ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira. (POL08)
.[Not only from a political$_{-1}$ point of view], but also [economically$_{+3}$], Greece has brought hope to other southern European countries, including the Basque Country.

In Example (8), the negation mark is *ezin* ("can not") and its range *hobeki* ("better") goes as far as the adjective. Since in the range of action there is only one word and it has a value, the range of action value is +2, as well as for the sentence. In this case, analyzing the influence of the negation mark, we realized that the negation mark *hobeki* ("better") consolidates the word with value, in fact, in intensity *ezin hobeki* ("perfectly") is stronger than *hobeki* ("better"). Therefore, we ranked the sentence itself and the negation mark's value in the group which they consolidate.

However, in Example (9), although it uses the same negation mark (*ezin* "can not"), we see its influence is different, for example, the range of the negation mark *ezin* ("can not") is *irabazi* ("win"), with value +2 and this weakens it. In other words, *irabazi ezinik* ("without winning") is weaker than *irabazi* ("win") in terms of sentiment value. Therefore, the negation mark and the sentence itself is part of a larger group, because it has been weakened. As can be seen with these two examples, it is possible for the same negation mark itself to be in two different functions.

Finally, in Example (10), the negation mark *ez ezik* ("not only"), with a sentiment value of −1, from a "political point of view" negates the word group, but from the perspective of sentiment analysis, no change occurs in the sentiment value of the group of words. In fact, in this case, after the negation, it is used to add information and not to negate the negation's range. Therefore, the sentiment value of words that are part of the structure is not influenced by the structure *ez ezik* ("not only") in its range.

Number	**Negation mark**	**Structure of the rule**
1	ezin	PM *ezin* + [adjective/adverb] (+ comparative suffix) PM
5	ezin	PM [(NP) + verb + *ezin*] PM PM [*ezin* (+ auxiliar verb) (+ NP) + verb (+ NP)] PM
10	ez	PM [NP] *ez ezik* PM

Table 2.9 – Proposed rules for identifying negation marks that have different effects on sentiment.

Although we wanted to add information about negation in tools like SO-CAL which are based on rules, it is necessary to identify negation-marks and their range. We, therefore, created and evaluated identification rules. We created and evaluated them in the *Constraint Grammar* approach (Karlsson et al., 2011).

Nevertheless, to identify negation marks and their range before creating rules, in the stage before this, we organized in a structured manner the information collected in the linguistic analysis, as can be seen in Table 2.9. When completing this table, we used (Altuna et al., 1985) as a reference, since it describes the position of negation marks in sentences.

One of the items collected in the fact that the structure for each rule is syntactic. The structure of rules in Table 2.9 has the following features: square brackets [] show the negation mark's range, brackets () show that the phrase is optional. Italics show the negation mark's or lexicalized structure and forward slash / indicates that different grammatical categories can appear in the scope of negation.

Other items that appear in the rules include groups of words and grammatical categories - NP and VP show noun- and verb-phrase, respectively - and adjectives, adverbs, verbs, auxiliary verbs, and phrases.

Finally, there is another item that appears in the rules: restricting punctuation marks (PM). This restriction in rules aims to place the range of the negation mark within the sentence and not use items from other phrases. The range can appear both before and behind the negation mark, and since we saw the risk of treating the range as a word we decided to add restricting punctuation marks to the rules. In the case of lexicalized structures, there is no such restriction. Since lexicalized structures have a stable structure (since they are groups of words which are repeated) this is not necessary.

In Table 2.9, after describing what type of features who has, we shall explain how those rules are created.

(11) (...) [ezin da baztertu$_{-1}$ ekaitz zaparradaren bat izatea.]
.[It can not be ruled out$_{-1}$ a storm shower]. (EGU35)

(12) [Irabazi$_{+2}$ ezinik] jarraitzen du Eibarrek. (KIR17)
Eibar continues [without winning$_{+2}$]

Through Examples (11) and (12), we shall look at Table 2.9's fifth rule's two variants. In Example (11), after *ezin* ("can not") being a negation mark, the auxiliary verb *da* ("is") and main verb *baztertu* ("rule out") appear, and

finally, a subject which is a noun phrase *ekaitz zaparradaren bat izatea* ("being a storm shower"). Therefore, the structure obtained for this phrase would be: *ezin* "can not" + auxiliary verb + main verb + NP. All the phrase appears in brackets and this shows the range of the negation mark. If we study Example (12), we will see the main verb *irabazi* ("win") appears before the negation mark *ezin* ("can not") and this will be the negation mark's range. After following the same procedure with 36 other instances of the negation mark *ezin* ("can not") in the corpus, we obtained the rules which appear in Examples (13) and (14).

(13) PM [(NP) + verb + *ezin*] PM

(14) PM [*ezin* (+ auxiliar verb) (+ NP) + verb (+ NP)] PM

As Examples (13) and (14) show, the main verb always appears before or after the negation mark *ezin*. In the case of Example (13), the appearance of the noun phrase is optional as well as in Example (14), where the same thing happens. The noun phrase can appear before or after the main verb (for example, *ezin partidua irabazi "can not win the match"* or *ezin irabazi partidua* "can not win the match"). The auxiliary verb can also appear after the negation mark.

Furthermore, to finish, we added the restriction of punctuation marks to the rule, so as not to use the part of the phrase before or after the negation mark. How the *Constraint Grammar* (Karlsson et al., 2011) adaptation rule was used to identify the negation mark, its range and lexicalized structure can be seen in Figure 2.1 below.

```
LIST PUNTUAZIOA = PUNT_PUNT PUNT_KOMA PUNT_BI_PUNT PUNT_GALD PUNT_ESKL
    PUNT_HIRU PUNT_PUNT_KOMA;

LIST EZ = "ez";
LIST BESTERIK = "beste";

# (2)
# besterik ez egiturak
MAP (!besterikezHAS) TARGET (DET) IF (0C BESTERIK)   (1C EZ);

(...)
```

Figure 2.1 – Examples of rules for identifying negation marks and their range in *Constraint Grammar* context (CG3).

Figure 2.1 shows how we adapted to the context of *Constraint Grammar* (Karlsson et al., 2011) :

- LIST. Here certain words or other items, for instance, punctuation mark, are listed. In Figure 2.1, punctuation signs (LIST PUNTUAZIOA) and negation marks (LIST EZ, LIST BESTERIK) are listed.
- MAP command (reflection rules). Using these rules, the grammar inspects the specified structures in the corpus and the structures in the corpus are tagged. The rules consist of the following:
  - Tag assigned. In Figure 2.1, the tag has the ! sign at the beginning and it is assigned to the target word.
  - The target word. In Figure 2.1, the target of the rule is a determinant (DET) (*beste*, "other"). In order to tag the target word, this target word must fulfill the rule's condition.
  - The rule's conditions. In order to assign the tag *!besterikezHAS* to a determinant, which is in position 0, the negation mark *ez* ("not") must be after this determinant (in position 1).

When carrying out the evaluation, we used 48 texts from the Basque Opinion Corpus. These 48 texts had to be processed to adapt them to the context of the *Constraint Grammar* (Karlsson et al., 2011) and to do this we used the Basque morpho-syntactic disambiguator *Constraint Grammar*.

```
"<,>"<PUNT_KOMA>"
    PUNT_KOMA
"<batere>"
     "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG \%SINT
"<harrotu>"   S:278/0
     "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG \%
         ADIKAT S:278 !gabeAUR
"<gabe>"   S:141/0
     "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> \%SINT S:141 !gabe
"<\$.>"<PUNT\_PUNT>"
    PUNT\_PUNT
```

Figure 2.2 – The corpus tagged with *Constraint Grammar*'s rules (Karlsson et al., 2011) .

Each rule in the *Constraint Grammar* (Karlsson et al., 2011) leaves its tag on in every word that fulfills the conditions. In Figure 2.2, for example, the

rules we created in the *Constraint Grammar* (Karlsson et al., 2011) tagged two words (*!gabeAUR* and *!gabe*).

After applying the rules to 144 corpus test sections, these rules were evaluated. Firstly, the correspondence between two annotators to carry out the evaluation was measured using a small part of the corpus and, then, one person did the entire evaluation.

When evaluating, the annotators had to follow the procedure below:

1- The annotator used three tags when evaluating: ETIK_ONDO, when the rule's tag is correct; ETIK_FALTA, when the word in the corpus needed a tag and when the rule does not assign any and, finally, ETIK_GAIZKI, when the rule assigns a tag to the corpus' word, but the word does not need a tag.

2- In the corpus, the annotator studied each word one by one and as the evaluation progressed gave the words a tag. If the word was tagged correctly, the annotator assigned ETIK_ONDO (correct tag). However, if the tag was missing, the annotator assigned the ETIK_FALTA (missing tag) status. Finally, if the word had an incorrect tag (false positive), then it was marked as ETIK_GAIZKI (incorrect tag).

```
"<,>"<PUNT_KOMA>"
    PUNT_KOMA
ETIK\_FALTA "<batere>"
    "batere" ADB ARR ZERO w39,L-A-ADB-ARR-13,lsfi48 @ADLG \%SINT
ETIK\_ONDO "<harrotu>"   S:278/0
    "harrotu" ADI SIN PART NOTDEK w40,L-A-ADI-SIN-38,lsfi49 @-JADNAG \%
        ADIKAT S:278 !gabeAUR
ETIK\_ONDO "<gabe>"   S:141/0
    "gabe" ADB ARR ZERO w41,L-A-ADB-ARR-14,lsfi50 @KM> \%SINT S:141 !gabe
"<\$.>"<PUNT\_PUNT>"
    PUNT\_PUNT
```

Figure 2.3 – The tags assigned by annotators to words.

Figure 2.3 presents an example of an evaluation. There are three words and all three should be tagged. Two out of three (*harrotu* "become arrogant" and *gabe* "without") are tagged as ETIK_ONDO, at the beginning of lines. However, this is not the case for the word *batere* ("not"). Although it is part of the range of the negation mark, because it influences the verb *harrotu* ("become arrogant"), it is not tagged.

Consequently, it is tagged as ETIK_FALTA, in this case too, at the beginning of the line. The aforementioned procedure was followed for agreement between two annotators as well as to evaluate the corpus test section.

3- Finally, to measure agreement, F1 score was used. As the results show, the score was 0.60.

2.2.3 Discourse level: rhetorical relations, their components, and central unit

At the discourse level, following the *RST* approach (Mann and Thompson, 1988), we focus on two aspects. On the one hand, we analyzed changes in sentiment valence in rhetorical relations. On the other hand, in the central units, that is, in the most important units of discourse in the text, we studied the relation between grammatical categories and opinion.

Rhetorical relations

In rhetorical relations, we have first defined our aims. In total, we set three goals:

- We want to measure the agreement of semantic orientation between the rhetorical relation and its components.
- Given the distance between rhetorical relations from the central unit, we want to measure the greatest agreement of the semantic orientation between rhetorical relations and the whole text.
- In opinion texts, we want to examine whether different types of rhetorical relations they appear in the same parts of texts. We will establish the position of rhetorical relations based on the distance to the central unit.

After setting our goals in rhetorical relations, we have taken a few steps. The following is a step-by-step methodology:

1- Assign semantic orientation to different components of the discourse level:

   – 240 opinion texts.

– The following rhetorical relations: EVALUATION/INTERPRETATION (140 instances), CAUSE subgroup (71), CONCESSION/ANTITHESIS (70), EVIDENCE/JUSTIFICATION (53), CONTRAST (26), CONDITION subgroup (18), ENABLEMENT/MOTIVATION (6). In total, 384 instances were assigned the semantic orientation.

– The components of the above rhetorical relations: semantic orientation has been assigned to the nucleus and the satellites as well as to the first and last EDU of the relations.

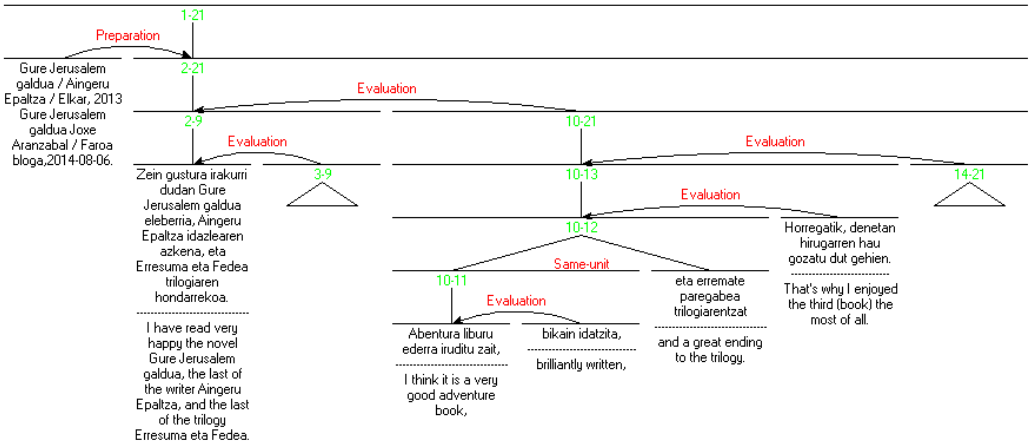


Figure 2.4 – Discourse-tree of the text (SENTFAR-01.)

2- Parameter analysis of rhetorical relations. After introducing the rhetorical relation of 28 literature texts in the database, we analyzed the following parameters. That is, in the second step of the methodology, in the manual labeling, we performed the parameter analysis on the 384 instances of the aforementioned rhetorical relations. We will use the EVALUATION relation (the EDUs 10-13) of Figure 2.4 to explain the parameters:

– Nuclearity. A rhetorical relation can be mononuclear or multinuclear. In Figure 2.4, the EVALUATION relation is N(ucleus)-S(atellite).

– Semantic orientation of rhetorical relations and EDUs. We have assigned three types of semantic orientation to rhetorical relations

and EDUs: positive, negative and neutral. First, we assign the semantic orientation to the EDUs and, then, to the whole rhetorical relation. In the EVALUATION of Figure 2.4 (the EDUs 10-12 and 13), the two EDUs and the whole relation have a positive semantic orientation.

- Distance to the central unit. The distance between the rhetorical relation and the central unit has been measured by the number of rhetorical relations between them. In our case, the distance of the EVALUATION relation (the EDUs 10-13) is +2.
- The type of rhetorical relation. The last parameter that we have taken into account is the type of rhetorical relation. The relation we take as an example is EVALUATION type. Consequently, the EDUs 10-12 presents a situation and the EDU 13 makes an evaluative comment about the situation.

3- In terms of semantic orientation, we measured the agreement among the semantic orientation labels. Agreement measurement has been made on rhetorical relations and their constituents. Manual evaluation was performed using F-measure.

Central unit

On the other hand, to find out the most common grammatical categories of the words and to analyze the distribution of words with semantic orientation in the central unit, we have performed the following steps:

1- Extract central units from the corpus. First, a linguist has selected central units from 192 texts of Basque Review Corpus. The test part of the corpus was not used. Before selecting the central units, the texts were segmented following the guidelines of the *RST* approach (Das and Taboada, 2018). On the other hand, we used the *RSTTool* tool (O'Donnell, 2000) for text segmentation.

In Figure 2.5, the text EGU01 can be seen in XML format. Each of these segments is an elementary discourse unit (EDU) and the task of the annotator, in connection with the next step, is to identify which of these EDUs is the central unit, in other words, the most important EDU in opinion texts.

```
<rst>
  <header>
    <relations>
    </relations>
  </header>
  <body>
    <segment id="1"> Gaurko eguraldia. HEGO EKIALDEKO HAIZEA ETA HODEI
        BATZUREKIN EPELDU EGINGO DA. </segment>
    <segment id="2"> Giro ona tokatuko zaigu gaur ere hego haize epelarekin
        .</segment>
    <segment id="3"> Borraska pixka bat hurbiltzeak haizea apur bat mugituko
         du</segment>
    <segment id="4"> baina </segment>
    <segment id="5">tenperatura igoaz, </segment>
    <segment id="6">giroa goxo mantenduko da leku gehienetan.</segment>
    <segment id="7"> Hodei zirrinta mehe batzuk agertuko dira zeru sabaian</
        segment>
    <segment id="8"> baina eguzkia apur bat lausotu arren,</segment>
    <segment id="9"> astro handia bistan edukiko dugu</segment>
    <segment id="10"> eta dotoreziak ez du bat ere galduko.</segment>
    <segment id="11"> Baditeke iluntze aldera ekaitz hodei batzuk garatzea
        eta euri zarrastaren batzuk botatzea agian. </segment>
    <segment id="12">Akats txiki horiek gora behera eguraldi bikaina oro har
        .</segment>
  </body>
</rst>
```

Figure 2.5 – Selection of central unit by two annotators in the EGU01 opinion text.

2- Inter-annotator agreement in central unit selection. To prove that the annotator has chosen the central unit correctly, another person has also selected the central units of these texts. To measure the agreement between two annotators, a total of 78 opinion texts (32.5% of the total corpus) were used. Percentage calculation was used to calculate the inter-annotator agreement.

Agreement	51	0.65
Disagreement	27	0.35
Total	78	1.0

Table 2.10 – Inter-annotator agreement when choosing a central unit in opinion texts.

As shown in Table 2.10, in 51 opinion texts of the 78 texts, two annotators considered the same EDU as a central unit. In the other

44 opinion texts, however, different EDUs have been chosen as central units. Therefore, the agreement in selecting the central unit of 78 opinion text was 0.65.

3- Analysis of the grammatical category of words in central units. In the next step, we analyzed the type of words in each central unit and whether one or more words appear at regular frequency. With this aim, we used the *Analhitza* tool (Otegi et al., 2017). We have analyzed these central units domain-by-domain.

4- Analysis of the results provided by *Analhitza* (Otegi et al., 2017). After that, we analyzed the results provided by the *Analhitza* tool (Otegi et al., 2017). In analyzing the results, we considered the following aspects:

   i) Frequency of words. We analyzed the 30 most frequent words of each grammatical category.

   ii) Grammatical category of words. Another feature of the words considered was their grammatical category. The tool directly classifies words into grammatical categories; which has made our work easier.

   iii) Domain of words. We also examined whether these words with high frequency are domain-related. In all domains, the classification was whether or not words belong to the domain (binary classification), except in weather. In weather, we also used the time concept in the domain classification, since we believe that the time is of particular importance. Therefore, in the weather domain, the word classification was trivial.

   For example, in words with the highest frequency in the weather domain, we classified the word “winter” as belonging to the domain and “appearance” as not belonging to the weather domain; because it is not directly related to this domain. Finally, since the word "weekend" indicates time, we classified it into a weather domain.

   iv) Assignation of sentiment valence to central units and their words. Another feature examined was the sentiment valence of the central units. We have used the Basque version of the SO-CAL tool with *Sentitegi* lexicon (Alkorta et al., 2018), which we created for the assignment of sentiment valence.

2.3 Lexicon-based document level sentiment classifier in Basque

The third major step of this work was to develop a document level sentiment classifier in Basque. With this aim, we wanted to develop the first Basque version of the SO-CAL tool (Taboada et al., 2011). Therefore, we used the *Eustagger* tool (Alegria et al., 2002) and the *SentiTegi* sentiment lexicon (Alkorta et al., 2018).

2.3.1 Integration of the *Eustagger* tool

The English version of SO-CAL contains lemmatized words with sentiment valence and some rules related to inflection. This is not valid for Basque, due to its rich inflection system and agglutinative nature.

To solve this difficulty, in the Basque version of SO-CAL, we have decided to integrate the *Eustagger* tool (Alegria et al., 2002) to lemmatize words of texts before assigning sentiment valence to them. In this way, the tool will first lemmatize the text and, then, check if the lemmatized word is present in the lexicon of the tool and if it is, it will assign a sentiment valence to the word in the text. Therefore, if a lemmatizer is integrated into the tool, the tool will first be able to lemmatize the text; then to check if the word lemmatized is in the lexicon, and finally, if the word is in the lexicon, to assign the sentiment valence to the word.

Figure 2.6 shows the changes made and a comparison of the structures of the SO-CAL tool in English and Basque. As can be seen, in English, there are some morphological rules (such as the deletion of the plural *-s* letter) and in this way, the words in the texts are transformed into lemmas. In Basque, meanwhile, because of morphological richness, the *Eustagger* lemmatizer (Alegria et al., 2002) is integrated to achieve the lemmatization. Then, the sentiment lexicon is applied and if the lemmatized words are in the lexicon, the tool assigns a sentiment valence to them. Finally, other rules are similar for English and Basque (assigning more weight to words with negative semantic orientation, not assigning sentiment valence in interrogative sentences, etc.) and they are useful for both languages.

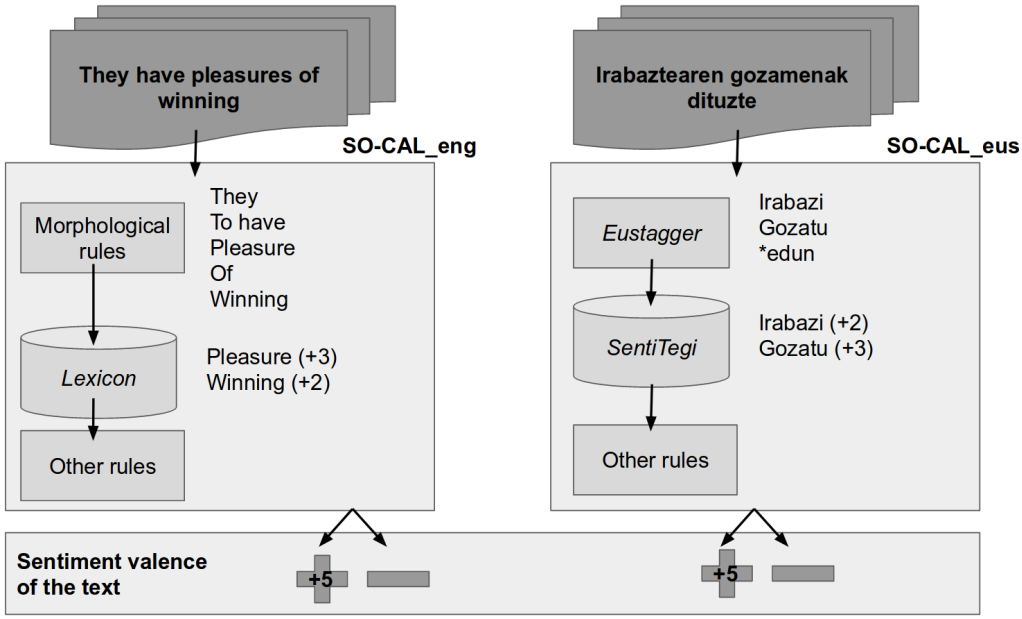


Figure 2.6 – Comparison of the structures of the English and Basque versions of the SO-CAL tool.

2.3.2 Integration of the *Sentitegi* sentiment lexicon

In the Basque version of the SO-CAL tool, the *Sentitegi* sentiment lexicon (Alkorta et al., 2018) has been integrated .

The tool itself contains a module for sentiment lexicons where we added Basque ones. For the tool to work, lexicons must be in *txt* format, with an entry and its valence in each line. Also, each grammatical category must have its file. The *Eustagger* lemmatizer (Alegria et al., 2002) will lemmatize the word in the text and identify its grammatical category, and, then, the tool will search that word in its grammatical category. We have added lexicons for four grammatical categories: noun, adjective, adverb, and verb.

Table 2.11 shows the change made. On the left, there is an English lexicon with a list of words and their sentiment valence. On the right, there is a part of the lexicon in Basque.

English		**Basque**	
Thriving	+3	Bikain	+5
Record-setting	+3	Maximo	+1
Leading	+3	Orokor	+3
Industrious	+3	Min	-2
Best-selling	+3	Polit	+4
Upset	+3	Gogor	-1
Clean	+2	Ahul	-2
Capacious	+2	Behar	-1
Cogent	+2	Txar	-3
Confident	+2	Zail	-2

Table 2.11 – Examples of sentiment lexicons in English and Basque.

2.3.3 Sentiment classifier evaluation

After integrating the *Eustagger* lemmatizer (Alegria et al., 2002) and the *Sentitegi* sentiment lexicon (Alkorta et al., 2018) in the sentiment classifier, we evaluated the classifier itself.

We have used 48 review texts from the Basque Opinion Corpus. As mentioned in this methodology section, the semantic orientation of the corpus opinion texts is annotated and we compared them with the results provided by the sentiment classifier. The tool gives numerical results and when evaluating, we considered the sign in the numerical result.

2.4 Summary

In this section, we describe the methodology of this thesis work. First, we created resources and tools for sentiment analysis. On the one hand, we built a Basque Opinion Corpus, collecting opinion texts from newspapers and specialized websites. Then we developed *SentiTegi* (Alkorta et al., 2018), a sentiment lexicon in Basque. To do this, we have translated the lexicon of the Spanish version of the SO-CAL tool using the *Elhuyar* (Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) bilingual dictionaries.

In the second step, we identified the valence shifters and measured their effect on words and phrases. We worked on phonological and morphological valence shifters. We collected information about them from bibliographic

sources and, then, we extracted their instances from a part of the corpus. In the next step, we assigned a sentiment valence to these instances with valence shifters using the *SentiTegi* lexicon (Alkorta et al., 2018). After that, we measured the effect of these instances in words. In other words, we examined if they reinforce, weaken, or do not affect the valence of words and phrases.

We also similarly dealt with negation marks. We collected a list of negation marks from bibliographic sources and searched them in a part of the corpus. Then, we assigned sentiment valence to the sentences with a negation mark and measured the effect of the negation marks. In the next step, using the above analysis, we developed rules for identifying negation marks and their scope and then adapted them to the *Constraint Grammar* (Karlsson et al., 2011) to evaluate them.

At the discourse-level, we obtained different kinds of rhetorical relations from a part of a corpus annotated with the *RST* approach (Mann and Thompson, 1988). Then, using both SO-CAL and *SentiTegi* sentiment lexicon (Alkorta et al., 2018), we assigned sentiment valence to the components of rhetorical relations and rhetorical relations. We also indicated some other features of rhetorical relations. With this information, we studied rhetorical relations from a sentiment analysis perspective. On the other hand, we also worked on the central unit by analyzing the grammatical categories of words that appear there and studying the distribution of words with sentiment valence.

In the next step, we adapted the English SO-CAL tool for sentiment classification into Basque. To this end, we first integrated the *Eustagger* tool (Alegria et al., 2002) for text lemmatization in the SO-CAL tool. Then, we replaced the English sentiment lexicon with the Basque *SentiTegi* lexicon (Alkorta et al., 2018).

RESULTS

3

Resources and tools

3.1 Article 1

Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis

***Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis*,**

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, IXA Group, University of the Basque Country

jon.alkorta@ehu.eus
koldo.gojenola@ehu.eus
mikel.iruskieta@ehu.eus.

Nowadays, it is very usual to read reviews about movies, products or tourist destinations before taking a decision. Reviews, as a particular genre, follow some genre constraints and also a specific discourse structure.

Following Taboada et al. (2016) discourse structure, along with syntax, is necessary to get a better account of sentiment analysis in review corpora. The aim of this paper is to present a corpus we have developed in order to study sentiment analysis in Basque. As far as we know, there is no polarity-balanced corpora for the study of sentiment analysis in Basque.

Corpus design

In order to fulfill this gap, we built a corpus for that purpose following this criteria:

1) Collect texts from specialized review websites (online magazines and newspapers).
   1.a) With a clear negative or positive evaluation.
   1.b) With rich syntactic structures and *opinionative* data.
   1.c) With balanced domains: 20 positive and 20 negative texts with similar word length.
2) Describe corpus information with: code, title, source, polarity and word length.
3) Analyze the corpus using different methods to measure the opinionative phenomena and evaluate its quality. Reliability has been measured comparing some characteristics of our corpus against other corpora.

Corpus

The corpus that we have built is composed of 240 texts in Basque corresponding to 6 domains (books, music, movies, weather, politics and sports). It contains 52,092 tokens and 3,711 sentences.

Regarding size, our corpus can be compared to other corpora built to analyze sentiment analysis: i) Emotiblog (Boldrini et al., 2010), that contains 100 texts for each language (Spanish, Italian and English; ii) the SFU Review Corpus (Taboada, 2008) is made up of 400 texts (8 domains), 289,270 tokens and iii) the Opinion Corpus for Arabic (OCA) (Rushdi-Saleh et al., 2011) has 500 texts and 215,948 tokens.

The quality of our corpus has been measured with respect to the following phenomena:

i) Presence of first person. Sentiments are usually expressed using the first person and, thus, we have measured if the frequency of the first person is different in objective and subjective corpora. A set of texts from Wikipedia has been taken as objective corpus (with the same topic and similar length as our corpus), because Wikipedia asks writers to include neutral or non-opinative texts. A language analyzer for Basque and English (ANALITZA), which is based on a set of tools for the automatic linguistic analysis based on IXA-pipes (Agerri et al., 2014), has been used to obtain the frequency of use of the first person.

With this tool we have measured the frequency of the first person of verbs (in Basque) and the frequency of pronouns (in English). Results demonstrate that the frequency of the first person is different in both corpora. While the presence of first person is about 0.12% (English Wikipedia) and 1.31% (Basque Wikipedia) in objective corpora, in subjective corpora the frequency increases up to 8.37% in our corpus and 11.80% in the SFU Review Corpus (Taboada, 2008). Results also show language differences: the first person is mostly plural in our corpus (5.10% plural and 3.27% singular) while the first person is mostly singular in the SFU Review Corpus (1.70% plural and 10.10% singular).

ii) Adjectives in the corpus. We have measured the frequency of adjectives, because adjectives are one of the most frequently used phenomena in sentiment analysis. However, we found that the frequency is similar (from 8% to 9%) in both languages and four corpora.

iii) Discourse markers. Relational discourse structure can change the polarity of a text, because there are coherence relations which describe the purpose or the conclusion of a text. Because of that, a text span with such relations is more important and, consequently, the polarity of the text span should be taken into account. In our corpus, we have seen some discourse marker signals, that signal *purpose* or *conclusion* discourse relations.

Some discourse markers that signal purpose are: *azken batean* 'in the end' (9), *laburbilduz* 'to sum up' (4), *azkenik* 'finally' (3), *amaitzeko* 'to finish' (3), etc. Moreover, the following discourse marker list signals the conclusion: *beraz* 'thus' (74), *ondorioz* 'consequently' (12), *hortaz* 'so' (4), etc.

The following example of our corpus shows the relevance of this phenomenon:

(1) *(...) aire masa* ***hotz*** *eta* ***ezegonkor*** *bat iritsiko zaigu (..) eguraldia benetan* ***gaiztoa*** *izango dugu. (...). Laburbilduz, etxean geratzeko moduko eguraldia*.

(...) **cold** and **unstable** air mass (...) very **bad** weather. (...). In short, the weather invites us to stay at home.

In Example (1), the first sentence states that the weather will be cold, bad and unstable and, after the underlined discourse marker, the prediction is summarized suggesting *to stay at home*.

Adversative discourse markers change the polarity of the previous text span. There are some adversative discourse markers in our corpus: *baina* 'but' (389), *ordea* 'however' (40), *hala ere* 'nevertheless' (38), *aldiz* 'while' (34), *berriz* 'whereas' (33), *dena den* 'even so' (16), *dena dela* 'anyway' (5), *haatik* 'though' (5), to cite some.

(2) *(...) Jada ezagutzen dugun istorioa, noski. Baina, horrek ez dio freskotasunik kendu (...).*

(...) We already know the story, of course. But this does not remove the freshness (...).

In Example (2), the first sentence may be said to have a negative polarity, but the discourse relation signaled with an adversative discourse marker in the second sentence inverts the polarity (from negative to positive).

iv) Irrealis Blocking. As Taboada et al. (2011) explains, there are some language forms that triggers an *irreal* context. We have found different examples in our corpus: a) conditionals and b) negative polarity items.

a) We found three types of conditionals in our corpus: i) non-hypothetical; ii) hypothetical and iii) unreal.

b) Besides, we have found various negative items for persons (*inor* 'nobody', 13 instances), things (*ezer* 'nothing', 29), mood (8), time (16) and space (5).

iv) Negation. Negation appears in different ways in our corpus: *ez* 'not' (718) modifying clauses; *gabe* 'without' (107) modifying noun phrases and *ezean* 'in the absence of / unless' (4) modifying subordinate clauses.

Conclusion and future work

In this work, we have created a Basque corpus for sentiment analysis and we have made a preliminary analysis of the data. The frequency of first person and discourse markers shows that the corpus is valid to study different opinionative phenomena. Moreover, we observe that the corpus has typical constructions (discourse marker, irrealis blocking and negation) analyzed in sentiment analysis which also suggest that our corpus contains opinionative data. In the future, we will tag this corpus using the frameworks of Rhetorical Structure Theory (RST, Mann & Thompson, 1988) and Appraisal Theory (Martin & White, 2005), to study how relational discourse structure modifies other language levels (semantic and syntactic) of sentiment analysis in Basque, following previous work (Alkorta et al., 2015).

References

Agerri, R., Bermudez, J., & Rigau, G. (2014, May). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC* (Vol. 2014, pp. 3823-3828).

Alkorta J., Gojenola K., Iruskieta M. & Perez A. (2015). Using relational discourse structure information in Basque sentiment analysis. 5th Workshop "RST and Discourse

Studies", in *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015), Alicante (España).*

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2010, July). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 1-10). Association for Computational Linguistics.

Mann, W. C., & 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.

Martin, J. & White, P. (2005) *The Language of Evaluation: Appraisal in English*. London: Palgrave McMillan.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, *62*(10), 2045-2054.

Taboada, M. (2008) The SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, *2*, 325-347.

3.2 Article 2

SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis

ISSN 2007-9737

SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta

IXA NLP Group, University of the Basque Country (UPV/EHU), Vizcaya, Spain

jon.alkorta@ehu.eus, koldo.gojenola@ehu.eus, mikel.iruskieta@ehu.eus

Abstract. The creation of a semantic oriented lexicon of positive and negative words is often the first step to analyze the sentiment of a corpus. Various methods can be employed to create a lexicon: supervised and unsupervised. Until now, methods employed to create Basque polarity lexicons were unsupervised. The aim of this paper is to present the construction and evaluation of the first semantic oriented supervised Basque lexicon ranging from $+5$ to -5. Due to the lack of resources, the Basque lexicon was created translating the SO-CAL Spanish dictionary by means of two bilingual dictionaries following specific criteria and then slightly corrected with the SO-CAL English dictionary and frequency data obtained from the Basque Opinion Corpus. Evaluation results show that the correlation between human annotators is slightly better than between a gold standard lexicon (obtained from human annotation) and the translated dictionary. This shows that the quality of the translated lexicon is satisfactory, although there is a space to improve it.

Keywords. Semantic oriented lexicon, manual translation method, Basque, sentiment analysis.

1 Introduction

Sentiment analysis is a task that classifies documents according to their polarity. This research area has had a big development in the last years due to social networks and Internet, which have increased the quantity of opinions and other types of text with emotion, and is in demand of methods for automatic processing.

There are many resources for sentiment analysis for the most used languages such as English [9], Chinese [15] and Spanish [5].

Additionally, competitions like SemEval [10] have greatly contributed to the development of resources and tools for sentiment analysis. However, the development is not symmetric on lesser used languages or languages in normalization process like Basque.

The semantic oriented lexicons are related to the lexical level and, so, they are useful and important in sentiment analysis. If the semantic orientation of the words is known, opportunities open up to calculate the semantic orientation of sentences and, therefore, the semantic orientation of texts taking into account syntax and discourse constraints.

The creation of the semantic oriented Basque lexicon has been semi-manual translating from the SO-CAL Spanish dictionary, and then enriching it with corpus analysis and the English SO-CAL dictionary. In the translation process, different bilingual dictionaries have been used. We have decided to use a semi-manual procedure to create our lexicon, in order to take into account some idiosyncratic characteristics of Basque language.

The aim of this paper is to present a semantic oriented lexicon for Basque. We will emphasize the process of creating this lexicon, and particularly the solutions adopted to solve the problems encountered.

The main contributions of this work are: i) the creation of a domain-specific semantic oriented Basque lexicon, ii) a description of a semi-manual technique to create the lexicon and iii) a thorough evaluation.

ISSN 2007-9737

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the methodology of the translation process. Then, Section 4 discusses the design decisions, while Section 5 describes the characteristics of the created lexicon in two stages. In Section 6 the quality of the lexicon is evaluated and, finally, Section 7 concludes the paper, also proposing directions for future work.

2 Related Work

There are various approaches for the creation of polarity lexicons, based on knowledge or on automatic methods. Each of the approaches has its advantages and drawbacks.

SO-CAL [14] is a dictionary-based tool to extract sentiment from texts. The dictionary was created manually, where words are annotated with polarity (positive or negative) and strength (semantic orientation: from ± 1 to ± 5). There are two versions of SO-CAL tool. The original version is the English SO-CAL and the Spanish version, the second one, is based on the previous version. The English and Spanish dictionaries (V1.11) contain 6,610 and 4,880 words, respectively.

A disadvantage of manually-created lexicons is the hard-work to make modifications. In contrast, they can be tailored to be domain-specific and, depending on the linguistic information used, they can treat a variety of different linguistic phenomena.

ML-SentiCon [6] is a multilingual polarity lexicon, where the lexicons have been automatically generated from an improved version of Senti-WordNet. It contains a Basque lexicon that contains 4,323 lemmas. The polarity values are situated between -1 and $+1$, in a continuous scale. Additionally, QWN-PPV tool [11] is able to generate multilingual polarity lexicons, including Basque. This unsupervised tool makes use of a corpus and WordNet.

The main disadvantage of these lexicons is that they are not domain-specific, so their results could vary from one domain to another. In contrast, their main advantage lies on the facility to create them.

Another characteristic of previous three works is that the sentiment value of words is in a scale, although the scale dimensions are different. However, there are works in which the sentiment value of words are not in scale. For example, in some works like [13], there are two non-numerical tags: *positive* and *negative*. Consequently, two words with different intensity are expressed with the same tag.

Methods to evaluate lexicons are different depending on each technique. Some works [3] use intrinsic methods where the result of the system is compared to a gold standard data set, predefined by evaluators. In contrast, there are other systems [4] which use extrinsic methods where the system is evaluated in an applied setting. Finally, some works [7] use both extrinsic and intrinsic methods.

The lexicon presented in this work differs from previous ones in several respects. SO-CAL dictionaries have also been manually created but, until now, they have dealt with languages which are not morphologically rich (Spanish and English) in contrast with Basque. Another relevant difference of this study has been the evaluation. We will apply an intrinsic evaluation and measure, using Pearson correlation, the agreement between two human annotators, and the reliability between the gold standard (based on human annotation) and the translated dictionary. Finally, the characteristic of the created lexicon is another interesting aspect. The words of the lexicon have the sentiment value in a scale from -5 to $+5$. This allows us to study how sentiment shifters of different linguistic levels (morphology, syntax and discourse) affect on sentiment analysis.

3 Methodology

In order to create a semantic oriented lexicon for Basque, we have adopted several decisions taking different factors into account:

i) **Time.** The creation of a semantic oriented lexicon for Basque is related to the project of linguistics-based Basque sentiment analysis and, for that reason, the time to create the lexicon is limited.

ii) **Resources.** The Basque language is still in a normalization process and this has some limitations to create corpora and to reuse computational resources. On the one hand, it is difficult to create a large opinion corpus of different topics. This situation could affect to the quality of the lexicon if the corpus is used for that. The collaboration of lexicographers would be ideal but it is a costly resource, not available. This situation adds a difficulty to create a semantic oriented Basque lexicon from zero.

iii) **Quality.** We want to develop the lexicon with the best possible quality (and in the less time possible) and with that aim we will first translate the lexicon, after that evaluate it and then improve our semantic oriented lexicon following an specific criteria.

3.1 Resources for Translation

We have used mainly four resources in the translation process.

i) **The SO-CAL Spanish Dictionary [14].** This dictionary is the source to create the Basque semantic oriented lexicon. It contains 4,880 words of five grammatical categories (noun, adjective, adverb, verb and intensifier).

ii) **Two Bilingual Dictionaries:** Spanish-Basque: Elhuyar dictionary [16] and Zehazki [12]. These dictionaries have been used to translate the Spanish SO-CAL dictionary. Moreover, they have also been used to check if the translated word is an entry of such dictionaries since we will work only with words which are entries of one of these dictionaries. Dealing with collocations and expressions is necessary but it is out of the scope of this work.

iii) **The Basque Opinion Corpus [1].** After getting the first version of the lexicon, each entry has been checked in the corpus to create a domain-based lexicon. The corpus contains 240 texts of six different domains.

iv) **The SO-CAL English dictionary [14].** This version which contains 6,610 words has been used to verify and enrich the already created domain-based lexicon.

Taking all the factors explained above into account and using the mentioned resources, we have decided to translate the SO-CAL Spanish dictionary to create the Basque SO-lexicon *Sentitegi*, following the methodology explained in Figure 1.

3.2 Translation Steps

Figure 1 shows the steps followed in the translation process. To begin with, a first version of a semantic oriented Basque lexicon has been created from the Spanish version of the SO-CAL dictionary. After that, the second version has been created enriching it with the English lexicon version (V1.11) and limiting it to the domains of Basque Opinion Corpus.

Some interesting phenomena have been detected in the translation process of SO-CAL dictionaries from Spanish and English versions (V1.11) to Basque. Table 1 shows these five phenomena.

- Phenomenon 1 (P1): the Spanish word is translated but the translation is not an entry of Elhuyar [16] and Zehazki [12] dictionaries, so we do not take it into account.
- Phenomenon 2 (P2): The Spanish word is translated, it is an entry of Elhuyar but the translation does not appear in the Basque Opinion Corpus. Consequently, it will appear in the first version (V1.0) but not in the second one (V2.O).
- Phenomenon 3 (P3): The Spanish word is translated, it is an entry, it appears in the corpus but it is not in the SO-CAL English dictionary. So, it will appear in the first version of the dictionary, but not in the second one.
- Phenomenon 4 (P4): The Spanish word is translated, it is an entry, it appears in the corpus and it is not present in the SO-CAL English dictionary. Then, it will be included in the (first and) second version.

ISSN 2007-9737

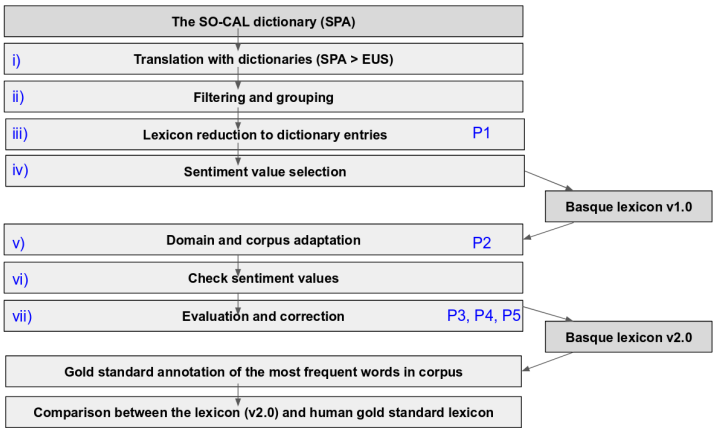


Fig. 1. Steps of the translation process. The enumeration in blue on the left indicates methodological steps. The blue code on the right (P1 to P5) indicates different phenomena in the translation process

– Phenomenon 5 (P5): The Spanish word is translated, it is an entry, it appears in the corpus and it also a word of the SO-CAL English dictionary. It will appear in the first and second versions. These last two phenomena are the same but the decision is different that depends on the characteristic of each word.

The translation process has been the following (see Figure 1):

i) **Automatic translation from Spanish into Basque.** The Spanish sentiment dictionary of SO-CAL has been translated using Elhuyar [16] and Zehazki [12] dictionaries. When one word of the dictionary has more than one entry, all the entries have been taken into account. The sentiment value of the Spanish word has been assigned to all the correlated elements in Basque.

   For example, the Spanish word *desacreditar* -2 "discredit" has been translated into Basque in different forms: *izena_kendu*, *ospea_kendu* and *sona_kendu* "discredit" with the same meaning. This example shows how one Spanish word could be translated in different forms to Basque. But these translations are not entries of the dictionary. Consequently, they have not been taken into account.

ii) **Filtering and grouping.** After translating all the words and transferring their sentiment values, the repeated words in Basque have been filtered and grouped.

   Table 1 shows how words in Basque (fourth column) can have one or more translations in Spanish (third column). The phenomena numbered 1, 2 and 4 have one translated word in Spanish whereas 3 and 5 have more than one.

   This phenomenon occurred because those words are polysemic. There are cases where two or more words in Spanish correspond to the same word in Basque and vice versa. Consequently, in some cases, each word in

Table 1. Words that belongs to five phenomena related to translation process

Phenomenon	SPA	SPA grouping	EUS	ENG	Value
P1	desacreditar “discredit”	desacreditar -2 “discredit”	ospea_kendu -2 “discredit”	-	-
P2	atrofiar “atrophy”	atrofiar -1 “atrophy”	atrofiatu -1 “atrophy”	-	-
P3	amago “feint”	amago “feint” -1 cicatriz “scar” -2	seinale “signal” -1	-	-
P4	franquismo “Francoism”	franquismo -2 “francoism”	frankismo -2 “francoism”	-	-2
P5	correcto “correct”	acertado “correct” +3 correcto “correct” +3 decente “decent” -2	zuzen +3 “correct”	right +1 correct +3	+3

Basque has several meanings and sentiment values in Spanish.

iii) **Dictionary entry: Check if the Basque translation is an entry in the Elhuyar [16] and Zehazki [12] dictionaries.** We have only accepted the translations which are entries of Elhuyar and Zehazki dictionaries. Consequently, Phenomenon 1 in Table 1 has occurred: *ospea_kendu* “discredit” is a collocation and not an entry, so we will not take it into account. In contrast, other words in the table are entries in the dictionary and they are maintained.

iv) **Sentiment value selection.** The value (and meaning in Spanish) of each word in Basque will be selected.

In order to choose the value, we have followed the following criteria:

- If the word in Basque has one translation (and value) in Spanish and if that translation is correct, the translation is selected. This is the case of phenomena 2 and 4 in Table 1. Sometimes the translation is not “correct” or “direct” as we will observe in Section 4.
- If the word in Basque has many translations (and values) in Spanish, the translation has been selected according to which translation is the best to use in the Basque Opinion Corpus [1]. We have analyzed the context of the words in the corpus using Key Word In Context (KWIC) format for concordance. This is the case of Phenomena 3 and 5 in Table 1.
- In the creation of the first version of the lexicon, there have also been cases where the word in Basque has not instances in the corpus. In these cases, the meanings that are used more frequently have been selected.

After these four steps, the first version of the Basque lexicon (V1.0) has been created. However, we detected some inconsistencies and we have felt the necessity to feed more information and, for that reason, we followed new steps to create the Basque lexicon (V2.0):

v) **Domain and corpus adaptation: New lexicon based on the Basque Opinion Corpus [1].** We have curated the first lexicon (Basque V1.0) and created the second version of this lexicon (Basque V2.0). This new lexicon has been curated with the information obtained form word frequencies we have extracted from the Basque Opinion Corpus.

The effects of this step are showed in Phenomenon 2 in Table 1. The word *atrofiatu* “to atrophy” does not appear in the corpus,

so it is not related to the domains of the corpus and, consequently, we do not take it into account. We do not take into account them because our work is limited to our corpus and we want to maintain as much as possible the coherence of SO values and avoid complexities which we do not see useful. In Table 1, Phenomena 3, 4 and 5 are not affected by this limitation while Phenomenon 2 is. With this procedure, the number of entries in the lexicon was reduced from 8,140 to 1,813 words, because it was manually checked and reviewed.

vi) **Curate and check SO values of each entry**: Find the English translations of each Basque entry in the SO-CAL English dictionary. Using the Elhuyar dictionary [16], we have translated the words in Basque to English and, after that, we have checked if the translated words are in the SO-CAL English dictionary. If the word is in this dictionary, we have maintain the dictionary entry and its value in the second version of the Basque dictionary. If the word is not in the English dictionary, almost in all cases. it was excluded from the second version in the Basque dictionary.

In Table 1, Phenomena 3 and 4 do not have any translation in the English dictionary and, consequently, their (English) column in Table 1 is empty. In contrast, Phenomenon 5 has two translations according to the English dictionary: *right* and *correct*.

vii) **Evaluation and correction**: Compare and choose the best translation and value. In this step, each word in Basque has the same value, most of the times, in Spanish and English (Basque V1.0).

There are 3 different cases in this situation:

- Phenomenon 3. There is not a word in the English version corresponding to the Basque word and the previous Spanish one is not accepted. In phenomenon 3, the word *seinale* "sign" has been assigned the value -1 (Table 1, fourth column) but there is not a corresponding value in the English version and, consequently, we have removed that value.
- Phenomenon 4. There is not a corresponding word in the English version for Basque and the previous Spanish translation and value are accepted. The word *frankismo* "francoism" is related to Spain and, for that reason, it appears in the Spanish version and not in English. In this case, we have maintained the assigned value.
- Phenomenon 5. The English translation and value are the same or better quality than the Spanish ones. Phenomenon 5 shows that the Spanish and English values agree, so we have assigned the value $+3$ to *zuzen* "correct". In other cases, the English and Spanish values differ. When this happens we decided that the English value will prevail to the Spanish one in the second version of the Basque dictionary, because the quality is slightly better in English as we previously report.

Phenomena 3 and 5 show how we have decided to give more relevance to the English version.[1]

4 Discussion

We explain in this section how we have solve the most fundamental problems we have found during the translation process:

i) **Source language is not always the preferred language.** English and Spanish could be the source language but we have chosen Spanish due to several reasons. The overall accuracy of the English SO-CAL is 76.62% while in the Spanish version is 71.81% [2]. In other words, the difference between them is not big enough. On the other hand, there are many more resources to translate

[1] Sometimes there is not a corresponding word in the English dictionary [16], an example and the explanation of what we have done in such cases is explained in Section 4.

Table 2. Examples of translations applying the coherence criteria

Criteria	EUS	Value	EUS	Value
A	errukigabe "ruthless"	-4	errukigabeko "(with) ruthless"	-4
B	tonto "stupid"	-3	tuntun "stupid"	-3
C	arduradun "responsible"	$+2$	arduragabe "irresponsible"	-2

the dictionary from Spanish to Basque than to translate from English to Basque. So, the translation from Spanish is more reliable and extended as shown in Table 1, where the phenomenon numbered 4 (*frankismo* "francoism") shows that although the English dictionary contains more items, there are some words in the Spanish dictionary that are not present in the English one.

In contrast, the English version has helped to check if the assigned value to the Basque word in the first version from Spanish is correct. In the cases where the value of the Spanish and English versions are different, we have preferred the English one as Phenomenon 3 (*seinale* "signal") shows. Due to this decision, the number of words of the lexicon has decreased from 1,813 to 1,237 entries.

ii) **Not one to one translation**. Another problem was presented when, in the translation, a Spanish word could be translated into Basque in different forms but with the same sense. We have decided to use all the translated words in Basque so as to get the higher recall possible. The first step, the automatic translation from Spanish into Basque, shows that one or more entries have been taken in Basque.

For example, the Spanish word *aparatoso* "showy, spectacular" has been translated into Basque in two different ways: *arranditsu* "spectacular" and *deigarri* "showy".

iii) **Domain adoptation of polysemic words**. There are some words that have opposite meanings according to their context. The best solution would be to create two entries but then it would be difficult to implement it in a system that does not distinguish between word senses. In this situation, we have decided to take only one meaning and we have used the Basque Opinion Corpus [1] to choose the meaning with the appropiate SO value.

For example, the Basque word *deigarri* "showy, spectacular" comes from Spanish *aparatoso* -3 "spectacular" or *llamativo* $+3$ "showy". Taking the context of the word in the corpus into account, we have disambiguated the word manually and chosen the value $+3$ for this word.

iv) **Coherence consistency.** In the process of choosing the value, we have to try (when the values match) to maintain the coherence of the values taking these criteria into account. Examples of the criteria are shown in Table 2.

A) Sometimes, the same word appears in different forms. For example, in the creation of the first version of the lexicon, it is usual that one word appears sometimes with genitive *-ko* "with" and other times with an elided genitive, and in both cases is a dictionary entry. In these cases, we decided to assign the same value. One of the cases is the adjective *berehala* "immediate". It appears with genitive suffix: *berehalako* "immediately" and without it *berehala* "immediate". We have assigned the same sentiment value $(+2)$ to both.

B) We assign (when the values match) the same value to words with similar meanings. For example, *tonto* "stupid" is used with man while *tuntun* with the same meaning is used with woman. We assign the value -3 to both.

Table 3. The semantic oriented Basque lexicons (V1.0 and V2.0)

	V1.0		V2.0	
Grammatical category	**Words**	**%**	**Words**	**%**
Noun	2,282	28.06	461	37.27
Adjectives	3,162	38.85	446	36.05
Adverbs	652	7.98	54	4.36
Verbs	1,657	20.36	276	22.32
Intensifiers	387	4.75		
Total	**8,140**	100	**1,237**	100

C) We also assign the same intensity (1 to 5), but opposite value (positive/negative) to antonymic words when the values coincide in Basque dictionary entry. In Basque, some prefixes (*des-* and *ez-* "dis-") and suffixes (*-ezin* "impossibility" "inability" and *-gabe* "without") are used to invert the meaning of the words and we have put special attention on these ones.

v) **"Incorrect" translations.** There have been some translations which are incorrect because of different factors. The Spanish word *provinciano* "backward" (-1) is employed to refer to people of Bizkaia and Gipuzkoa provinces. The Elhuyar dictionary [16] has defined the word as "inhabitant of Bizkaia or Gipuzkoa", a translation which is not useful for our purpose.

vi) **"Indirect" translations.** There have been some translations that we have considered as indirect. They are correct translations but since they have an extensive meaning and they are used in limited situations, they are not useful for us.

For example, the word *beltz* "black" could have two meanings: i) a color ii) "black, sad; gloomy, depressing" (figurative meaning). The figurative use of that word is less usual, there are other words with the same meaning and, taking into account that the word could complicate the correct sentiment value assignation of texts, we have decided not to assign any SO value.

The explained problems show the difficulty to translate a semantic oriented lexicon semi-automatically. This translation process is large and very detailed where the translation of the lexicon has different phenomena.

5 Results

As a result of the translation process, two versions of the semantic oriented Basque lexicon have been created. Table 3 shows the characteristics of these two versions.

The first version (V1.0) is the result of the first four steps in the translation process (Figure 1). It is translated directly from the Spanish SO-CAL dictionary with a strict criteria. But, unlike the second version (V2.0), the first version is not subject to the restrictions of being an entry of the Basque bilingual dictionaries and it was not improved taking into account the English SO-CAL dictionary, the Basque Opinion Corpus and other kind of features that work differently such intensifiers are considered as dictionary entries.

As a result of these considerations, the first versions have 8,140 entries and the second version 1,237, respectively. In both cases, nouns and adjectives are the grammatical categories with more entries. Verbs and adverbs are least frequent entries, whereas intensifiers have not been taken into account in the second version because they affect to other words, so we think that it is better to analyze differently assigning different values that does not go from -5 to -5 values.

Table 4. Examples of parallel lexicon

Word in lexicon	Value	SPA	Value	ENG	Value
bikain	$+5$	excepcional	$+5$	excellent	$+5$
on	$+2$	buen	$+2$	-	-
eskas	-1	escaso	-2	insufficient	-1
txar	-3	adverso	-3	bad	-3

Another interesting characteristic of the created lexicon is that it is parallel. That means that each word of the lexicon has it translations in English and Spanish and the sentiment values in each language also are included. This information appears in an orderly manner in the resource.

In Table 4, there are four examples showing the parallel lexicon. Sometimes, four sentiment values do not match because the Spanish and English SO-CAL lexicons have been created in different way. But the Basque word always matches with one of them. The examples of Table 4 are adjectives and they show how the sentiment values are in a scale.

Once we have implemented this lexicon in the Basque SO-CAL preliminary version, the created semantic oriented lexicon is useful to assign sentiment value to words as well as sentences, as is shown in the following examples:

(1) [*Halere, pentsa litekeenaren aurka, gaien urritasunak eta diskurtso errepikakorrak*$_{-6}$ *ez dakarte ñabardura aberastasunik, are gutxiago argumentu-mailako sakontasunik.*]$_{-6}$
(However, contrary to what is thought, the scarcity of problems and the repetitive$_{-6}$ discourses do not imply rich nuances, much less a plot depth.)$_{-6}$

(2) [*Arazo nagusia*$_{+2}$*, nire ustez, gaien*$_{+4}$ *emankortasun zalantzazkoan eta ekintzaren bilakaera eskasean*$_{-3}$ *datza.*]$_{+3}$
(The main$_{+2}$ problem is, I believe, the uncertain fertility of the topics$+4$ and the slow$_{-3}$ evolution of the action.)$_{+3}$

(3) *(...) [Emaitza ezustekorik*$_{-1.5}$ *gabeko istorio bat da, irakurlea epel*$_{-1.5}$ *uzteko arrisku dezente duen tonu arras moderatu batean emana.*]$_{-3}$
(The result is an unsurprising$_{-1.5}$ story, given in a moderate tone with a risk to leave the reader cold$_{-1.5}$.)$_{-3}$

As we show in the three examples the words of the dictionary have a SO value at the end of the word. To mention one, in Example 1, the Basque version of SO-CAL tool assigns the value -6 to the word *errepikakor* "repetitive". There is no another word with sentiment value according to lexicon, so the sentiment value of the sentence is also -6.[2] The methodology to calculate the semantic orientation of the sentence is similar in Examples 2 and 3.

6 Evaluation

In this section, we want to evaluate two aspects of the translation task. On the one hand, we want to evaluate the difficulty of the task. We think that the annotation of sentiment polarity is a difficult task because there is not a guide to follow and subjective perceptions must be, first, measured and, last, corrected if possible. On the one hand, the inter-annotator agreement of SO value annotation has been evaluated between two linguists annotators. On the other hand, we also want to measure the quality of the translated lexicon. With these in mind, a gold standard annotation has been created from the previous annotation and discussion by both annotators.

[2] In this sentence, the sentiment value of the word *errepikakor* "repetitive" in the lexicon is -4. But in SO-CAL tool, there are some mathematical operations related to linguistic phenomena that increase or decrease the sentiment value of the words. In this case, the sentiment value has increased to -6.

Table 5. Pearson correlation measurement and contingency table between two annotators

Grammatical category	Pearson 1	Pearson 2
Noun	0.87	0.59
Adjectives	0.71	0.60
Adverbs	0.93	0.82
Verbs	0.87	0.76
Total	**0.79**	**0.73**

Total categories			
	0	**NEG**	**POS**
0	187	12	27
NEG	14	42	5
POS	39	5	69

In order to evaluate these two aspects, we have extracted the most frequent 400 words (100 per each grammatical category) using Analhitza [8] from the Basque Opinion Corpus [1]. We have used Pearson correlation [17] to evaluate both tasks. Pearson correlation has been used in two different ways: i) Pearson 1: the correlation is measured taking into account only the annotated words by both annotators and ii) Pearson 2: the correlation is measured taking into account all words in the corpus.[3]

6.1 Correlation between annotators

We have decided to measure the correlation of two annotators to create the gold standard, taking into account the results achieved in the correlation coefficient. Table 5 shows the coefficient for each grammatical category, together with a contingency table.

Pearson 1 value shows that the correlation coefficient is high (0.79). This means that the value assigned is similar in a big percentage of the annotated words. The coefficients for different grammatical categories are situated between 0.71 and 0.93. In a similar way, Pearson 2 also shows high correlation, although it is slightly lower (0.73), with values between 0.59 and 0.82.

The contingency table of Table 5 shows that the biggest difference comes when one annotator has assigned a value to one word and the other one had not assigned any value and vice versa (90.19 % of all discrepancies 92 of 102).

After calculating this correlation, two annotators have discussed about their differences and after reaching consensus, a gold standard has been created.

[3] This means that there are cases where one word has been annotated by one annotator or by none of the them. When it happens, the un-annotated words value is 0 in order to calculate the Pearson correlation.

6.2 Correlation between the lexicon and gold standard

The correlation between the human gold standard lexicon and the translated lexicon shows some differences compared to the correlation between two annotators as presented in Table 6.

With Pearson 1, the cases in which the dictionary and gold standard contain an annotation for the word show similar correlation when compared to the results of two annotators (0.79). The correlation is high since the coefficients for the different grammatical categories are situated between 0.69 and 0.96. In contrast, Pearson 2 shows a lower correlation (0.54) and the coefficients of grammatical categories are situated between 0.47 and 0.59.

The interpretation of these results is that the values assigned to the dictionary and gold standard are similar (Pearson 1). But the difference from the previous result in Pearson 2 is created when the semantic oriented lexicon assigns value to the word and the annotator does no do it. This situation does not occur in the correlation between two annotators.

The contingency table shows us how the gold standard and the created dictionary differ. The discrepancy here also comes from the difficulty to assign a positive or negative value to a word. The difference is similar: 89.83 % of all discrepancies (106 of 118) are related to the decision to assign sentiment polarity to words. But here, in contrast with correlation between two annotators, the last version of the lexicon is more conservative, because the gold standard

Table 6. Pearson correlation measurement and contingency table between the gold standard and the Basque semantic oriented lexicon (V2.0)

Grammatical category	Pearson 1	Pearson 2
Noun	0.96	0.59
Adjectives	0.78	0.56
Adverbs	0.75	0.47
Verbs	0.69	0.54
Total	**0.76**	**0.54**

Total categories			
	0	**NEG**	**POS**
0	195	2	15
NEG	30	34	8
POS	59	4	53

annotates much more words than the lexicon does, decreasing the correlation in Pearson 2.

To sum up, the evaluation shows a high correlation in Pearson 1 in the case of two annotators and the lexicon and gold standard. The correlation coefficient is 0.79 and 0.76, respectively. In the case of Pearson 2, the correlation between two annotators remains high (0.73) but the correlation measure falls between the lexicon and gold standard (0.54).

7 Conclusion and Avenues for Future Work

In this paper we presented the first semi-manually created semantic orientation lexicon for Basque[4]. Time factor, few resources and quality pushed us to translate the SO-CAL Spanish dictionary to Basque.

The translation process has followed several steps. To summarize the steps, the English and Spanish SO-CAL dictionaries have been translated into Basque using two bilingual dictionaries. After that, the groups of words with the same meaning have been grouped and the best sentiment values according to the context of the Basque Opinion Corpus have been chosen. Finally, the created lexicon has been adapted to the domains of the Basque Opinion Corpus. The Basque sentiment lexicon has its limitations, since polysemy and figurative meaning phenomena were not considered and therefore are not totally solved.

Pearson correlation shows that the agreement coefficient is high between both annotators with respect to the following two factors: $i)$ assigning a value and $ii)$ deciding if a word has any value. In contrast, in the case of the comparison between human gold standard and translated lexicon, the correlation coefficient is high when the value is assigned but not in the case of deciding if the word has a value or not, which results has been lower. This lower coefficient appears mainly because there are less words annotated in our translated lexicon V2.0.

At present, the second version of semantic oriented lexicon is implemented in the Basque SO-CAL. In a foreseeable future, our aim is to improve this lexicon but considering morphosyntactic and discourse phenomena. This lexicon will be the basis of this system and we will consider how to enrich the system with sentence level and text level information.

[4] The semantic oriented Basque lexicon is available at: `http://ixa.si.ehu.es/node/11438`

Acknowledgements

Jon Alkorta's work is financed by a Basque Government scholarship (code: PRE_2017_2_0041). Koldo Gojenola's work is partially financed by the PROSA-MED (TIN2016-77820-C3-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). Mikel Iruskieta's work is partially financed by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish Ministerio de Economia, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). We would like to offer our special thanks to Maite Taboada, Arantxa Otegi and Oier Lopez de Lacalle.

References

1. **Alkorta, J., Gojenola, K., & Iruskieta, M. (2016).** Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. *Proceedings of Fourth International Workshop on Discourse Analysis (IWoDA16)*, pp. 58–62.
2. **Brooke, J., Tofiloski, M., & Taboada, M. (2009).** Cross-linguistic sentiment analysis: From english to spanish. *Proceedings of the international conference RANLP-2009*, pp. 50–54.
3. **Chetviorkin, I. & Loukachevitch, N. (2012).** Extraction of russian sentiment lexicon for product meta-domain. *Proceedings of COLING 2012*, pp. 593–610.
4. **Chetviorkin, I. & Loukachevitch, N. (2014).** Two-step model for sentiment lexicon extraction from twitter streams. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 67–72.
5. **Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008).** Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. *Procesamiento del lenguaje natural*, Vol. 41, No. 0.
6. **Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014).** Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, Vol. 53, pp. 113–120.
7. **Goyal, A. & Daumé III, H. (2011).** Generating semantic orientation lexicon using large data and thesaurus. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, pp. 37–43.
8. **Otegi, A., Imaz, O., de Ilarraza, A. D., Iruskieta, M., & Uria, L. (2017).** Analhitza: a tool to extract linguistic information from large corpora in humanities research. *Procesamiento del Lenguaje Natural*, , No. 58, pp. 77–84.
9. **Pak, A. & Paroubek, P. (2010).** Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, volume 10, pp. 1320–1326.
10. **Rosenthal, S., Farra, N., & Nakov, P. (2017).** Semeval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518.
11. **San Vicente, I., Agerri, R., & Rigau, G. (2014).** Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 88–97.
12. **Sarasola, I. (2005).** *Zehazki: gaztelania-euskara hiztegia*. Alberdania.
13. **Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966).** The general inquirer: A computer approach to content analysis.
14. **Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011).** Lexicon-based methods for sentiment analysis. *Computational linguistics*, Vol. 37, No. 2, pp. 267–307.
15. **Tan, S. & Zhang, J. (2008).** An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, Vol. 34, No. 4, pp. 2622–2629.
16. **Zerbitzuak, E. H. (2013).** Elhuyar hiztegia: euskara-gaztelania, castellanovasco. usurbil: Elhuyar.
17. **Zou, K. H., Tuncali, K., & Silverman, S. G. (2003).** Correlation and simple linear regression. *Radiology*, Vol. 227, No. 3, pp. 617–628.

Article received on 18/12/2017; accepted on 15/02/2018.
Corresponding author is Jon Alkorta.

4

Contextual valence shifters

4.1 Article 3

Saying no but meaning yes: negation and sentiment analysis in Basque

Saying no but meaning yes: negation and sentiment analysis in Basque

Jon Alkorta
Computer Languages and Systems
IXA group (UPV/EHU)

Koldo Gojenola
Computer Languages and Systems
IXA group (UPV/EHU)

Mikel Iruskieta
Didactics of Language and Literature Department
IXA group (UPV/EHU)

{jon.alkorta,koldo.gojenola,mikel.iruskieta}@ehu.eus

Abstract

In this work, we have analyzed the effects of negation on the semantic orientation in Basque. The analysis shows that negation markers can strengthen, weaken or have no effect on sentiment orientation of a word or a group of words. Using the Constraint Grammar formalism, we have designed and evaluated a set of linguistic rules to formalize these three phenomena. The results show that two phenomena, strengthening and no change, have been identified accurately and the third one, weakening, with acceptable results.

1 Introduction

Negation is a morphosyntactic operation in which a lexical item denies or inverts the meaning of another lexical item or language construction (Loos et al., 2004). The effect of the negation can be the change of semantic orientation (SO) and, according to Liu (2012), negation is called sentiment shifters because they change the semantic orientation of a word or a sentence.

With the aim of calculating the semantic orientation, the first step is to build a lexicon, but this is not enough, to grasp the correct SO-value of Example 1.

(1) [*Irabazi*$_{+2}$ *ezinik jarraitzen du Eibarrek*]$^{+2}$. (KIR17)
[(The soccer team) Eibar continues without winning$_{+2}$]$^{+2}$.

Following the semantic lexicon Sentitegi (Alkorta et al., 2018)[1], the semantic orientation of the word *irabazi* ("to win") is $+2$, and consequently, of the sentence also is $+2$. But we can notice that the semantic orientation of the sentence is clearly negative. The negator *ezin* ("can not") turns the positive oriented word *irabazi*$_{+2}$ ("to win") into a negative oriented one. Therefore, we think that addressing this phenomenon is crucial to obtain better results in the calculation of the SO of texts.

[1]The semantic lexicon is available on the web at: http://ixa.si.ehu.es/node/11438

The main aim of this work is to study how negation expressions and syntactic structures can change the semantic orientation of words, and to design a set of linguistic rules by means of Constraint Grammar (Karlsson et al., 2011) in order to identify these phenomena. According to our corpus study, different negation language forms can strengthen, weaken or have no effect on semantic orientation. These results go in the same direction as (Jiménez-Zafra et al., 2018b) where effects of negation within its scope are studied. We have centered our study on negation markers that unlike negation in verbs and nouns and negative polarity items, they only share information about negativity while others can share more information like aspect of action (e.g. *they denied going to the city*).

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes methodological steps. Then, Section 4 presents theoretical framework, while Section 5 gives a linguistic analysis. Section 6 shows results and error analysis, concluding with Section 7 and proposing directions for future work.

2 Related Work

There is a variety of works about negation and sentiment analysis in different languages and from different approaches.

For English, Liu and Seneff (2009) have presented a work where a parse-and-paraphrase paradigm is used to assign sentiment polarity for product reviews. If negation is detected, its polarity will be reversed (switch negation). If it has a value of $+5$, it will be reversed to -5, and vice versa. Following this, they have improved results (recall was improved in 45 %). The treatment of negation has been different in Taboada et al. (2011). In their work, when a negator is identified, the polarity value is not reversed; instead it is shifted toward the opposite polarity by a fixed amount. This approach is called shift negation. In

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 85–90
Brussels, Belgium, October 31, 2018.
https://doi.org/10.18653/v1/P17

Text	Text span	Dictionary words	SO value
MUS20	*Pogostkinak [ezin hobeki]*$^{+}$ *atera zituen* Pogostkina took them out [in an unbeatable way]$^{+}$	*hobeki* best, better	$+2$
KIR17	[*Irabazi ezinik*]$^{-}$ *jarraitzen du Eibarrek*, Eibar continues [without winning]$^{-}$	*Irabazi* to win	$+2$

Table 1: Polarity extraction of words (step 2) and linguistic analysis (step 3).

the creation of the semantic orientation calculator (SO-CAL tool), Taboada et al. (2011) have also treated negation in combination with other linguistic phenomena (like irrealis or intensifiers).

In Spanish, there are several works related to negation and sentiment analysis. In the case of Jiménez Zafra et al. (2015), firstly, they have analyzed what the effects of different negators in different sentences are. After that, they have created linguistic rules defined by the previous analysis. Finally, they developed a module that has been included in their polarity classifier system, improving results between 2.25 % and 3.02 % depending on the resource. Vilares et al. (2015) have used a syntactic approach for opinion mining on Spanish reviews. This system treats negation taking into account the scope and polarity flip caused by negation. According to their results, there is an improvement, due to the implementation of negation, among other reasons.

Our work is related to (Taboada et al., 2011) and (Jiménez Zafra et al., 2015) since it is based on a linguistic analysis and also because a set of rules that detect the negation language forms are created. As far as we know, there is not any work which analyzes negation in connection with sentiment analysis in Basque.[2]

3 Methodological steps

1- Negation corpus. We have extracted 359 negation instances of seven[3] negation markers. They were extracted from a total of 96 reviews of six different topics: movies, music, literature, politics, sports and forecast. We have selected those negation markers because they are the most frequent in the corpus.

2- Polarity extraction of every instance. We have created a polarity tagger, based on a POS tagger (Ezeiza et al., 1998) to enrich the corpus with POS information on a semantic oriented lexicon for Basque (Alkorta et al., 2018), to assign the semantic orientation value (SO value, between -5 and $+5$) to words, as shown in Table 1. There, the adverb *hobeki* ("best", "better") and the verb *irabazi* ("to win"), have a SO value of $+2$ in the lexicon.

3- Linguistic analysis. We have analyzed whether the negation markers can change the semantic orientation and the SO value of sentences. We have also tried to identify whether there are other phenomena related to negation with or without effects on semantic orientation. In Table 1, in MUS20, the negation marker appears near *hobeki* ("best"), an adverb. The result of this combination is strengthening. In contrast, in KIR17, the verb *irabazi* ("to win") is before the negator and the result is weakening. These two examples show the different performances of *ezin(ik)* ("can not"). Consequently, in Table 1, for example, this negation marker appears in two different groups. The same methodology has been used with other negation markers.

4- Constraint Grammar (CG3) rules for negation. Several rules have been proposed to detect each group, in order to identify the effects of negation based on the linguistic analysis presented in Section 5.

5- Evaluation. We use F_1 to evaluate the results using a different set of 46 reviews from the same corpus (Alkorta et al., 2016)[4].

4 Theoretical framework

In this section, we explain the three most important concepts, regarding our analysis: i) scope (negation analysis) and ii) switch and iii) shift negation (sentiment analysis approach to negation).

(2) *Berez pianorako konposatutako poliptiko txiki honek* **ez** *du bere naf kutsua galtzen*$_{-2}$ *bertsio orkestratuan.* (MUS01)
This small polyptych composed for the piano does **not** lose$_{-2}$ its naive sense in the orchestral version.

[2] Altuna et al. (2017) also analyze negation but their point of view is different, since they analyze events in Basque texts.

[3] The extracted negation markers from the Basque Opinion Corpus (Alkorta et al., 2016) are the following: *ez* ("not"), *gabe* ("without"), *ezin* ("can not"), *salbu* ("except (for)"), *izan ezik* ("except (for)"), *ezta* ("not even, not either") and *ezean* ("in the absence of" or "unless").

[4] A part of the corpus is available on the web at: http://ixa2.si.ehu.es/diskurtsoa/fitxategiak.php

(3) *−maitasun istorio konbentzional bat, grazia*$_{+3}$ *handirik*$_{+1}$ ***gabe****a−*. (LIB07)
*−*a conventional love story, **without** great$_{+1}$ grace$_{+3}$.[5]

According to Huddleston and Pullum (2002), the scope of negation is the part of the meaning that is affected by the negation marker, changing or not their SO value. In the examples above, the scope is underlined. As our study shows, there can be two kinds of semantic orientation in scope and these can be changed by negation markers. In Example 2, the SO value of the verb *galdu* ("to lose") and of its scope is -2. The negation weakens the SO value of the verb, reversing its SO. But, in Example 3, the SO values of the noun *grazia* ("grace") $+3$ and the adjective *handi* ("great") $+1$ assign a SO value of $+4$ to the scope which is positive. The negator *gabe* ("without") weakens the SO value.

According to Taboada et al. (2011), there are two approaches in sentiment analysis to weaken the negative SO value: i) switch negation and ii) shift negation.

(4) This pub is [**not** good$_{+3}$]$^{-3(switch)}_{-1(shift)}$ but the music from there is good$_{+3}$.

In the switch negation approach, the SO value of Example 4 is reversed. The SO value of the adjective *good* is $+3$ while the reversed SO value is -3. However, this criteria has a problem: if *excellent* is $+5$; *not excellent* would be more positive ($+1$) than *not good* (-2), but the SO value points to the contrary (*not excellent* is more negative than *not good*).

Otherwise, in the shift negation, the different negators have their own SO value and the results depend on the interaction of both SO values (the value of negation marker and negated word). Taking into account Example 4, the SO value of the negation *no* is -4 in the dictionary; so, when it modifies the word *good*, which has a SO value of $+3$, the sum value of scope is -1. This is the way how the shift approach solves the problem we describe in Example 4. We have decided to use the shift negation approach assigning a ± 4 SO value to the negators.

5 Linguistic analysis

In the theoretical framework of the shift negation, it has been considered that negation markers only weakens the SO value. Nevertheless, we have identified two other functions of these negation markers with low frequency, but relevant anyway from our point of view as the works of (Jiménez-Zafra et al., 2018a) and (Jiménez-Zafra et al., 2018b) show. As we observed in this study, the negation markers can strengthen, weaken or have no effect in the SO value of its scope as Figure 1 shows.

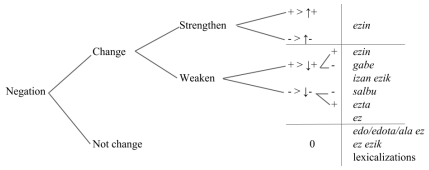


Figure 1: The effects of negation on semantic orientation according to negation markers.

The majority of negation markers usually weaken the semantic orientation of scope. But as we can see in Figure 1, the negation marker *ezin* ("can not"), for example, can strengthen or weaken the semantic orientation of scope. The weakening can be understood in two ways: i) if the word or scope of the semantic orientation is $+5$, $+4$, -5 or -4, their semantic orientation will not become negative because according to our methodology (shift negation), due to our SO value of the negators is ± 4. In contrast, ii) if the semantic orientation of scope or sentence is between -3 and $+3$, their semantic orientation will be reversed. iii) Finally, negation with conjunction, contrastive negation and lexicalized structures do not change the SO value of the scope.

5.1 Negation strengthening the SO

Among all the negation instances, we have observed some cases where the semantic orientation has been strengthened (1.96 %: 7 of 359). This happens when the negation marker *ezin* ("can not") modifies adjectives or adverbs.

(5) *Dena nahasten da maisulan* ***ezin*** *ederragoa*$_{(+4)}$ *osatzeko.* (MUS21)
Everything is mixed to create a masterpiece that can **not** be more beautiful$_{(+4)}$.

In Example 5, the negator modifies the adjective and, in this case, the negation with an adjective in a comparative structure is used to reinforce the positive SO value. The result

[5] **Bold** is used to mark the negator, underline means the scope of negation.

Example	Negation marker	Categorization	Instances
6	ez	[(NP +)] *ez* [+ aux. (+ NP) + verb (+ NP)]	214
		[(NP +) verb +] *ez*	18
		[NP +] *ez*	13
		ez [+ NP +] *ez* [+ NP] (...) (repetitive)	2
7	gabe	[NP/VP/clause +] *gabe*	41
	ezin	[(NP) + verb +] *ezin*	19
		ezin [(+ NP) +] verb [(+ NP)]	5
	salbu	[NP] + *salbu*	2
	izan ezik	[NP/clause] + *izan ezik*	1
	ezta	*ezta* + [NP/clause]	1
	ez, ezin	with any clear pattern	7
	Total		**323**

Table 2: Negation weakening the semantic orientation.

of negating a positive chunk *can not be more beautiful* is to be even more positive. In this case, the *masterpiece* is *very beautiful*.

5.2 Negation weakening the SO

In the majority of cases, the SO value is weakened due to negation. Several negation markers can weaken the semantic orientation. In our corpus, 89.98 % of cases (323 of 359) show a weakening of scope.

(6) *Horrek* ***ez*** *die eragotzi*$_{(-2)}$ *ordea, 57 milioi euro ematea San Mames klub pribatuari!.* (POL30)
It does **not** prevent$_{(-2)}$ them, however, to give 57 milion euros to San Mames private club!

(7) *Irabazi*$_{(+2)}$ ***ezin****ik jarraitzen du Eibarrek, baina oso puntu ona eskuratu du Getaferen zelaian.* (KIR17)
Eibar continues **without** winning$_{(+2)}$, but it has achieved a very good point in Getafe's (football) field.

In Example 6, the default word order of Basque (main verb + auxiliary verb) was reversed in a typical negation structure (*ez* "not" + auxiliary verb + main verb). In this example, the negation marker *ez* ("not") has an effect on all the words of the sentence, including the verb *eragotzi* ("prevent") which has a negative SO value (-2), weakening its SO value. In Example 7, the negation marker *ezin* "can not" negates the verb *irabazi* ("win"). Therefore, the negation marker *ezin* ("can not") works like an intensifier does with adjectives and adverbs (Example 5) while it has the opposite function with verbs and nouns (Example 7). Therefore, weakening negators can have a positive or negative (± 4) SO value, if the modified chunk (scope) has a positive or negative SO value. The same happens if the SO value is positive $+5$, because the result of the weakening (-4) will not change the polarity and the SO value will still be positive $+1$. In contrast, if the SO value of the modified chunk $+3$ or -3 or lower, the SO value will be reversed to a ± 1. This happens in Example 6 and Example 7. In the first example, the SO value of the scope is $+2$ (*eragotzi* ("prevent") -2 + *ez* ("not") $+4$ = $+2$). In the second one, the SO value of the scope is -2 (*irabazi* ("win") $+2$ + *ez* ("not") -4 = -2).

5.3 Negation with no effect

Negation with no effect on semantic orientation has happened in 8.08 % of our sample (27 of 359). In these cases, the negation does not modify any word with a SO value assigned. This can happen due to three reasons: i) the negator appears with a conjunction, ii) the negator is a part of contrastive negation and iii) the negator is part of a lexicalized structure (structures with their own meaning and sometimes also corresponding to dictionary entries). The scope concept is applicable only in the case of contrastive negation and the particle *ez* ("no") with a conjunction.

(8) *Ikuspuntu politikotik*$_{(-1)}$ *ez ezik, ekonomikotik*$_{(+3)}$ *ere Greziak esperantza ekarri du Europako hegoaldeko beste herrietara, tartean Euskal Herrira.* (POL08)
Not only from the political point of view, but also from the economic point of view, Greece has also hoped for other parts of southern Europe, including the Basque Country.

(9) *Sei puntu baino ez dituela, hamaseigarren postuan da Reala sailkapenean.* (KIR27)
With only six points, Real is in the sixteenth position in the classification.

Example 8 shows a contrastive negation with additive function (Silvennoinen, 2017). In other words, the negation mark does not negate the noun phrase, as in *ikuspuntu politikotik*$_{(-1)}$ ("from the political$_{(-1)}$ point of view"), actually it functions as conjunction

Example	Negation marker / lexicalized structure	Instances
	[verb/*bai* "yes"] + *edo/edota/ala ez* (*ez* with conjuction)	3
8	[NP] + *ez ezik* (contrastive negation)	2
9	*baino/besterik ez*	11
	Others lexicalized structures	13
	Total	**29**

Table 3: Negation without effects on semantic orientation.

and adds new information: *ekonomikotik ere* ("also from the economic point of view"). Structures of Table 3 have their own SO value, they can be considered as dictionary entries and they can appear in different positions in the sentence. In Example 9, the structure *baino/besterik ez* ("only") is an adverb.

6 Evaluation

6.1 Evaluation methodology

To tag the negation changes of the SO value, we have created negation rules based on previous studies.Rules have been implemented using Constraint Grammar (CG3) (Karlsson et al., 2011) to assign the correct value to the negated structures. The corpus of 96 texts has been tagged using the Basque morphosyntactic disambiguator based on the CG formalism (Aduriz et al., 1997). Then, a different set of 48 texts of the Basque Opinion Corpus has been used as test dataset to evaluate the rules. After that, the results have been analyzed manually, observing if the words have been annotated or not and, when annotated, whether they have the correct annotation.

Negation effects	**Prec.**	**Rec.**	**F_1**
Strengthen	1.00	1.00	1.00
Weaken	0.93	0.80	0.86
No effect	0.97	1.00	0.98
Total	0.93	0.80	0.86
Negated elements	**Prec.**	**Rec.**	**F_1**
Negation markers	1.00	0.96	0.98
Lexicalized structures	0.96	1.00	0.98
Scope	0.91	0.75	0.82
Total	0.93	0.80	0.86

Table 4: General results of negation effects and negated elements.

Most of the corpus was evaluated by one linguist, but with the aim to know the reliability of this evaluation a piece of the corpus (10 %) has been annotated by two linguists. Both annotators have followed a guideline to evaluate the output of CG3 rules. According to the results, the Cohen's kappa score is 0.93 for the annotation of the words that belong to negation and the kappa score is 0.69 for the annotation of words that have been annotated correctly, badly or is missed (which can be considered as *substantial* in (Landis and Koch, 1977)).

6.2 Results and error analysis

According to general results, the F_1 of the negation rules identifying elements related to negation is 0.86 (Precision is 0.93 while recall is 0.80).

In accordance with weakening and scope error analysis, these elements show lower F_1 score because they behave more irregularly. The components as well as the length in scope are more unpredictable. Moreover, some negators apply to lists of words with comma and, as some constraints in CG3 rules correspond to punctuation marks, they have not been detected. This suggests that the rules need more precision. So, the punctuation mark constraint is not enough. Therefore, some syntactic information is needed to detect these kind of structures.

7 Conclusions and Future Work

This work presents a negation analysis for Basque sentiment analysis based on Constraint Grammar rules. According to this study, the negation can affect the semantic orientation (SO value) in different ways: i) strengthening, ii) weakening or iii) having no effect. According to our evaluation to measure the identified words, the overall precision is 0.93, the recall 0.80 and the F_1 score 0.86. In line with error analysis, the punctuation mark constraint is not enough and more precise rules are needed in the negation weakening. In the near future, i) we want to implement these negation rules in a tool for automatic Basque sentiment analysis and ii) we want to continue with the analysis of negation: analyzing the scope in a bigger corpus and especially based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), studying if the position of negator in rhetorical structure has any effect on sentiment analysis.

References

Itziar Aduriz, José María Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Koldo Gojenola, and Montse Maritxalar. 1997. Morphosyntactic disambiguation for basque based on the constraint grammar formalism. *Proceedings of Recent Advances in NLP (RANLP97)*, pages 282–288, Tzigov Chark (Bulgary).

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2016. Creating and evaluating a polarity - balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis*, pages 58–62. Santiago de Compostela (Spain).

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2018. SentiTegi: building a semantic oriented Basque lexicon. In *Proceedings of the CICLing 2018*. Hanoi (Vietnam).

María Jesus Aranzabe Begoña Altuna and Arantza Díaz de Ilarraza. 2017. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria and Patxi Salaberri (pub.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea*, pages 127–134. Udako Euskal Unibertsitatea (UEU), Bilbo (Spain).

Nerea Ezeiza, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics.

Rodney Huddleston and Geoffrey Keith Pullum. 2002. The Cambridge grammar of English. *Language. Cambridge: Cambridge University Press*.

Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, M. Dolores Molina-González, and L. Alfonso Ureña-López. 2018a. Relevance of the SFU Review SP-NEG corpus annotated with the scope of negation for supervised polarity classification in Spanish. *Information Processing & Management*, 54(2):240–251.

Salud María Jiménez Zafra, Eugenio Martínez Cámara, María Teresa Martín Valdivia, and María Dolores Molina González. 2015. Tratamiento de la Negación en el Análisis de Opiniones en Espanol. *Procesamiento del Lenguaje Natural*, (54).

Salud María Jiménez-Zafra, Mariona Taulé, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and M. Antónia Martí. 2018b. SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Jingjing Liu and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 161–169. Association for Computational Linguistics.

Eugene Emil Loos, Susan Anderson, Dwight H. Day, Paul C. Jordan, and J. Douglas Wingate. 2004. *Glossary of linguistic terms*, volume 29. SIL International Camp Wisdom Road Dallas.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Olli O. Silvennoinen. 2017. Not only apples but also oranges: Contrastive negation and register. *In Turo Hiltunen, Joe McVeigh and Tanja Sily (edit.), Big and Rich Data in English Corpus Linguistics: Methods and Explorations, VARIENG, Helsinki (Finland)*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2015. A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, 21(1):139–163.

4.2 Article 4

Using relational discourse structure information in Basque sentiment analysis

Using relational discourse structure information in Basque sentiment analysis

El uso de la información de la estructura retórica en el analisis de sentimiento

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta y **Alicia Perez**
IXA Group. University of the Basque Country
jon.alkorta@ehu.eus, koldo.gojenola@ehu.eus, mikel.iruskieta@ehu.eus, alicia.perez@ehu.eus

Resumen: En este artículo presentamos un estudio sobre el análisis del sentimiento que explota información extraida de la estructura relacional discursiva en un corpus en euskera sobre crítica literaria. Para el análisis discursivo hemos utilizado la *Rhetorical Structure Theory* (RST) y para la polaridad el método QWN-PPV. Los resultados preliminares demuestran que el análisis del discurso es efectivo para el análisis de opiniones.
Palabras clave: Análisis del sentimiento, polaridad, relaciones de coherencia, unidad central, RST, crítica literaria

Abstract: This paper presents a study in sentiment analysis which exploits information of the relational discourse structure in a Basque corpus consisting of literature reviews. The QWN-PPV method was employed to label all the texts at element level and the *Rhetorical Structure Theory* (RST) was used to extract discourse structure information. The preliminary results show that discourse structure is effective for opinion mining.
Keywords: Sentiment analysis, polarity, coherence relations, central unit, RST, literary criticism

1 Introduction

Sentiment analysis is nowadays a well known topic where the opinion, sentiment or subjectivity (Pang and Liu, 2008) are studied. The opinion about films (Pang and Lee, 2004), the success of politicians (Tumasjan et al., 2010) and the opinion of consumers about products are some of the topics studied.

Different levels of language has been studied in Sentiment Analysis. Hatzivassiloglou and McKeown (1997) studied the word level, Yu and Hatzivassiloglou (2003) the sentence level and Pang, Lee, and Vaithyanathan (2002) the discourse level. These are some examples of these three levels extracted form our corpus[1]:

i) Lexical level: where words[2] and entities have their own polarity[3], as in Example (1).

(1) [...] *literatura ona*$_{(+)}$ *sortu ahal izateko.* BER01
[...] to create good$_{(+)}$ literature.

In the example the word *ona* (good) has a positive polarity, because its entry in a dictionary has a positive value.

ii) Syntactic level: where the function in the word ordering or the clause's syntactic function can change the polarity assigned in the lexical level.

(2) *xede*$_{(+)}$ *onak*$_{(+)}$ *ez dira nahikoak*$_{(+)}$ *izaten literatura ona*$_{(+)}$ *sortu ahal izateko.* BER01
good$_{(+)}$ goals$_{(+)}$ are not enough$_{(+)}$ to create a good$_{(+)}$ literature.

In Example (2), the negation *ez* (not) changes the polarity assigned by the dictionary entries to a negative polarity determined by the negation of an otherwise positive statement.

[1] References and links to see the annotated text are at the end of the examples.

[2] For example, SentiWordNet (Esuli and Sebastiani, 2010) is a lexical resource for opinion mining which assigns three sentiment scores to each synset of WordNet: positivity, negativity, objectivity.

[3] In the following examples the polarity will be marked with: (+) when positive, (−) when negative and (∗) when neutral. All the examples were analyzed with the QWN-PPV method, explained below.

iii) Discourse level: where the coherence relations can highlight or even change a clause polarity (micro-structure), or the overall polarity of a text (macro-structure).

In Example (3) the change of polarity is out of the sentence scope, at discourse level.

(3) *Dokumentazio lana, esan bezala, nabarmena*$_{(+)}$ *da, eta baliabideen erabileran idazleak duen ahalmena eta egindako lana bereziki azpimarratzekoak*$_{(+)}$ *dira. Baina, horiek horrela izanik ere, emaitza zalantzagarria*$_{(*)}$ *da.* BER04
The documentation work, as mentioned before, is spectacular$_{(+)}$, and the capacity of the writer to use the resources and the work done especially are to underline$_{(+)}$. But, although that is so, the result is doubtful$_{(*)}$.

In this example, there are several words with a positive polarity, but the polarity of the example is not positive because a contrast discourse relation signaled by the adversative connector *baina* (but) has changed it.

This example demonstrates the importance of rhetorical relations, which can change the polarity of the sentence. For that reason, it is necessary, from our point of view, to also consider the discourse structure information in sentiment analysis.[4]

Currently there exists an Opinion Mining system for Basque. We have used this tool[5], that assigns automatically a positive or negative polarity to words[6]. The system makes use of the QWN-PPV method (Vicente, Agerri, and Rigau, 2014), that automatically generates polarity lexicons annotated at synset and lemma level. For that purpose, QWN-PPV uses a Lexical Knowledge Base (WordNet) and a list of positive and negative elements. QWN-PPV is a method that detects elements from lexical level and, consequently, the method is unable to correctly detect the polarity of the examples mentioned before —see examples (2) and (3).

[4]Example (4) below also shows the importance of discourse structure considering the main topic of the text and the rhetorical relation, when assigning a polarity score.

[5]Ber2Tek Opinion Mining can be tested at `http://iritzierauzketa.ber2tek.eus/`

[6]The method does not signal a neutral polarity.

With the aim of fulfilling this gap, we want to develop a method based in two language levels: the lexical and the discourse level. To that end, we will estimate the importance that the discourse structure has in sentiment analysis. Our study, which is based on a theoretical approach based on discourse, exploits information from the relational discourse structure in a Basque corpus consisting of literature reviews. The theory we employ to that end is the *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987)[7]. This theory describes the structure and coherence of text and it has been useful in Sentiment Analysis and in other many NLP advanced tasks (Taboada and Mann, 2006). In this respect, this work is a first approximation to Opinion Mining using discourse structure in Basque. This study, as far as we know, is the first work on Basque and sentiment analysis from a discourse point of view. Therefore, it fullfils this gap in Basque, but it is also relevant for RST, because it includes a different language to some recent works like (Trnavac and Taboada, 2014) in English and (Zhou et al., 2011) in Chinese.

The rest of the paper is structured as follows. Section 2 lays out the related work. Section 3 sets out the methodology we used and Section 4 presents the results. Finally, Section 5 presents a discussion and establishes directions for future work.

2 Related work

As we have explained before, Sentiment Analysis has three linguistic different levels, while we will focus on the lexical level and the discourse level interaction. Inside the discourse level, there are various categorizations according to different viewpoints.

In the discourse level there are two possible methods: language model and knowledge-based model. The first one determines if a span of a text is subjective, while the second one finds words with its polarity in a dictionary and calculates a sentiment score for all the text with an algorithm.

i) In the language model approach, Alistair and Diana (2005) use Support Vector Machines to classify sentiment expressed by movie reviews. Firstly, they use unigram fea-

[7]Other theories worth to mention are the *Segmented Discourse Representation Theory* (SDRT) (Asher, 1993) and the *Penn Discourse TreeBank* (PDTB) (Miltsakaki et al., 2005).

tures and then, they couple bigrams. The bigrams are composed by a valence shifter and another word. The results are acceptable and adding a term-counting method helps get better results.
ii) The knowledge-based approximation can be divided in four approaches (Cambria et al., 2013): keyword spotting, lexical affinity, statistical methods, and concept-level techniques. In the experiments, we will use the QWN-PPV method, which is an (almost) unsupervised method, i.e., a statistical method.

According to (Zhou, 2013), from a theoretical point of view, there are two approaches to Sentiment Analysis: discourse-based and aspect-based.
a) In a discourse-based approach not all the sentences have the same importance. Several researchers have tried to measure the contribution of sentences or phrases to the polarity of the text. Discourse Structure based Sentiment Analysis is divided in two approaches: rule based and weight based ones. In both approaches the results have improved with the addition of discourse relations. Moreover, these works have shown that the combination of two paradigms can bring an overall improvement.

In the rule based approach, Somasundaran et al. (2009) use a supervised collective classification and a supervised optimization framework in order to improve polarity classification. Text spans are extracted according to their importance in discourse structure.

Vanzo, Croce, and Basili (2014) assign a sentiment polarity to entire tweet sequences using a Markovian formulation of the Support Vector Machine discriminative model, SVM_{hmm}. In contextual information, they take into account two aspects: the conversation and the user attitude or the overall attitude of the last tweets. The individual perspective is independent in context, so they consider the tweet as a multifaceted entity. Consequently, each vector contributes in one aspect of the overall representation. The evaluation shows that sequential tagging effectively improves the detection precision approximately 20% in F1 measure.

In the weight based approach, Polanyi and Zaenen (2006) demonstrate that the structure of the text gives important information to extract the opinion. They have found that connectors increase or decrease the intensity of polarity. In this way, discourse relations can also increase or decrease the intensity.

Inspired in this previous work, Taboada, Voll, and Brooke (2008) extract the most important spans of a text and then, they calculate the semantic orientation in two ways, where the most important spans weight more. First, they use RST and they extract all the nuclei of the text. After that, they use a topic classifier based in support vector machines, improving results.

The rhetorical information of a text can be extracted automatically with a discourse parser and then this information can be used in sentiment analysis to determine the polarity of the text in a more reliable way. For example, Taboada, Voll, and Brooke (2008) and Heerschop, Goossen, and Hogenboom (2011) use the Sentence-level PArsing of DiscoursE (SPADE) tool in order to extract the discourse relations automatically from the text (Soricut and Marcu, 2003).
b) In Aspect-based Sentiment Analysis, the subject of a review is important because it helps to predict the "relevant" polarity expressed in the text. In this way, words related with the aspect help to give more accurate results.

Lim and Buntine (2014) combine language model and aspect creating the LDA-based[8] Twitter Opinion Topic Model. This model uses strong sentiment words, hashtags, mentions and emoticons to predict the opinion modeling the target.

The OpeNER project (Opener, 2013) makes use of three components: *i*) opinion express, *ii*) opinion holder and *iii*) opinion target. There are four tag levels in the Annotation Tool of the project. Tagging is based in three parameters: positive / negative attitude, sentence "on-topic" and "to-the-point"[9]. In the project, topic sentence can be a touristic attraction, a restaurant or a hostel. The opinions indirectly linked with the reviewed entity are also considered as "on-topic". The "to-the-point" category implies that a reviewer gives an opinion of the annotation object and then expresses a lot of details of it.

The Replab project developed the *Reputation online*. Spina, Gonzalo, and Amigó (2014) added Twitter signals to content sig-

[8]Latent Dirichlet Allocation is a opinion model used for Opinion Mining

[9]The OpeNER project can be consulted at `http://www.opener-project.eu/`.

nals to improve topic detection and they learnt a similarity function in order to supervise the topic detection clustering process. The last aim of this work is to solve the reputation monitoring problem automatically. They made use of different entities (*Maroon 5, Yamaha, Ferrari*) in the task. The difference with the current work is that their text is unordered (tweets) while ours is ordered (literary criticism).

Our method is a combination of two approaches: based on discourse structure and based on aspect. On the one hand, our approach is based on discourse structure because we will use RST in order to put different weights to Elementary Discourse Units (EDUs) according to their position in a discourse tree. On the other hand, our approximation is also aspect-based because we want to identify the words related with the main topic in order to get better results.

We think that the implementation of discourse structure together with the QWN-PPV method can improve the results. In other words, the polarity of the text will be better assessed. In this paper, we will do a first approach analyzing discourse topic and its influence on structures of attitude. In the following Example (4), we show the results of the QWN-PPV tool on the whole AIZ02 text.

(4) Number of words containing sentiment found: 7
Polarity score: 0.22
Polarity (if threshold > 0.0): positive
Gustura$_{(+)}$ *irakurtzen da nobela, protagonistaren*$_{(+)}$ *joko*$_{(+)}$ *bikoitza nola bukatuko ote den, nahiz eta amaiera horren zantzuak aurretik eskaintzen*$_{(+)}$ *dizkigun*$_{(+)}$ *idazleak. Idazkerak ere laguntzen*$_{(+)}$ *du aurrera plazerez*$_{(+)}$ *egiten.* AIZ02
The novel is read with pleasure$_{(+)}$, how is going to finish the$_{(+)}$ double$_{(+)}$ game$_{(+)}$ of$_{(+)}$ the$_{(+)}$ protagonist$_{(+)}$, although the narrator previously gives$_{(+)}$ us$_{(+)}$ some clues of the ending. Writing also helps to go along with pleasure$_{(+)}$.

The QWN-PPV method determines any positive word with a positive value (+1) and any negative word with a negative value (−1). Then, a polarity score (0.22) is estimated for the text. To do so, both positive and negative words are counted and divided by the number of the words in the text. If there are more positive words, as in Example (4), the polarity score will be higher (0.23) than the threshold (zero) and, therefore, the overall polarity of the text will be positive.

On the other hand, the QWN-PPV method estimates a lower polarity score (0.11) for all the AIZ02 text. So, the example shows how coherence relations related to the discourse topic can contribute to a better assignment of the text polarity.

3 Methodology

We have used the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to achieve the rhetorical information of the text. The main concepts of RST are nuclearity of text spans[10] and coherence relations among text spans. With these two concepts a hierarchical tree structure (RS-tree) of coherence can be build, where all the text spans have a function in the tree, because relations are recursive (one relation can work as one of the spans in another relation). RST relations are hypotactic —hierarchical relations with one nucleus (N) and one satellite (S), e.g.: ELABORATION, JUSTIFY, EVALUATION, CAUSE...— and paratactic —discourse coordination where all the discourse units are nucleus, e.g. CONTRAST, DISJUNCTION, CONJUNCTION...[11]

In a hierarchical RS-tree there is always a Central Unit which is the most salient EDU (Iruskieta, Diaz de Ilarraza, and Lersundi, 2014). For example, in scientific abstracts authors often show explicitly the discourse topic as follows: "the principal aim of this paper is to investigate..." (Paice, 1980).

To show an example of the discourse structure we want to use, a partial RS-tree of AIZ02 in Figure 1 is presented. The central unit of the tree structure is represented with straight vertical lines (the unit 2-2 in the example). The annotator interpreted the RS-tree as follows:

a) PREPARATION for the article, by means of the title ([1−1 > 2−10]).

[10] Discourse structure is recursive so there are formally two different units: *a*) Elementary Discourse Unit and *b*) group of EDUs.

[11] A more detailed explanation of RST can be found at http://www.sfu.ca/rst/ and in Basque at http://www.sfu.ca/rst/07basque/index.html.

b) with the highest EVALUATION linked to the central unit she interprets that it is evaluating all the propositions mentioned before, that can be taken as all the work ([2−14 < 15−20]).

c) with the lowest EVALUATION linked to the central unit, she is evaluating an aspect of the work, only the proposition mentioned in the central unit ([6−6 < 7−7]).

These are the steps taken to carry out this study:

i) Building a corpus. We have collected a corpus of 28 texts, where 19 of them will be used for training and the remaining 9 for testing.[12] The texts are reviews of Basque Literature works. The size of the texts is not uniform: the shortest one has 106 words and the longest one 485 words. A corpus description is presented in Table 1 and the annotated corpus can be consulted in the Basque RST Treebank at `http://ixa2.si.ehu.es/diskurtsoa/en/` (Iruskieta et al., 2013).

Text	Doc.	EDUs	Words
CRITICS	28	1038	8823

Table 1: Corpus description

ii) EDU segmentation of texts. Before preparing the experiment, we have processed the texts. Firstly, we used EusEduSeg (Iruskieta and Zapirain, 2015) a discourse segmenter to segment all the texts automatically.[13] After that, the segmentation has been corrected manually to avoid losing rhetorical information in subsequent phases.

iii) Corpus annotation. After segmentation, we have annotated the most salient EDU or the central unit and after it we have tagged all rhetorical structures of the text with the RSTTool (O'Donnell, 1997) using the Basque extended relations of RST.

iv) Central Unit (CU) and Polarity gold standard. To do so, we have made up a questionnaire based on Google Forms, where 20 annotators participated in the annotation. This was done in order to have a gold standard which we could use to compare the results.
Our gold standard was collected as follows: *i*) the central unit must be selected at least by four participants. If not, we selected the three most voted central units.[14] *ii*) The polarity of each text was conformed with the average of all the annotators, in two ways:
- Polarity 1 (P1): quantitative polarity annotation from 1 to 5.
- Polarity 2 (P2): qualitative polarity description with three values: negative, neutral and positive.

v) Manual extraction of text spans composed with the text of the central unit and the EVALUATION relation. We have manually built different features based on the rhetorical structure tree:
- ALL (F1): the result of QWN-PPV on the full text.
- CU (F2): only the central unit.
- CU-H-EV (F3): the central unit and the highest EVALUATION relation linked to it.
- CU-ALL-EV (F4): the central unit and all the EVALUATION relations linked to it.

Table 2 shows all the glosses we have used to perform the analysis.

Gloss	Meaning
P1	Polarity of five categories
P2	Polarity of three categories
F1	QWN-PPN for all the text
F2	The central unit
F3	The central unit and the highest EVALUATION relation
F4	All the EVALUATION relations of the text

Table 2: Glosses of the predicted variables and features.

vi) Lemmatization and polarity extraction with the QWN-PPV. Before running QWN-PPV, we run Eustagger (Ezeiza et al., 1998) to divide the sentences into unambiguous tokens.[15]
After that, we run the QWN-PPV

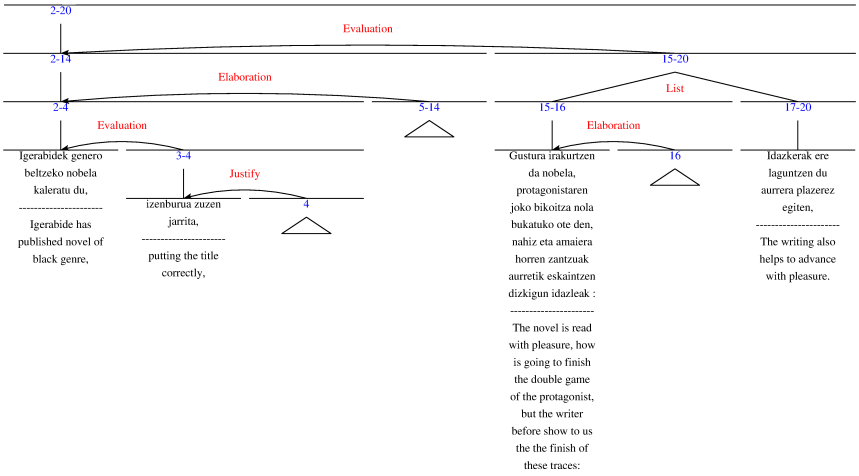
[12] All the texts are available in *Kritiken Hemeroteka* at `http://kritikak.armiarma.eus/`

[13] EusEduSeg can be tested at `http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl`.

[14] Hearst (1997) considered that a subtopic boundary was true if at least three out of seven (42.86%) annotators placed a boundary mark.

[15] Eustagger is a lemmatizer and tagger for Basque based on Stochastic and Rule-Based Methods. Eustagger can be tested at `http://ixa2.si.ehu.es/demo/analisimorf.jsp`.

Figure 1: A partial RS-tree of AIZ02



method (Vicente, Agerri, and Rigau, 2014) and obtained the polarity for each of the four features.

vii) Analysis of results. We have used two methods in order to analyze the results: *Logistic Regression* (LR) and *Sequential Minimal Optimization* (SMO).
They are well known though efficient techniques that have often been used as baseline. The first is adequate for regression problems as in this case, where the P1 class is a numeric polarity from 1 to 5. The second one tackles classification, that is, the class to guess (P2) is nominal and it was obtained by a straightforward discretization mechanism (positive, negative and neutral). Both methods are implemented with open libraries. We calculate percent agreement and precision, recall and f-measure as follows:

$$precision = \frac{correct_{polarity}}{correct_{polarity} + excess_{polarity}}$$

$$recall = \frac{correct_{polarity}}{correct_{polarity} + missed_{polarity}}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

where $correct_{polarity}$ is the number of correct *polarity items*, $excess_{polarity}$ is the number of overpredicted *polarity items* and $missed_{polarity}$ is the number of *polarity items* the system missed to tag.

A summary of the methodology we have employed is presented in Figure 2.

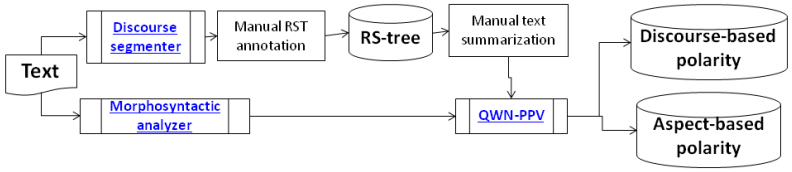
4 Results

Firstly, we present the results of all the features when trying to guess P1 (five categories). The results of each feature and the best combinations are presented in Table 3.

When individual features were considered, F1 with LR obtained the best results (0.37), while F2, F3 and F4 obtain lower results, with F4's contribution near to zero. But when combinations were considered, using features F1, F3 and F4 together (F134) with SMO obtained a better result (0.40). The results show that, in guessing P1, there is a gain employing combinations of features based on discourse structure.

Secondly, we will test the same algorithms to try to guess P2, based on three categories. The results are presented in Table 4. When using individual features, F1 and F2 with LR obtain the best results (0.47). Looking at the combinations, F1234 with SMO gives the

Figure 2: System architecture



Method	Feature	Fm
LR	**F1**	**0.37**
	F2	0.21
	F3	0.32
	F4	-
	F134	0.28
	F1234	0.30
SMO	**F1**	0.23
	F2	0.19
	F3	0.27
	F4	-
	F134	**0.40**
	F1234	0.38

Table 3: Results guessing P1 (five categories, cross-validation on the development set).

Feature	Method	Fm
LR	**F1**	**0.47**
	F2	**0.47**
	F3	0.44
	F4	-
	F134	**0.52**
	F1234	0.39
SMO	**F1**	0.37
	F2	0.36
	F3	0.36
	F4	-
	F134	0.50
	F1234	**0.53**

Table 4: Results on P2 (three categories, cross-validation on the development set).

best results (0.53). Overall, the results show that when using discourse structure (combinations), the results on P2 improve considerably.

To sum up, in the case of P1 with five categories (see Table 3) SMO is the best single algorithm for prediction. In contrast, SMO gave the best results when using a combination of features, with an F-measure of 0.40.

In the case of P2 with three categories (see Table 4), LR continues to be the best method using a single feature with an F-measure of 0.47. In combinations, SMO gives the best results (0.53).

After examining the results on the development set, we test the best methods (LR on the individual features and SMO on the combinations) on the test set. We show the results in Table 5.

	Feature	Method	Fm
P1	**LR**	**F1**	0.09
	SMO	**F134**	0.09
P2	**LR**	**F1**	0.59
	LR	**F134**	**0.84**
	SMO	**F1**	0.40
	SMO	**F1234**	0.73

Table 5: Test set results for P1 and P2

When guessing P1, the best results are obtained with F1 using LR and H134 using SMO. That means that the algorithms based in discourse structure we have used are not able to guess P1 accurately, possibly because the small size of the corpus to deal with five categories. In contrast, the results to guess P2 (three categories) using combinations of features based on discourse structure are considerably better than considering all the text (F1), with 0.84 and 0.73 for LR and SMO, respectively. Therefore, it seems that the implementation of discourse features can improve the results in opinion mining.

Performing a first error analysis, the confusion matrix of F134 guessing P2 with SMO shows that a text with neutral polarity has been classified as having a negative polarity. The error is not specially important, as a matter of fact, the QWN-PPV method puts only positive or negative polarity to the texts. So, the difference is that the method has two main categories and we use three categories.

If we analyze Example (5), we can see that the Central Unit of the text is neutral but the method considers it a negative text.

(5) *Aho gustu*$_{(+)}$ *gazi-gozoa utzi*$_{(*)}$ *dit*$_{(+)}$ *[...]ren bigarren ipuin liburuak.* [...] BER03
The second storybook of [...] has$_{(+)}$ left$_{(*)}$ me a sweet-and-sour taste$_{(+)}$ in the mouth.[...]

In the example, the word *gazi-gozoa* 'sweet-and-sour' is a neutral word but the QWN-PPV does not detect it. Moreover, some words of the remaining text are not tagged with their correct polarity.

5 Conclusions and future work

We have presented a set of algorithms in which we have tried to examine the importance of discourse structure information in Opinion Mining. Firstly, we have concluded that guessing the polarity of the text of literary reviews based on three categories gives better results than the one with five categories for Opinion Mining. This could be due to the fact that the task is easier and also to the reduced size of the training set.

The second conclusion is that combining several discourse structures is the best method for Logistic Regression giving an F-measure of 0.84 (and also for SMO with an F-measure of 0.73). An error analysis has shown that the errors were soft: a text with neutral polarity has been misclassified as having negative polarity. Looking at the importance of each individual feature, we can say that an important weight related to polarity is situated in the EVALUATION discourse relation.

In future works, our aim is to:

- Annotate a bigger corpus. The perliminary experiments performed in this work should be validated using a bigger corpus.
- Implement a full set of experiments on combinations of the central unit and the EVALUATION relation. We want to study if there is any difference with the relations which are attached to the central unit and the relations which are not.
- Test other phenomena of discourse structure such as the nuclearity (satellite vs. nucleus) and INTERPRETATION relations, to check if they have any influence on polarity.
- Automatize all the system, by testing in partial RS-trees, where only the central unit and EVALUATION relations linked to it were considered.
- Study which syntactic and discourse structures are more important (i.e., they change the polarity of lower levels).
- Implement an automatic annotator of word level polarity based on a supervised dictionary, to solve some problems observed in the QWN-PPV.

Bibliografía

[Alistair and Diana2005] Alistair, Kennedy and Inkpen Diana. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*.

[Asher1993] Asher, Nicholas. 1993. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

[Cambria et al.2013] Cambria, Erik, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. In *IEEE Intelligent Systems*, volume 28, pages 15–21, Piscataway, NJ, USA, March.

[Esuli and Sebastiani2010] Esuli, A. and F. Sebastiani. 2010. SentiWordNet: A publibly available lexical resource for opinion mining. In *Proceedings of 5th International Conference on LREC*, pages 417–422.

[Ezeiza et al.1998] Ezeiza, Nerea, Itziar Aduriz, Iñaki Alegria, Jose Mari Arriola, and Ruben Urizar. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *COLING-ACL'98*, volume 1, pages 380–384, Canada, August 10-14.

[Hatzivassiloglou and McKeown1997] Hatzivassiloglou, Vasileios and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for*

computational linguistics, pages 174–181. Association for Computational Linguistics.

[Hearst1997] Hearst, M. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. In *Computational Linguistics*, volume 23, pages 33–64, March.

[Heerschop, Goossen, and Hogenboom2011] Heerschop, B., F. Goossen, and A. Hogenboom. 2011. Polarity Analysis of Texts using Discourse Structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1061–1070, Glasgow, Scotland, UK, October 24–28.

[Iruskieta et al.2013] Iruskieta, Mikel, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.

[Iruskieta, Diaz de Ilarraza, and Lersundi2014] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475. Dublin City University and ACL, Dublin, Ireland.

[Iruskieta and Zapirain2015] Iruskieta, Mikel and Beñat Zapirain. 2015. EusEduSeg: A Dependency-Based EDU Segmentation for Basque. In *SEPNL*, Alicante, September 16-18.

[Lim and Buntine2014] Lim, Kar Wai and Wray Buntine. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1319–1328. ACM.

[Mann and Thompson1987] Mann, William C and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

[Mann and Thompson1988] Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. In *Text*, volume 8, pages 243–281.

[Miltsakaki et al.2005] Miltsakaki, Eleni, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005), Barcelona, Spain, December*.

[O'Donnell1997] O'Donnell, Michael. 1997. Rst-tool: An rst analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, Duisburg, Germany.

[Opener2013] Opener. 2013. OpeNER annotation guidelines. Annotation of Costumer Reviews in the Touristic Domain V5.0. pages 1–19.

[Paice1980] Paice, Chris D. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *3rd annual ACM conference on Research and development in information retrieval*, pages 172–191, Cambridge, June. Butterworth and Co.

[Pang and Liu2008] Pang and N. Liu. 2008. Opinion Mining and Sentiment Analysis. In *Foundations and trends in information retrieval*, volume 2, pages 1–135.

[Pang and Lee2004] Pang, B. and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimun cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271.

[Pang, Lee, and Vaithyanathan2002] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Polanyi and Zaenen2006] Polanyi, Livia and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect*

in text: Theory and applications. Springer, pages 1–10.

[Somasundaran et al.2009] Somasundaran, Swapna, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 170–179. Association for Computational Linguistics.

[Soricut and Marcu2003] Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

[Spina, Gonzalo, and Amigó2014] Spina, Damiano, Julio Gonzalo, and Enrique Amigó. 2014. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 527–536. ACM.

[Taboada, Voll, and Brooke2008] Taboada, M., K. Voll, and J. Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. In *Simon Fraser Univeristy School of Computing Science Technical Report*.

[Taboada and Mann2006] Taboada, Maite and William C Mann. 2006. Applications of rhetorical structure theory. *Discourse studies*, 8(4):567–588.

[Trnavac and Taboada2014] Trnavac, Radoslava and Maite Taboada. 2014. Discourse structure and attitudinal valence of opinion words in sentiment extraction. *LSA Annual Meeting Extended Abstracts*.

[Tumasjan et al.2010] Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. T. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*, number 10, pages 178–185.

[Vanzo, Croce, and Basili2014] Vanzo, Andrea, Danilo Croce, and Roberto Basili. 2014. A context-based model for Sentiment Analysis in Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2345–2354, Dublin, Ireland, August.

[Vicente, Agerri, and Rigau2014] Vicente, Iñaki San, Rodrigo Agerri, and German Rigau. 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, April 26-30.

[Yu and Hatzivassiloglou2003] Yu, Hong and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics.

[Zhou et al.2011] Zhou, Lanjun, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.

[Zhou2013] Zhou, Yudong. 2013. Fine-grained Sentiment Analysis with Discourse Structure. In *Master Thesis. Saarland University*, October.

4.3 Article 5

Using lexical level information in discourse structures for Basque sentiment analysis

Using lexical level information in discourse structures for Basque sentiment analysis

Jon Alkorta
IXA research group
UPV-EHU
jon.alkorta@ehu.eus

Koldo Gojenola
IXA research group
UPV-EHU
koldo.gojenola@ehu.eus

Mikel Iruskieta
IXA research group
UPV-EHU
mikel.iruskieta@ehu.eus

Maite Taboada
Discourse Processing Lab
Simon Fraser University
mtaboada@sfu.ca

Abstract

Systems for opinion and sentiment analysis rely on different resources: a lexicon, annotated corpora and constraints (morphological, syntactic or discursive), depending on the nature of the language or text type. In this respect, Basque is a language with fewer linguistic resources and tools than other languages, like English or Spanish. The aim of this work is to study whether some kinds of discourse structures based on nuclearity are sufficient to correctly assign positive and negative polarity with a lexicon-based approach for sentiment analysis. The evaluation is performed in two phases: i) Text extraction following some constraints on discourse structure from manually annotated trees. ii) Automatic annotation of semantic orientation (or polarity). Results show that the method is useful to detect all positive cases, but fails with the negative ones. An error analysis shows that negative cases have to be addressed in a different way. The immediate results of this work include an evaluation on how discourse structure can be exploited in Basque. In the future, we will also publish a manually created Basque dictionary to use in sentiment analysis tasks.

1 Introduction

Sentiment analysis is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2012, p. 7).

Automatic sentiment analysis is an area in continuous development. It first started with the identification of subjectivity (Wiebe, 2000) and, after that, polarity identification and measurement of strength have become the center of new developments (Turney, 2002). The objectives of sentiment analysis are evolving as well, as different types of information are used. For instance, initially, entity- and aspect-based information was used (Hu and Liu, 2004) but, later, new types of information, such as discourse structure information, have been used (Polanyi and Zaenen, 2006).[1]

This study is the first work that examines lexical and discourse structure information for sentiment analysis of Basque. The main aim is to evaluate which discourse structures can help in polarity detection following a lexicon-based approach. Our hypothesis is that some discourse structures are more related to opinions than others and we want to identify and study how they can help in a sentiment analysis task.

The paper is organized as follows: Section 2 discusses related works. Section 3 explains the methodology of the study and Section 4 presents the results and error analysis. Finally, conclusions and future work are given in Section 5.

2 Related Work

Various studies from different theoretical approaches analyze the influence of nuclearity and some rhetorical relations in sentiment analysis tasks. For example, Zhou et al. (2011) use discursive in-

[1] See a detailed review of sentiment analysis in Taboada (2016).

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 39–47,
Santiago de Compostela, Spain, September 4 2017.

formation in Chinese to eliminate noise at the intra-sentence level, improving not only polarity classification but also the labeling of rhetorical relations at sentence level.

Wu and Qiu (2012) analyze sentiment analysis based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in Chinese texts. They split texts in segments and, then, they train weights taking into account relations and nuclearity, showing that CONTRAST, CAUSE, CONDITION and GENERALIZATION have a more important role in this task than other discourse relations. Bhatia et al. (2015) use a simpler classification of relations into CONTRAST or NON-CONTRAST, and they show that the distinction improves the results of bag-of-words classifiers using Rhetorical Recursive Neural Networks.

Chardon et al. (2013) rate documents using three approaches: i) bag-of-words, ii) partial discourse information and iii) full discourse information. The discursive approach gives the best result in the framework of Segmented Discursive Representation Theory (SDRT).

Trnavac et al. (2016) propose that a few rhetorical relations have a significant effect on polarity: CONCESSION, CONTRAST, EVALUATION and RESULT. They also conclude that nuclei tend to contain more evaluative words than satellites.

Alkorta et al. (2015) analyze which features perform better in order to detect the polarity of texts using machine learning techniques on Basque texts. Their results show that discourse structure is needed to improve results along with other types of features. They use a dictionary created by automatic means with an unsupervised method (Vicente et al., 2017). The dictionary values of their work are binary (-1 for negative polarity and $+1$ for a positive one).

In this work, we analyze which coherence relations could help to improve lexicon-based sentiment analysis, so that we can assign different weights to discourse structures following Bhatia et al. (2015) when calculating sentiment analysis for a whole text. For this task, we use the RST framework.

The main contributions of this work are: i) A fine-grained dictionary, manually created for Basque with 5 different negative values and 5 different positive ones, ranging from -5 to $+5$. ii) A study of how discourse structure interacts with this polarity lexicon.

3 Methodology

The subsections below detail the main steps followed in the present study.

3.1 Extraction of discourse structures

In the first phase, different discourse structures were compared. They will be used to determine which ones can be helpful in sentiment analysis. To extract as many discourse structures as possible, we use the corpus described in Alkorta et al. (2016), annotated for discourse relations according to RST.

The corpus contains 29 book reviews. Regarding polarity, it is a balanced corpus, with 14 positive reviews and 15 negative ones. The majority of reviews were collected from a website specialized in Basque literary reviews (Kritiken Hemeroteka).[2]

The following subcorpora were created, following some discourse constraints:

- Full text, containing all the RS-tree of the text.
- Texts extracted from central units (CU)[3] of the text.
- Text spans extracted from the CU of the text and from the central subconstituent (CS)[4] of some rhetorical relations (see Table 1).

Relation	CS	Relation	CS
ELABORATION	34	CONCESSION	2
EVALUATION	32	RESTATEMENT	2
PREPARATION	32	SUMMARY	2
BACKGROUND	13	ANTITHESIS	1
CIRCUMSTANCE	8	PURPOSE	1
INTERPRETATION	6	MOTIVATION	1
CAUSE	4	JUSTIFY	1

Table 1: Number of central subconstituents (CS) in the corpus per relation type linked to the CU.

We extracted 139 instances of rhetorical relations from our corpus. For some relations, such as ELABORATION and PREPARATION (66 of 139), we do

[2] http://kritikak.armiarma.eus/.

[3] Central units are defined as the most important EDU (Elementary Discourse Unit), and it is the main nucleus when tree structure is constructed (Iruskieta, 2014).

[4] Central subconstituents are "the most important unit of the modifier span that is the most important unit of the satellite span" (Iruskieta et al., 2015, p. 5).

not expect them to contain important polarity information, because these relations only add extra information to the central unit. In fact, Mann and Thompson (1988, p. 273) mention that in the case of ELABORATION "R(eader) recognized the situation presented in S(atellite) as providing additional detail for N(uclei). R(eader) identifies the element of subject matter for which detail is provided". Similarly, in PREPARATION "R(eader) is more ready, interested or oriented for reading N(uclei)". We did not take into account relations with low frequency (a single instance), such as MOTIVATION, JUSTIFICATION, ANTITHESIS and PURPOSE. Consequently, we will work with a subcorpus containing 69 relations, where almost half of them are central subconstituents of EVALUATION.[5]

3.2 Polarity extraction and evaluation

Polarity was extracted from all the discourse structures using a dictionary (v1.0) of words annotated with their semantic orientation: polarity (positive or negative) and strength (from 1 to 5). To do so, the Spanish SO-CAL dictionary (Taboada et al., 2011) was translated using the Elhuyar (Zerbitzuak, 2013) and Zehazki (Sarasola, 2005) bilingual Spanish-Basque dictionaries. Our dictionary contains information about grammatical categories: nouns, adjectives, verbs and adverbs.

Dictionary	Words	SO(-)	SO(+)
Nouns	2,882	1,635	1,247
Adjectives	3,162	1,733	1,429
Adverbs	652	225	427
Verbs	1,657	1,006	651
Total	**8,353**	**4,599**	**3,754**

Table 2: Characteristics of the Basque dictionary.

As Table (2) shows, the dictionary contains a total of 8,353 words. The majority of words are nouns and adjectives. In terms of polarity, there are more negative words (almost one thousand more).

We created a polarity tagger, based on this dictionary. The polarity tagger used the output of Eustagger (Aduriz et al., 2003), which is a robust and wide-coverage morphological analyzer and a Part-of-Speech tagger (POS) for Basque, to enrich the text with a POS analysis information and to assign polarity to every lemma of the dictionary that matches with the lemma and category of the text. With the aim of comparing the results of the system, a linguist annotated the polarity (positive, negative or neutral) of all the discourse structures described in Section (3.1).

Figure 1 shows a portion of the RST tree of one text (LIB28).[6] After the full RST analysis was performed for each text, we extracted the following discourse structures: i) the text of the central unit (EDU_2), as shown in Example (1), and ii) the central subconstituent of the EVALUATION ($EDU_{21,22,23,25}$), in Example (2).

(1) XIX. mendean Gasteiz inguruak izutu$_{(-3)}$ zituen Juan Diaz de Garaio Sacamantecas pertsonaia hartu$_{(+2)}$ du Aitor Aranak (Legazpi, 1963) bere azken eleberrian$_{(+2)}$. (LIB28_CU)
English: Aitor Arana (Legazpi, 1963) has taken$_{(+2)}$ in his last novel$_{(+2)}$ the character Juan Diaz Garaio Sacamantecas who scared$_{(-3)}$ the surroundings of Gasteiz in the 19th century.

(2) Hala ere, nahiko$_{(+2)}$ planoa da nobela$_{(+2)}$, erritmoa falta$_{(-1)}$ zaio eta bortxaketen kontaketak aspergarriak$_{(-3)}$ ere bihurtzen$_{(-2)}$ dira, Bestalde, alabaren ikuspuntua$_{(+2)}$ ez da batere argi geratzen$_{(-2)}$, (...). (LIB28_EVA)
English: However, the novel$_{(+2)}$ is fairly$_{(+2)}$ flat, it lacks$_{(-1)}$ rhythm, and the stories of rapes also become$_{(-2)}$ boring$_{(-3)}$. On the other hand, the point of view$_{(+2)}$ of the daughter is not clear$_{(-2)}$ (...)

The classifier then assigns polarity to each word in the dictionary, as shown in Table 3 and in examples (1) and (2). The table shows that the semantic

[5] All the reviews of the corpus were coded, assigning the domain LIB (for literature review) and a number, and each discourse structure extracted from them was also coded: CU stands for text that only contains the central unit of the text, CAUS for texts that contain CAUSE relation, INT for INTERPRETATION, ELAB for ELABORATION, CIR for CIRCUMSTANCE, BACK for BACKGROUND and finally, EVA for EVALUATION. In addition, if the same relation appears more than once in each text, we added letters (e.g., a, b, c) to each relation, to indicate their order of appearance.

[6] Size constraints prevent us from showing the entire tree.

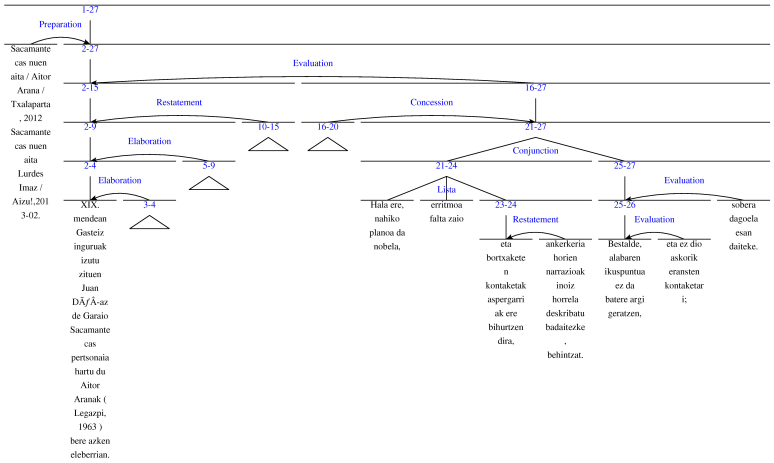


Figure 1: Central unit and the central subconstituent of EVALUATION in text LIB28

orientation of the central unit (LIB28_cu) is positive, while the semantic orientation of the central subconstituent (LIB28_EVA) is negative.

Ex.	CS ID	Classifier	SO	Manual
1	LIB28_cu	$-3+2+2$	$+1$	Neutral
2	LIB28_EVA	$+2+2-1-3-2$	-2	Negative

Table 3: Semantic orientation of LIB28_cu and LIB28_EVA: results of the classifier and of the manual annotation.

3.3 Normalization of semantic orientation results

We normalized the results obtained with the classifier to compare the different discourse structures, as in the following examples:

(3) Gure izaeraz$_{(+3)}$ hausnartzeko$_{(+1)}$ manual gisa eta, etxetik ibiltzeko$_{(+2)}$ dosi psikoanalitiko ttipi$_{(-1)}$ moduan$_{(+1)}$ hautematen$_{(+4)}$ dut nik. (LIB26_INT)
English: I consider$_{(+4)}$ it is like a manual with a small$_{(-1)}$ dose of psychoanalysis, a domestic$_{(+2)}$ consideration$_{(+1)}$ to reflect about$_{(+1)}$ our being$_{(+3)}$.

(4) Nolanahi$_{(-2)}$ den dela, saihestezina da gatazka$_{(-4)}$. (LIB13_CIR)
English: In any case$_{(-2)}$, the conflict$_{(-4)}$ is inevitable.

The results obtained by the classifier are $+112$ (LIB10),[7] $+10$ (LIB26_INT) and -6 (LIB23_CIR), as shown in Table 4. To compare those results among them, we normalized the frequencies dividing these results by the number of the words in each discourse structure. We show the normalized frequencies in Table 4.

Ex.	CS ID	SO	Words	NV
	LIB10	$+112$	418	$+0.27$
3	LIB26_INT	$+10$	17	$+0.59$
4	LIB13_CIR	-6	8	-0.75

Table 4: Examples of semantic orientation results after normalization (NV = Normalized Value).

Table 4 shows how normalization helps to better adjust the weight of the automatically assigned polarities. As a matter of fact, the values are adjusted

[7] Remember that this notation, LIB10, represents the entire text.

to a smaller range and, therefore, they are more easily comparable.

4 Results and error analysis

The results show that using a simple classifier with a manually built dictionary, along with different rhetorical structures, helps to identify the strength of such structures. For example, the result obtained in the central subconstituent of EVALUATION is strong.

(5) Guztiz$_{(+3)}$ gomendagarria$_{(+3)}$. (LIB26_EVA)
English: Highly$_{(+3)}$ recommended$_{(+3)}$.

(6) Liburu$_{(+5)}$ sano gomendagarria$_{(+3)}$ da, (LIB23d_EVA)
English: It's a very recommendable$_{(+3)}$ book$_{(+5)}$,

In Examples (5) and (6) the strength is higher than 1: $+2$ ($+6$ / 3 = 2) and $+1.6$ ($+8$ / 5 = $+1.6$), respectively, while the strength in other relations is lower.

(7) Izugarri$_{(+5)}$ gustura irakurri dut Bertol Arrietaren Alter ero narrazio bilduma. (LIB26_CAUS)
English: I have read very$_{(+5)}$ comfortable the Alter ero narration collection of Bertol Arrieta.

(8) Udako giro$_{(-2)}$ sapa horretan gertatzen diren kontakizun xumeak$_{(+3)}$ ekarriko dizkigu idazleak. (LIB15_CIR)
English: The writer will bring us the common$_{(+3)}$ stories that happen in that sticky atmosphere$_{(-2)}$ of summer.

The strength of CAUSE shows in Example (7) a value lower than 1 ($+5$ / 11 = $+0.45$). In Example (8) the central subconstituent of INTERPRETATION shows a value lower that 1 with a value of $+0.08$ ($+1$ / 12 = $+0.08$) and lower value than in Example (7).

We have analyzed the discourse structure with the aim of determining the strongest discourse structures of our corpus and therefore the structures that contribute most to improving sentiment labeling. Most of the values are between -1 and $+1$, but in 11.59% of the relations (8 of 69 relations), the values are higher than one (see Table 5).

RR	**Total**	**Total (<1)**	**%**
EVALUATION	32	6	18.75
INTERPRETATION	6	1	16.67
BACKGROUND	13	1	7.69
Others	18	0	0.00
Total	69	8	11.59

Table 5: Polarity strength ($< +1$ and > -1) of central subconstituents.

The most frequent and strongest value is obtained in EVALUATION (18.75%, 6 of 32). After that, the second strongest relation is INTERPRETATION with 16.67% (1 of 6). And, finally, BACKGROUND is once above one (7.69%, 1 of 13).

As examples (9, 10, 11, 12, 13) show, these relations have similar characteristics: short central subconstituents with many and strong evaluative words.

(9) berriz, zuzenean$_{(+3)}$ egin$_{(+2)}$ dut. (LIB14a_EVA)
English: whereas, I have done$_{(+2)}$ it directly$_{(+3)}$.

(10) Abentura$_{(+2)}$ liburu$_{(+5)}$ ederra$_{(+3)}$ iruditu$_{(+1)}$ zait, eta erremate paregabea$_{(+4)}$ trilogiarentzat. (LIB14b_EVA)
English: It seemed$_{(+2)}$ to me a beautiful$_{(+3)}$ adventure$_{(+2)}$ book$_{(+5)}$, and extraordinary$_{(+4)}$ finish for the trilogy.

(11) izenburua zuzen$_{(+3)}$ jarrita$_{(+1)}$, (LIB29a_EVA)
English: the title set$_{(+1)}$ correctly$_{(+3)}$,

(12) Intrigazko$_{(+2)}$ argumentua garatu$_{(+1)}$ nahi$_{(+3)}$ da. (LIB01b_EVA)
English: You want$_{(+3)}$ to develop$_{(+1)}$ an argument of intrigue$_{(+2)}$.

(13) Folklorean ikusi$_{(+4)}$ nahi$_{(+3)}$ ditu idazleak komunitate$_{(+1)}$ baten bizi$_{(+2)}$ nahi$_{(+3)}$ eta indarra$_{(+3)}$. (LIB35_INT)
English: The author wants$_{(+3)}$ to see$_{(+4)}$ in the folklore the strength$_{(+3)}$ and the desire$_{(+3)}$ to live$_{(+2)}$ of one community$_{(+1)}$.

Consequently, their value is higher than one, as shown in Table (6).

Ex.	**CS ID**	**NV**
9	LIB14a_EVA	1
10	LIB14b_EVA	1.36
11	LIB29a_EVA	1
12	LIB01b_EVA	1
13	LIB35_INT	1.33

Table 6: Central subconstituents and their value ($<+1$).

In contrast, we did not see any case of other central subconstituents with a value higher than one. If we compare partial discourse structures with the results obtained with all words of a text, the strength is lower in all cases. This is because polarity words do not have the same frequency in other rhetorical relations and, as a consequence, the concentration of words with semantic orientation is smaller. The highest value across the texts is $+0.50$ (LIB35), and the lowest value is -0.1 (LIB28).

These results suggest that opinions and, consequently, words with semantic orientation, are mainly found in the central subconstituent of EVALUATION, INTERPRETATION and BACKGROUND.

Apart from helping to identify the strongest central subconstituents, we have observed that the dictionary together with some central subconstituents can help in sentiment analysis. In fact, assigning a weight to some CSs could help to improve sentiment analysis results, as in text LIB34.

(14) "Behi eroak$_{(-3)}$" bilduman, ordea, egileak aurrekoan izan zituen arazoak$_{(-1)}$ konpondu$_{(+3)}$ ditu. Zoritxarrez$_{(-4)}$ bilduma honek batzuetan xelebrekeria$_{(-1)}$ merketik$_{(+3)}$ badu nahiko$_{(-2)}$. (LIB34b_EVA)
English: However, in "Behi eroak$_{(-3)}$" collection, the author has solved$_{(+3)}$ the problems$_{(-1)}$ that he had before. Unfortunately$_{(-4)}$, this collection has enough$_{(-2)}$ cheap$_{(+3)}$ eccentricity$_{(-1)}$.

The human annotator marked LIB34 as a negative review and the system assigns a value of $+0.15$ for the entire text, but a negative value of -0.2 (-5/25=-0.2) for LIB34b_EVA, Example (14). If the proper weight was assigned to this CS (LIB34b_EVA), the semantic positive orientation of the entire text (LIB34) would be corrected and tagged as negative.

We analyzed the previous finding in all the CSs of EVALUATION, but taking the results of the human annotator, instead of the classifier. In total, in 29 texts, there are 32 CSs of EVALUATION and in 24 of them, the human annotation of polarity of CSs and texts agree. So, the agreement happens in 75% of CSs and 86.20% of texts (25 texts).

Even though most of the times there is agreement between the annotated polarity of CSs and texts, this does not happen in all cases. For example, in other cases, the same text has one positive central subconstituent and another negative central subconstituent of EVALUATION. These cases are 12.50% of central subconstituents and 6.89% of texts (LIB03ab and LIB12ab).

Finally, there are two cases in which the polarity of the central subconstituent of EVALUATION and the polarity of all text are the opposite (LIB02ab and LIB19ab).

(15) eta apustu ausarta$_{(+3)}$ egin$_{(+2)}$ du bertan. (LIB19a_EVA)
English: and has made$_{(+2)}$ a strong$_{(+3)}$ bet there.

(16) Batetik, idazleak goi-literaturaren jokalekua hautatu duelako —liburuaren$_{(+5)}$ erlazio estratestualak eta baliatutako$_{(+1)}$ errekurtso andana$_{(-1)}$ lekuko—. Bestetik, borgestarretik asko duen jokoa$_{(-4)}$ delako liburuan$_{(+5)}$ dagoena. (LIB19b_EVA)
English: On the one hand, because the writer has chosen a scene from high literature —extratextual relations and a lot$_{(-1)}$ of resources used$_{(+1)}$ in the book$_{(+5)}$ as proof—. On the other hand, because there is a game$_{(-4)}$ that has a lot of Borges in the book$_{(+5)}$.

In this case, the text LIB19 is negative, whereas examples (15) and (16) are positive. We observe that the change of polarity happens in the EVALUATION situated inside an ELABORATION coherence relation.

(17) Baina, horiek horrela izanik ere, emaitza$_{(+1)}$ zalantzagarria$_{(-1)}$ da. Izan ere, liter-

aturan, baliabide$_{(+2)}$ orok medio izan behar$_{(-1)}$ du, eta irakurleak ikusi$_{(+4)}$ behar$_{(-1)}$ du errekurtsoak literaturaren mesedetan$_{(+3)}$ daudela "baita metaliteraturaz ari$_{(+2)}$ garenean ere". Hemen, ordea, medioak emaitza$_{(+1)}$ estaltzen$_{(-2)}$ du maiz$_{(+1)}$: literaturaren mekanismoekin egindako$_{(+2)}$ jokoek$_{(-4)}$ ipuinetan$_{(+2)}$ dauden istorioak$_{(-1)}$ indartu$_{(+1)}$ beharrean$_{(-1)}$, higatu$_{(-2)}$ egiten$_{(+2)}$ dituzte. Aldamioa oso$_{(+1)}$ nabarmena$_{(+4)}$ da, idazle askok beretzat nahi$_{(+3)}$ lukeen ahalmenez$_{(+2)}$ jasoa$_{(+2)}$. Haatik, hartatik sortzen$_{(+2)}$ den literatura ez da hain ikusgarria$_{(+4)}$. (LIB19_ELAB)
English: But, they being so, the result$_{(+1)}$ is doubtful$_{(-1)}$. In fact, in the literature, all resources$_{(+2)}$ need$_{(-1)}$ to be the medium, and the reader needs$_{(-1)}$ to see$_{(+4)}$ that resources are in favor$_{(+3)}$ of literary, "also when we are talking$_{(+2)}$ about metaliterature." But here, the medium hides$_{(-2)}$ the result$_{(+1)}$ in many times$_{(+1)}$: games$_{(-4)}$ made$_{(+2)}$ by literary devices wear away$_{(+2)(-2)}$ the tales$_{(-1)}$ of the stories$_{(+2)}$ instead$_{(-1)}$ of strengthening$_{(+1)}$ them. The scaffolding is very$_{(+1)}$ evident$_{(+4)}$, built$_{(+2)}$ with capacity$_{(+2)}$ as many writers would like$_{(+3)}$. However, the literature created$_{(+2)}$ is not very impressive$_{(+4)}$.

In Example (17), there are some discourse markers (*but, however*) and words (*doubtful, wear away, not very impressive*) that suggest a change of polarity that affects all text. Consequently, this example shows that, apart from central constituents of EVALUATION, a deeper analysis of nuclearity assigning different weighs could be necessary in order to improve sentiment analysis.

4.1 Error analysis

In this section, we will analyze the errors that can affect accurate detection of sentiment analysis, and specially the ones that were relevant in this study: i) errors in negative reviews, and ii) errors related to syntax.

4.1.1 Errors in negative reviews

Brooke et al. (2009) mention that lexicon-based sentiment classifiers show a positive bias because humans tend to use positive language (see also Taboada et al. (2017)). We also found this problem by examining the results of the classifier.

As Table (2) shows, the majority of the words in the dictionary are negative. Therefore, it is expected that we will detect more negative words in the texts. However, the results of the classifier with our dictionary show a tendency to classify texts as positive in different discourse structures of the texts.

For example, this tendency is observed in results of the CS of EVALUATION[8] (see Table 7).

CS of EVALUATION	Total	Guess	%
Positive	20	19	95.00
Negative	11	4	36.36
Neutral	1	0	0.00
Total	32	23	71.88

Table 7: Positive polarity tendency in central subconstituents of EVALUATION.

Table 7 demonstrates that the classifier tends to consider as positive the majority of central subconstituents of this rhetorical relation. In fact, 26 of 32 central subconstituents have been classified as positive. Consequently, the correct guess rate in CSs is higher in positive (95%) versus negative (36.36%).

A tendency to positive semantic orientation is higher if we analyze the results of all texts instead of just central subconstituents of EVALUATION as shown in Table 8.

Texts	Total	Guess	%
Positive	14	14	100
Negative	15	1	6.67
Total	29	15	51.72

Table 8: Positive polarity tendency in texts of the corpus.

As a consequence of this positive bias, our classifier guesses easily the texts with positive polarity and the correct guess rate is 100%. In contrast, the rate is very low in negative texts, as a matter of fact, there is only one right guess in text LIB28 (-0.1) and consequently, the correct guess rate is 6.67%.

[8]We have analyzed this relation and not others because it accounts for almost half of all the studied rhetorical relations.

However, if we compare the results of central subconstituents and texts, we can observe another tendency. The rate of correct assignments in positive texts is higher (95% vs. 100%) on the full texts (long text), while for negatives it is higher (36.36% vs. 6.67%) in central subconstituents (short text). This suggests that the tendency to positive semantic orientation is stronger using our dictionary as a bag-of-words approach as the text is longer.

In summary, the dictionary classifier shows the same problem already described in previous research, as there is a strong tendency towards positive semantic orientation, which increases as the text is longer.

4.1.2 Errors related to syntax

As we mentioned in Section 4.1.1, there is a tendency towards positive polarity caused by the use of positive language and, for that reason, the correct guess rate is lower in negative texts. However, it is not the only reason, and information at the syntactic level also affects the results. As an example, we will discuss one particular problem, negation. Due to negation, the polarity of a sentence is changed and it is necessary to take this characteristic into account in sentiment analysis.

(18) (...) narrazioak ere ez du arretarik bereganatzen$_{(+4)}$ (...) (LIB18_EVA).
English: (...) the narration also does not get attention$_{(+4)}$ (...)

In Example (18), the semantic orientation of the sentence would be negative but our classifier regards it as positive. The classifier has detected *bereganatu* 'to get hold of' as a positive word ($+4/7=+0.57$). But, in this case, a correct analysis should assign it a negative value.

In a first study of our subcorpus of CSs of different rhetorical relations, we estimate that this affects to 11.43% of the constituents, since 8 of 70 CSs have some type of negation.

5 Conclusions and future work

This study has analyzed whether combining a semantic oriented dictionary with some discourse structure constraints is helpful in sentiment analysis of Basque.

The results show that i) the central subconstituents (CS) of EVALUATION, INTERPRETATION and BACKGROUND are the units with the strongest semantic orientation, and ii) the CSs of EVALUATION could help in improving semantic orientation of the texts, given that the results of the human annotation of polarity of CSs and the full text text agree in 75% of the cases.

On the other hand, error analysis has shown that there are some aspects that should be addressed: i) a tendency to positive semantic orientation, and ii) sentence and more discourse level constraints are needed.

In the near future, we plan to pursue the following aspects:

i) Do reviews have a specific discourse structure? We hypothesize that reviews have a specific structure and, consequently, the same discourse relations will be repeated with high frequency, and they will appear in the same place.

ii) How we can weigh properly the central subconstituents of EVALUATION and INTERPRETATION, and neutralize the positive tendency, to improve the results for negative reviews?

iii) Are other CSs not linked to the CU important for sentiment analysis?

Acknowledgments

We thank Arantxa Otegi for assistance with the lexicon-based polarity tagger. Jon Alkorta's work is funded by a PhD grant (PRE_2016_2_0153) from the Basque Government and Mikel Iruskieta's work is funded by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish *Ministerio de Economia, Comercio y Competitividad*.

References

[Aduriz et al.2003] Itziar Aduriz, Izaskun Aldezabal, Inaki Alegria, J Arriola, Arantza Dıaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite state applications for basque. In *EACL2003 Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

[Alkorta et al.2015] Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, and Alicia Prez. 2015. Using rela-

tional discourse structure information in basque sentiment analysis. In *SEPLN 5th Workshop RST and Discourse Studies. ISBN: 978-84-608-1989-9. https://gplsi.dlsi.ua.es/sepln15/en/node/63*.

[Alkorta et al.2016] Jon Alkorta, Koldo Gojenola, and MIkel Iruskieta. 2016. Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela , September 29 th - 30 th . Extended Abstracts. ISBN: 978 - 84 - 608 - 9305 - 9*.

[Bhatia et al.2015] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

[Brooke et al.2009] Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50–54.

[Chardon et al.2013] Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–37. Springer.

[Hu and Liu2004] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Iruskieta et al.2015] Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

[Iruskieta2014] Mikel Iruskieta. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea*.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Mann and Thompson1988] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[Polanyi and Zaenen2006] Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

[Sarasola2005] Ibon Sarasola. 2005. *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

[Taboada et al.2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

[Taboada et al.2017] Maite Taboada, Radoslava Trnavac, and Cliff Goddard. 2017. On being negative. *Corpus Pragmatics*, 1(1):57–76.

[Taboada2016] Maite Taboada. 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

[Trnavac et al.2016] Radoslava Trnavac, Debopam Das, and Maite Taboada. 2016. Discourse relations and evaluation. *Corpora*, 11(2):169–190.

[Turney2002] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

[Vicente et al.2017] Inaki San Vicente, Rodrigo Agerri, and German Rigau. 2017. Q-wordnet ppv: Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *arXiv preprint arXiv:1702.01711*.

[Wiebe2000] Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

[Wu and Qiu2012] Fei Wang1 Yunfang Wu and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. In *24th International Conference on Computational Linguistics*, page 1311.

[Zerbitzuak2013] Elhuyar Hizkuntza Zerbitzuak. 2013. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. usurbil: Elhuyar.

[Zhou et al.2011] Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.

4.4 Article 6

Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus

Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus

Jon Alkorta Ixa Group / UPV/EHU
Koldo Gojenola Ixa Group / UPV/EHU
Mikel Iruskieta Ixa Group / UPV/EHU
{jon.alkorta, koldo.gojenola, mikel.iruskieta}@ehu.eus

Abstract

Discourse information is crucial for a better understanding of the text structure and it is also necessary to describe which part of an opinionated text is more relevant or to decide how a text span can change the polarity (strengthen or weaken) of other span by means of coherence relations. This work presents the first results on the annotation of the Basque Opinion Corpus using Rhetorical Structure Theory (RST). Our evaluation results and analysis show us the main avenues to improve on a future annotation process. We have also extracted the subjectivity of several rhetorical relations and the results show the effect of sentiment words in relations and the influence of each relation in the semantic orientation value.

1 Introduction

Sentiment analysis is a task that extracts subjective information for texts. There are different objectives and challenges in sentiment analysis: i) document level sentiment classification, that determines whether an evaluation is positive or negative (Pang et al., 2002; Turney, 2002); ii) subjectivity classification at sentence level which determines if one sentence has subjective or objective (factual) information (Wiebe et al., 1999) and iii) aspect and entity level in which the target of one positive or negative opinion is identified (Hu and Liu, 2004).

In order to attain those objectives, some resources and tools are needed. Apart from basic resources as a sentiment lexicon, a corpus with subjective information for sentiment analysis is indispensable. Moreover, such corpora are necessary for two approaches to sentiment analysis. One approach is based on linguistic knowledge, where a corpus is needed to analyze different linguistic phenomena related to sentiment analysis. The second approach is based on statistics and, in this case, the corpus is useful to extract patterns of different linguistic phenomena.

The aim of this work is to annotate the rhetorical structure of an opinionated corpus in Basque to check out the semantic orientation of rhetorical relations. This annotation was performed following the *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988). We have used the Basque version of SO-CAL tool to analyze the semantic orientation of this corpus (Taboada et al., 2011).

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the theoretical framework, the corpus for study and the methodology of annotation as well as the analysis of the corpus carried out. Then, Section 4 explains the results of the annotation process, the inter-annotator agreement and the results with regard to analysis in the subjectivity of the corpus. After that, Section 5 discusses the results. Finally, Section 6 concludes the paper, also proposing directions for future work.

2 Related work

The creation of a specific corpus and its annotation at different linguistic levels has been very a common task in natural language processing. As far as a corpus for sentiment analysis is concerned, information related to subjectivity and different grammar-levels has been annotated in different projects.

Refaee and Rieser (2014) annotate the Arabic Twitter Corpus for subjectivity and sentiment analysis. They collect 8,868 tweets in Arabic by random search. Two native speakers of Arabic annotated the tweets. On the one hand, they annotate the semantic orientation of each tweet. On the other hand, they also annotate different grammatical characteristics of tweets such as syntactic, morphological and semantic features as well

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 144–152
Minneapolis, MN, June 6, 2019.

as stylistic and social features. They do not annotate any discourse related feature. They obtain a Kappa inter-annotator agreement of 0.84.

The majority of corpora for sentiment analysis are annotated with subjectivity information. There are fewer corpora annotated with discourse information for the same task. Chardon et al. (2013) present a corpus for sentiment analysis annotated with discourse information. They annotate the corpus using Segmented Discourse Representation Theory (SDRT), creating two corpora: i) movie reviews from *AlloCinéf.fr* and ii) news reaction from *Lemonde.fr*. They collect 211 texts, annotated at EDU and document level. At the EDU level, subjectivity is annotated while at the document level, subjectivity and discourse relations are annotated. Results in subjectivity show that, at EDU level, Cohen's Kappa varies between 0.69 and 0.44 depending on the corpus and, at the document level, Kappa is between 0.73 and 0.58, respectively. They do not give results regarding the annotation of discourse relations.

Asher et al. (2009) create a corpus with discourse and subjectivity annotation. They categorize opinions in four groups (REPORTING, JUDGMENT, ADVISE and SENTIMENT), using SDRT as the annotation framework for discourse. Exactly, they use five types of rhetorical relations (CONTRAST/CORRECTION, EXPLANATION, RESULT and CONTINUATION). They collect three corpora (movie reviews, letters and news reports) in English and French. 150 texts are in French and 186 texts in English. According to Kappa measure, in opinion categorization, the inter-annotator agreement is 95% while in discourse segmentation it is 82%.

Mittal et al. (2013) follow a similar methodology. By the annotation of negation and discourse relations in a corpus, they measure the improvement made in sentiment classification. They collect 662 reviews in Hindi from review websites (380 with a positive opinion and 282 with a negative one). Regarding discourse, they annotate violating expectation conjunctions that oppose or refute the current discourse segment. According to their results, after implementing negation and discourse information to HindiSentiWordNet (HSWN), the accuracy of the tool increases from 50.45 to 80.21. They do not mention the inter-annotating agreement of violating expectation conjunctions.

To sum up, this section gives us a general overview about discourse-based annotated corpora for sentiment analysis. Corpora have been made for specific aims, annotating only some characteristics or features related to discourse and discourse relations. This situation differs from our work, because our work describes the annotation process of the relational discourse structure and how the function in the rhetorical relation affect to the analysis in the semantic orientation.

3 Theoretical framework and methodology

3.1 Theoretical framework: Rhetorical Structure Theory

We have annotated the opinion text corpus using the principles of *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988; Taboada and Mann, 2006), as it is the most used framework in the annotation of discourse structure and coherence relations in Basque where there are some tools (Iruskieta et al., 2013, 2015b) to study rhetorical relations. According to this framework, a text is coherent when it can be represented in one discourse-tree (RS-tree). In a discourse-tree, there are elementary discourse units (EDU) that are interrelated. The relations are called *coherence relations* and the sum of these coherence relations forms a discourse-tree. Moreover, the text spans present in a discourse relation may enter into new relations, so relations can form compound and recursive structures.

Elementary discourse units are text spans that usually contain a verb, except in some specific situations. The union of two or more EDUs creates a coherence relation. There are initially 25 types of coherence relations in RST. In some cases, one EDU is more important than other one and, in this case, the most important EDU in the relation is called *nucleus*-unit (basic information) while the less important or the auxiliary EDU is called *satellite*-unit (additional information). Coherence relations of this type are called *hypotactic relations*. In contrast, in other relations, EDUs have the same importance and, consequently, all of them are nucleus. The relations with EDUs of same rank are called *paratactic relations*. The task that selects the nucleus in a relation is called *nuclearity*.

Hypotactic relations are also divided into two groups according to their effect on the reader. Some relations are *subject matter* and they are re-

lated to the content of text spans. For example, CAUSE, CONDITION or SUMMARY are subject matter relations. On the other hand, the aim of other relations is to create some effect on the reader. They are more rhetorical in their way of functioning. EVIDENCE, ANTITHESIS or MOTIVATION belong to this group.

Figure 1 presents a partial discourse-tree of an opinion text (tagged with the code LIB29). The text is segmented and each text span is a discourse unit (EDU). The discourse units are linked by different types of rhetorical relations. For instance, the EDUs numbered with 15 and 16 are linked by an ELABORATION relation and the EDUs ranging from 15 to 20 are linked by LIST (multinuclear relation). On the other hand, the EDU numbered 2 is the central unit of this text because other relations in the text are linked to it and this text span is not attached to another one (with the exception of multinuclear relations).

According to Taboada and Stede (2009), there are three steps in RST-based text annotation:

1- Segmentation of the text in text spans. Spans are usually clauses.

2- Examination of clear relations between the units. If there is a clear relation, then mark it. If not, the unit belongs to a higher-level relation. In other words, the text span is part of a larger unit.

3- Continue linking the relations until all the EDUs belong to one relation.

Following Iruskieta et al. (2014) we think that it is recommendable, after segmenting the corpus, to identify first the central unit, and then mark the relations between different text spans.

3.2 The Basque Opinion Corpus

The corpus used for this study is the *Basque Opinion Corpus* (Alkorta et al., 2016). This corpus has been created with 240 opinion texts collected from different websites. Some of them are newspapers (for instance, Berria and Argia) while others are specialized websites (for example, Zinea for movies and Kritiken Hemeroteka for literature).

The corpus is multidomain and, in total, there are opinion texts of six different domains: sports, politics, music, movies, literature books and weather. The corpus is doubly balanced. That is, each domain has the same quantity of opinion texts (40 per domain) and each semantic orientation (positive or negative subjectivity) has the same quantity of opinion texts per each domain (20 positive and 20 negative texts per domain). We extract preliminary corpus information using the morphosyntactical analysis tool Analhitza (Otegi et al., 2017): 52,092 tokens and 3,711 sentences.

We made preliminary checks to decide whether the corpus is useful for sentiment analysis. The opinion texts are subjective, so the frequency information of the first person should be high. The results show that the first person appearance is of 1.21% in a Basque objective corpus (Basque Wikipedia) whereas its appearance is of 8.37% in the Basque Opinion Corpus. As far as the presence of adjectives is concerned, both corpora show similar results. From all the types of grammatical categories, 8.50% of the words correspond to adjectives in Basque Wikipedia and 9.82% in the corpus for study. Other interesting features for sentiment analysis, such as negation, *irrealis blocking* and discourse markers, have also been found in the corpus.

3.3 Methodological steps

We have followed several steps to annotate the Basque Opinion Corpus using the RST framework:

	A1	**A2**	**Total**
Movie	21 + 9	9	30
Weather	10 + 5	5	15
Literature	5	20 + 5	25
Total	50	39	70

Table 1: Number of texts annotated by two annotators. The number after the sum sign indicates the quantity of texts with double annotation.

1- Limiting the annotating work. Annotating 240 texts needs a lot of work and time. For that reason, we have thought to annotate some part of the corpus initially and, if the results of the annotation are acceptable, continue with the work. Taking into account the previously described data, both annotators have worked with 70 texts (29.16%) of three different domains. 21 texts from the movie domain have been annotated by one annotator and other 9 texts have been annotated by the two annotators. 10 texts from weather have been annotated once and other 5 texts of the same domain by two annotators. Finally, 25 texts

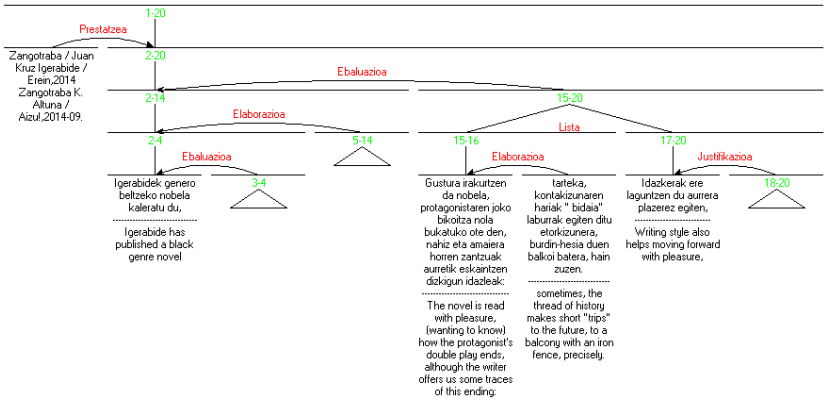


Figure 1: Part of a discourse-tree of the LIB29 review annotated with the RST framework.

of literature reviews have been annotated by one annotator and other 5 texts from the same domain by two. In total, 19 texts from 70 (27.14%) have been annotated by two annotators.

2- Annotation procedure and process. We decided to follow the annotation guidelines proposed by Das and Taboada (2018). Each person annotated four or five texts per day during two or three weeks. The time to annotate documents varied according to the domain. The texts corresponding to the weather domain are shorter and, consequently, easier to annotate while texts about movies as well as those of the literature domain are more difficult because their writing style is more implicit (less indicators and relation signals) and complex (longer at least). Approximately, each weather text was annotated in 20 minutes while movie and literature texts were annotated in one hour.

3- Measurement of inter-annotator agreement. In order to check the quality of the annotation process, inter-annotator agreement was measured. This was calculated manually following the qualitative evaluation method (Iruskieta et al., 2015a) using F-measure. In this measurement, in contrast with the automatic tool, the central subconstituent factor was not taken into account.

4- Semantic orientation extraction. Using the Basque version of the SO-CAL tool (Taboada et al., 2011), we have extracted the subjective information of rhetorical relations in the three domains of the corpus in order to check how the type of rhetorical relation affects their sentiment valence. SO-CAL needs a sentiment lexicon where words have a sentiment valence between -5 and $+5$. The Basque version of the sentiment lexicon contains 1,237 entries.

We have extracted the sentiment valence of 75 instances if CONCESSION and EVALUATION relations. From the 75 CONCESSION relations, 16 come from the weather domain, 34 from literature and 25 from movies. In the case of EVALUATION, 19 come from weather, 31 from literature and 25 from weather.

5- Results. On the one hand, we have calculated the percentage of rhetorical relations with the same label annotated by two persons. On the other hand, we have measured accumulated values of sentiment valences in nuclei and satellites in texts of different domains.

4 Results

4.1 Inter-annotator agreement

Table 2 shows the inter-annotator agreement of rhetorical relations (RR) between both annotators. This agreement was calculated following the qualitative method (Iruskieta et al., 2015a). According to these results, the highest agreement has been reached in the domain of weather where 17 of 39 relations (43.59%) have been annotated with the same relation label. After that, inter-annotator agreement in literature is 41.67% (70 from 168). Finally, the domain of movies obtained the lowest results, since the agreement is 37.73% (83 of

220). Taking all domains into account, 39.81% of the rhetorical relations have been annotated in the same way (170 relations of 427). The disagreements are due to different reasons: i) both annotators have to train more to reach a higher agreement and to obtain better results. ii) opinionative texts are more open than news or scientific abstracts. Therefore, there is more place for different interpretations.

Domain	Agreement (%)	Agreement (RR)
Weather	43.59	17 of 39
Literature	41.67	70 of 168
Movies	37.73	83 of 220
Total	**39.81**	**170 of 427**

Table 2: Inter-annotator agreement in different domains of the corpus measured by hand.

4.2 Subjectivity extraction from rhetorical relations

The annotation of the corpus using Rhetorical Structure Theory allows us to check the usefulness of the corpus. We have extracted the subjectivity from different types of rhetorical relations using the Basque version of the SO-CAL tool and we have been able to check the distribution of words with sentiment valence in each type of rhetorical relation and domain.

We have analyzed how words with sentiment valence appear in nuclei as well as satellites of CONCESSION and EVALUATION[1] in three domains. The results[2] are presented in Table 3. In the case of CONCESSION, the presence of words with sentiment valence in nuclei (47.21%) and satellites (52.79%) is similar in the three domains, although satellites show a higher proportion. In contrast, in the case of EVALUATION, words with sentiment valence are more concentrated on satellites (55.00%) in comparison with nuclei (45.00%). The only exception is weather, where nucleus prevail over satellites as far as the concentration of words with sentiment valence is concerned[3].

This information contrast between discourse and sentiment analysis provides us the option to understand what happens there. For example, in **CONCESSION**, the nucleus presents a situation affirmed by the author and the satellite shows a situation which is apparently inconsistent but also affirmed by the author (Mann and Taboada, 2005). In other words, the probability of an opinion appearance is similar in both. The sentiment valence of the nucleus prevails over the satellite but the application of Basque SO-CAL does not give the correct result because the tool does not apply any discourse processing and, consequently, in this CONCESSION relation, nuclei as well as satellite are given the same weight.

(1) [S[Puntu ahulak izan arren,]$_{-1.5}$ N[film erakargarri eta berezia da Victoria.]$_{+6}$]$_{+4.5}$ (ZIN19)
[S[Although it has weak points,]$_{-1.5}$ N[Victoria is an entertaining and special movie.]$_{+6}$]$_{+4.5}$

(2) [N[Joxek emaztea eta lagunak dauzka,]$_{-1.5}$ S[gaizki tratatzen baditu ere.]$_{-4.5}$]$_{-2.5}$ (SENTAIZ02)
[N[Joxe has a wife and friends,]$_{+2}$ S[although he treats them badly]$_{-4.5}$]$_{-2.5}$

(3) [S[Eta Redmaynen lana oso ona bada ere,]$_{+1}$ N[Vikanderrena bikaina da.]$_{+5}$]$_{+6}$ (ZIN15)
[S[Although Redmayn's work is very good]$_{+1}$, N[Vikander's is excellent.]$_{+5}$]$_{+6}$

In Example (1), the semantic orientation of the nucleus is positive while the semantic orientation of the satellite is negative. The sum is positive and, in this case, SO-CAL correctly assigns the semantic orientation of the overall rhetorical relation. In contrast, in Example (2), according to SO-CAL, the sentiment orientation of the relation is negative but it should be positive, because the semantic orientation of the nucleus is positive. This example clarifies how discourse information is needed in lexicon-based sentiment classifiers such as SO-CAL. Finally, in Example (3), the nucleus as well as the satellite and the rhetorical relation have positive semantic orientation and SO-CAL assigns correctly the semantic orientation.

Another type of rhetorical relation is **EVALUATION**, where the satellite makes an evaluative comment about the situation presented in the nucleus (Mann and Taboada, 2005). That means that the words with subjective information are more likely to appear in the satellite.

[1] We decide to choose these rhetorical relations, because we think they are more related to opinions and emotions.

[2] In order to measure the presence of words with subjectivity, we have calculated the sum of all the sentiment valences without taking into account their sign.

[3] In the weather domain, one of rhetorical relations has a very long nucleus compared to satellite. This situation may have influenced the results. In other cases, the length of nucleus and satellites has been similar.

Sum of sentiment valences	CONCESSION		EVALUATION	
	Nucleus	**Satellite**	**Nucleus**	**Satellite**
Weather	39.41	39.75	49.86	33.35
Literature	61.02	68.73	53.13	80.30
Movies	13.98	19.45	26.01	45.58
Total	114.41 (47.21 %)	127.93 (52.79 %)	128.99 (45.00%)	159.23 (55.00%)

Table 3: Accumulated values of sentiment valences in nuclei and satellites for each domain.

(4) [N[Arrate Mardarasek bere lehen liburua argitaratu du berriki, Pendrive,]$_0$ S[eta apustu ausarta egin du bertan.]$_{+3}$]$_{+3}$ (SENTBER04)
[N[Arrate Mardaras has published her first book recently, Pendrive,]$_0$ S[and she has made a daring bet there.]$_{+3}$]$_{+3}$

(5) [N[Bada, erraz ikusten den filma da "The danish girl".]$_{+1}$ S[Atsegina da, hunkigarria, entretenigarria]$_{+6}$]$_{+7}$ (ZIN15).
[N[So, "The danish girl" is a film easy to watch.]$_{+1}$ S[It is nice, touching, entertaining.]$_{+6}$]$_{+7}$

(6) [N[Talde lana izatetik pertsonaia bakarraren epika izatera pasako da erdialdetik aurrera]$_{+0.5}$ S[eta horretan asko galduko du filmak.]$_{-3.9}$]$_{-3.4}$ (ZIN39)
[N[It is going to pass from being team work to epic of one person]$_{+0.5}$ S[and in that, the film will lose a lot.]$_{-3.9}$]$_{-3.4}$

Here, we can see some specific characteristics of each rhetorical relation. Unlike CONCESSION, there is a concentration of words with sentiment valence in the satellite while words with sentiment valence have little presence in the nucleus. In fact, the sentiment valence of nuclei is never higher than $+1$ whereas satellites have a higher sentiment valence than ± 3 in all the cases. In these three Examples (4, 5 and 6), the Basque version of the SO-CAL tool guesses correctly the semantic orientation of rhetorical relations. For example, in Example (6), the semantic orientation of nucleus is positive and of satellite is negative. The sum of the two EDUs is negative and SO-CAL correctly assigns a -3.4 sentiment valence. This does not happen in all cases because the tool has not implemented any type of discourse information processing. Anyway, the tool provides information about semantic orientation that is necessary to study the relation between sentiment analysis and rhetorical relations.

5 Discussion

5.1 Inter-annotator agreement

Regarding inter-annotator agreement (Table 2), the agreement goes from 37.73% to 43.59%. However, some domains do not show regularity regarding agreement. For example, in the case of reviews (domain of literature), inter-annotator agreement is situated between 38% and 48%, except in two texts where the agreement is lower (26% and 30%). In the same line, in the weather domain, some texts show higher agreement than the average in the domain.

If we evaluate this doubly annotated corpus by automatic means in a more strict scenario (if and only if the central subconstituent is the same) following Iruskieta et al. (2015a), we can observe and evaluate other aspects of rhetorical structure, such as:

- **Constituent (C)** describes all the EDUs that compose each discourse unit or span.
- **Attachment point** is the node in the RS-tree to which the relation is attached.
- **N-S** or nuclearity specifies if the compared relations share the same direction (NS, NS or NN).
- **Relation** determines if both annotators have assigned[4] the same type of rhetorical relation to the attachment point of two or more EDUs in order to get the same effect.

Another aspect to take into consideration is that the manual and automatic evaluation does not show the same results with regard to inter-annotator agreement of the type of relation. According to a manual evaluation, inter-annotator

[4]If the central subconstituent is not described with the same span label and compared position (NS or SN), there is no possibility of comparing relations.

	Constituent		Attachment		N-S		Relation	
Domain	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**	**Match**	**F1**
Weather	20 of 37	0.54	9 of 37	0.24	22 of 37	0.59	15 of 37	0.41
Literature	84 of 155	0.54	67 of 155	0.43	105 of 155	0.68	48 of 155	0.31
Movies	112 of 221	0.56	88 of 221	0.40	147 of 221	0.67	68 of 221	0.31
Total	**216 of 413**	**0.52**	**164 of 413**	**0.40**	**274 of 413**	**0.66**	**131 of 413**	**0.32**

Table 4: Inter-annotator agreement results given by the automatic tool.

agreement is 39.81% while the automatic evaluation shows an agreement of 31.72%. As we have noted before, this difference comes due to the fact that the automatic comparison is made in a strict scenario and some relations are not compared, because the description of the central subconstituent of such relations is slightly different.

The inter-annotator agreement results given by the automatic tool offer complementary information related to the annotation of the corpus. As Table 4 shows, the inter-annotator agreement is low in the case of type of relation but the results are better in other aspects of rhetorical relations such as constituent and nuclearity. The agreement in attachment point achieves 0.40 that is low still but constituent as well as nuclearity have achieved the inter-annotator agreement of 0.52 and 0.66, respectively.

On the other hand, another interesting aspect is that there is no difference between domains as far as the agreement of different aspects related to writing style is concerned. It is surprising because the type and the way to express opinions are very different for each domain. In the weather domain, texts are short and clear and the language is direct. In contrast, in literature and movies, texts are longer, more diffuse and they use figurative expression many times. Even so, the weather domain obtains lowest results in three aspects mentioned in Table 4 but the type of relation obtains a better result compared to other domains.

The interpretation of inter-annotator agreement suggests that in the evaluation of some rhetorical relations the agreement is lower while other aspects related to rhetorical relations like constituent and nuclearity obtain a better agreement. We have also discovered that specially ELABORATION, EVALUATION and some multinuclear relations show higher disagreement.

5.1.1 Relevant RR disagreement: confusion matrix

In order to know the differences of these disagreements, we have also measured the type of rhetorical relations with the highest disagreement. With that aim, we have calculated a confusion matrix, and then we have identified the most controversial rhetorical relations. Results are shown in Table 5.

A1	**A2**		
RRs		**#**	**Total**
ELABORATION	MOTIVATION	9	19
ELABORATION	INTERPRETATION	6	
RESULT	ELABORATION	4	
INTERPRETATION	JUSTIFICATION	4	4
CONCESSION	CONTRAST	6	14
EVALUATION	CONTRAST	4	
LIST	CONJUNCTION	4	

Table 5: Disagreement in rhetorical relations.

According to Table 5, ELABORATION has been used by one annotator whereas the other has employed a more informative relation. In two cases, the first annotator (A1) has annotated an EVALUATION relation while the other annotator (A2) has annotated MOTIVATION and INTERPRETATION. In other case, A2 has annotated ELABORATION whereas A1 has tagged RESULT. In total, there are 19 instances in which ELABORATION has been annotated by one of the annotators. Moreover, there are 4 instances of disagreement between INTERPRETATION and JUSTIFICATION. Finally, there are also disagreements in multinuclear relations. While A2 has annotated CONTRAST in 10 relations, A1 has employed CONCESSION and EVALUATION. There are also 4 instances of disagreement between LIST and CONJUNCTION.

Our interpretation of this results is that one annotator (A1) tends to annotate more general rhetorical relations (e. g. ELABORATION) while other annotator (A2) annotates more precise relations. When it comes to multinuclear relations, it seems that A1 annotator has a tendency to not an-

notate multinuclear relations.

5.2 Checking the usefulness of the corpus for sentiment analysis

The second aim of this work has been to check the usefulness of the corpus for sentiment analysis. Firstly, the results have shown that in some cases the Basque version of SO-CAL does not assign a suitable semantic orientation to all the rhetorical relations, even when the semantic orientation of EDUs of the relation is correct. This means that the information of rhetorical relations would be needed in order to make a lexicon-based sentiment classification. In other words, this suggests that it would be recommendable to assign weights to EDUs of rhetorical relations to model their effect on sentiment analysis. Each type of rhetorical relation has different characteristics and, consequently, the way to assign weights to EDUs in each relation must be different.

For that reason, we have made a preliminary study with the purpose of checking how different types of rhetorical relations present a semantic orientation and what is the distribution of words with sentiment valence in rhetorical relations. The study of CONCESSION has shown that *i)* the probability of sentiment words appearing in nuclei as well as satellites is similar, and that *ii)* nucleus always prevails over the satellite and, consequently, the semantic orientation of nucleus must be the semantic orientation of all the rhetorical relation. However, the semantic orientation of the satellite must be also taken into consideration in the semantic orientation of all the rhetorical relation. Although comparing with nucleus, satellite has to be less important.

The opposite situation happens in EVALUATION. Here, we can see that words with sentiment valence concentrate more on the satellite while there are fewer words with sentiment valence in the nucleus. That means that the weight must be assigned to the satellite because that part of the relation is more important from the point of view of sentiment analysis.

This interpretation of the results suggests that the Basque Opinion Corpus annotated using RST can be useful for different tasks of sentiment analysis, in fact, the preliminary analysis made with rhetorical relations shows some characteristics and differences that are related to rhetorical relations.

6 Conclusion and Future Work

In this work, we have annotated a part of the Basque Opinion Corpus using Rhetorical Structure Theory. Then, we have measured inter-annotator agreement. The manual evaluation of the results shows that the inter-annotator agreement of the type of rhetorical relations is 39.81%. On the other hand, using an automatic tool we have obtained more fine-grained results regarding aspects of relations and attachment, as well as nuclearity, with an inter-annotator agreement higher than 0.5. We have also identified that ELABORATION, EVALUATION and some multinuclear relations show the highest disagreement.

On the other hand, we have also checked the usefulness of this annotated corpus for sentiment analysis and the first results show that it is useful to extract subjectivity information of different rhetorical relations. In CONCESSION relations, the semantic orientation of the nucleus always prevails but the valence of the satellite must also be taken into consideration. In EVALUATION relations, words with sentiment valence concentrate on satellite.

In future, firstly, we plan to build extended annotation guidelines to annotate the corpus with more reliability. This would be the previous step before annotating the entire corpus. On the other hand, we would like to continue analyzing how the subjective information is distributed in relations.

Acknowledgments

This research is partially supported by a Basque Government scholarship (PRE_2018_2_0033), the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) project PROSA-MED (TIN2016-77820-C3-1-R), University of the Basque Country project UPV/EHU IXA Group (GIU16/16) and project Procesamiento automático de textos basado en arquitecturas avanzadas (PES18/28).

References

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. 2016. Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis*, pages 58–62.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of opinion expressions in

discourse. *Lingvisticæ Investigationes*, 32(2):279–292.

Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–37. Springer.

Debopam Das and Maite Taboada. 2018. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Mikel Iruskieta, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015a. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Mikel Iruskieta, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 466–475.

Mikel Iruskieta, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2015b. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 11(2):303–334.

William C Mann and Maite Taboada. 2005. RST web site.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. 2013. Sentiment Analysis of Hindi Reviews based on Negation and Discourse Relation. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 45–50.

Arantxa Otegi, Oier Imaz, Arantza Diaz de Ilarraza, Mikel Iruskieta, and Larraitz Uria. 2017. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis. In *LREC*, pages 2268–2273.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Maite Taboada and Manfred Stede. 2009. Introduction to RST (Rhetorical Structure Theory). *ESSLLI2016*.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 246–253.

4.5 Article 7

The role of hierarchical structure and coherence relations in sentiment analysis: a study in Basque (preprinted)

The role of hierarchical structure and coherence relations in sentiment analysis: A study in Basque

Jon Alkorta
UPV/EHU

Maite Taboada
Simon Fraser University

Koldo Gojenola
UPV/EHU

Mikel Iruskieta
UPV/EHU

Sentiment analysis is now a mature area of study, with a number of tools and dictionaries in multiple languages. The goal of most automated systems for sentiment analysis is to determine whether a text (broadly construed, i.e., a review, a headline, a tweet) expresses positive or negative opinion towards its subject matter. The majority of the existing sentiment analysis systems follow a bag-of-words approach and, until now, syntactic and discourse constraints have been excluded, because it has been assumed that this information was not necessary or it was too complex to process adequately. We believe that contextual discourse information is crucial to accurate sentiment analysis and, to this end, the main aim of this work is to study the most relevant features of discourse relations vis-à-vis sentiment analysis. In order to add discourse features, we need to know how relational discourse structure phenomena affect or modify the polarity of text spans (what we define as semantic orientation of text spans), and then add this information to a sentiment analysis system.
Specifically, we identify the discourse structures that have a role in the interpretation of sentiment and opinion in text. We will base our study on Basque, a language that has not been widely addressed in sentiment analysis. We have annotated a corpus of book reviews in Basque with two different types of features: a) the relational structure of texts following Rhetorical Structure Theory (RST); and b) the semantic orientation (positive or negative polarity) of the annotated discourse relations. We analyze how the semantic orientation of such discourse relations interacts with different aspects of the relation (relation type, nuclearity) and with the overall text structure (position, distance from the text's central unit). The results suggest that there is evaluative prosody at the local level of discourse structure: The semantic orientation of a rhetorical relation tends to match that of neighbouring nuclei (as opposed to satellites) and it also tends to match the semantic orientation of the following spans (as opposed to previous spans).
Moreover, discourse relations near the central unit match the semantic orientation of the text overall more often than those that are farther away from the central unit, whether before or after the central unit. Finally, the results show that discourse relations are not evenly distributed across the text, but rather tend to appear in specific places in discourse trees and are sensitive to the semantics of the discourse relation in question.

Keywords: discourse relations, evaluative text, sentiment analysis, Rhetorical Structure Theory, Basque

Introduction: Sentiment analysis and discourse

Sentiment analysis aims at extracting subjectivity and opinion from language, often expressed as polarity, that is, whether the semantic orientation of a text, sentence or word is positive or negative. The most basic approach to sentiment analysis consists of using a sentiment lexicon considering the texts as 'bags-of-words' or unordered sequences of words, without taking into account syntactic or discursive features of the text (Taboada, 2016; Benamara, Taboada, & Mathieu, 2017). This technique is applied to various types of texts, such as tweets, headlines, news articles, on-line comments, reviews or performance evaluations. For instance, the polarity of the Basque sentence in Example (1) can be determined by first assigning a polarity or semantic orientation to each of the words, as shown in Example (2). The values of each of the words are extracted from a Basque sentiment lexicon (Alkorta, Gojenola, & Iruskieta, 2018).[1]

(1) Irabazi ezinik jarraitzen du Eibarrek. (KIR17)
win cannot continue AUX Eibar.
"(The soccer team) Eibar continues without winning."

(2) Ezin_0; Eibar_0; Irabazi_{+2}; Jarraitu_0
Cannot_0; Eibar_0; Win_{+2}; Continue_0

If we add up all the values of Example (2), the sentiment value of the sentence would be 0.5 (+2 divided by 4 words). In this case, the semantic orientation given by the bag-of-words approach is not completely right. The sentiment orientation of the sentence should be negative because the soccer team has not won the match. This change in the semantic orientation of the word *irabazi* +2 ('to win') is due to the negation marker *ezin* ('can not'), a contextual valence shifter (Polanyi & Zaenen, 2006) that changes the semantic orientation of words, phrases or sentences. There are many contextual valence shifters in Basque, in addition to negation markers. Some of them operate at the discourse level, whereas some others operate at the morphological or even phonological level. For example, *gozo* means "gently" referring to people which has a positive semantic orientation. Its form with expressive palatalization is *goxo* (change from [z] to [x]). It is a hypocoristic form, changing the semantic meaning and strengthening the semantic orientation because the meaning has more intensity.

The goal of this work is to describe the role of contextual discourse valence shifters in a Basque corpus annotated with discourse relations and their semantic orientation. Specifically, we focus on the role of discourse relations in the interpretation of opinion. Discourse relations hold between two non-overlapping text spans, indicating a connection between the propositions in those spans. Possible labels for those connections include Condition, Elaboration, Cause or Evaluation.

(3) [[Zahartuz zihoala, gero eta zailagoa egiten zitzaion basatikeriak egitea,]$_N$ → [[*NEG*]
[eta horri esker harrapatu zuten.]$_S$] → [POS]POS] (SENTAIZ03)
[[As he got older, it became more and more difficult for him to do wild things,]$_N$ →[[POS]
[and by that he was caught.]$_S$] → [POS]POS]

Example (3) shows a CAUSE rhetorical relation extracted from the opinion text SENTAIZ03. In the first line of the example, the nucleus-unit (N) of the relation presents a negative semantic orientation. On the other hand, the satellite-unit (S) of the relation appears in the second line with a positive semantic orientation. The union of both discourse units is made by a CAUSE rhetorical relation with positive ~~semanti~~c orientation.

[1] In the examples, the code in parentheses represents the domain (KIR = sport) and the text's number. The texts are part of a Basque corpus (Alkorta, Gojenola, & Iruskieta, 2016). Underlined words are valence shifters, i.e., expressions that can change the semantic orientation of the words in their scope.

Previous work has shown that nuclearity (the relative status of spans in a relation), the type of relation or the position of the relation in the text have an effect on the evaluative interpretation of that relation, and also on the overall polarity for the text (Taboada, Voll, & Brooke, 2008; Chardon, Benamara, Mathieu, Popescu, & Asher, 2013); see also next section. Following this work, and applying it to our corpus of Basque texts, we formulated the following research questions, which link the most significant discourse structure phenomena to their role in sentiment analysis:

1. Related to relations: Which span in the discourse relation tends to convey the semantic orientation for the entire relation? Spans in most relations are asymmetric, one being more important or salient (the nucleus) and the other one contributing to the information in the nucleus (the satellite). In other words, does the semantic orientation of the entire relation tend to match the semantic orientation of the nucleus or the satellite?

2. Related to the text: Which part of the text tends to convey the semantic orientation for the entire text? Do relations at different levels of the discourse tree contribute more or less to the semantic orientation for the entire text?

3. Related to the text: What is the relationship between central unit for the text and semantic orientation?

The work is organized as follows: after discussing related work, we describe data and methodology. We then proceed to present and discuss the results of the analysis, concluding with possible directions for future work.

Related work

Polanyi and Zaenen (2006) proposed that contextual information plays a role in the interpretation of evaluative words when performing automatic sentiment analysis. Their proposal emphasized phrase-level information, such as intensifiers (*great* in *great movie*) or negation (*not such a great movie*), but also considered conjunctions and discourse adverbs such as *although* or *however*. This naturally leads to considering the entire discourse relation that those connectives signal, and how it affects the meaning of evaluative words in the relation. For instance, Trnavac and Taboada (2012, p. 305) discuss Example (4), pointing out that the positive meaning conveyed by *fun* and *good* is tempered, as it were, by the conditional relation those words are part of.

(4) Fun is good, but only if you know when to stop.

Continuing this line of research, Trnavac, Das, and Taboada (2016) explore not only the types of relations that play a role in changing the polarity of the words they contain, but also which part, nucleus or satellite, most directly conveys evaluative meaning in discourse relations. They found that there is a group of relations that tend to intensify the polarity of evaluative words, namely Concession, Elaboration, Evaluation, Evidence and Restatement. They also observed a strong tendency for evaluative words to be found in the nucleus of a discourse relation.

Other research in sentiment analysis has made use of either manually annotated or automatically extracted discourse relations, to ascertain whether incorporating discourse information improves the results of sentiment analysis systems. Different discourse relation frameworks have been deployed in this work.

Some of this research has followed Rhetorical Structure Theory (Bhatia, Ji, & Eisenstein, 2015; Heerschop et al., 2011; Zirn, Niepert, Stuckenschmidt, & Strube, 2011), whereas a different strand employs Segmented Discourse Representation Theory (SDRT) for annotation and use in sentiment analysis systems (Chardon et al., 2013; Asher, Benamara, & Mathieu, 2008). Table 1 summarizes some of this work, by framework, type of discourse information annotated and granularity of the sentiment annotation.

Authors	Approach	Discourse information	Sentiment information
Mukherjee and Bhattacharyya (2012)	-	Discourse markers, connectives and Conditional discourse relations	Semantic orientation of tweets
Heerschop et al. (2011)	RST	Nuclearity: nucleus / satellite, RST relation types	-
Chardon et al. (2013)	SDRT	Partial discourse information, full discourse information	Semantic orientation at discourse unit and document level
Bhatia et al. (2015)	RST	Discourse unit position, Recursive neural networks, RST parser	Semantic orientation at discourse unit and document level
Asher et al. (2008)	SDRT	Relations: CONTRAST, CORRECTION, SUPPORT, RESULT and CONTINUATION	Four groups: reporting, judgment, advise and sentiment
Somasundaran, Namata, Wiebe, and Getoor (2009)	Dialog Acts	Types of frame relations (reinforcing and non-reinforcing relations)	Polarity (positive, negative, neutral) of dialogue units
Taboada et al. (2008)	RST	Nuclei, topic sentences	Semantic orientation of nuclei and texts
Zirn et al. (2011)	RST	Relations: CONTRAST (CONCESSION and CONTRAST), NON-CONTRAST.	Semantic orientation of discourse relations
Zhou, Li, Gao, Wei, and Wong (2011)	-	A set of cue-phrase-based patterns	-

Table 1
Previous work on discourse-based sentiment analysis.

Similarly, corpora used in these projects range in genre or text type, size and type of annotation. The most common text type is reviews, given the natural interest in extracting sentiment from online reviews. Other text types include tweets (Mukherjee & Bhattacharyya, 2012) or news (Chardon et al., 2013). Regarding the size of corpora, it ranges from 120 reviews in Zirn et al. (2011) to 1,000 reviews in Heerschop et al. (2011). For shorter texts, the corpus presented in Bhatia et al. (2015) consists of 32,000 tweets. Finally, some corpora are polarity-balanced, with the same number of positive and negative texts (Heerschop et al., 2011; Zirn et al., 2011).

With respect to methodology, there is a great amount of variation. Most of these projects annotate the corpus under consideration following annotation guidelines, leading to a gold standard that was created by humans (Chardon et al., 2013). Some research, on the other hand, makes use of automatic or semi-automatic methods to perform annotation. For instance, Heerschop et al. (2011) and Zirn et al. (2011) use sentiment lexicons to assign sentiment valence to words in the text. For annotation of discourse phenomena, automatic methods include discourse reweighing and Rhetorical Recursive Neural Networks (Bhatia et al., 2015), discourse-based bag-of-words models (Mukherjee & Bhattacharyya, 2012) and Sentiment Classifier Discourse Parser (Zirn et al., 2011).

Once a corpus is annotated with both sentiment and discourse information, the crucial question is whether the discourse information helps in automatically determining the sentiment. This is the evaluation process. In some cases, the evaluation measures the effects of linguistic phenomena. For instance, Chardon et al. (2013) use a bag-of-segments and discourse information, Mukherjee and Bhattacharyya (2012) evaluate a lexicon-based classification with and without discourse information, Heerschop et al.

(2011) evaluate the effect of word position and, finally, Zirn et al. (2011) evaluate sentence classification with nucleus or satellite information with RST relations. On the other hand, statistical approaches have used Support Vector Machine models (Mukherjee & Bhattacharyya, 2012) and Markov logic networks (Zirn et al., 2011). In most cases, the use of discourse information leads to positive results, i.e., an increase of accuracy for sentiment analysis.

This work presents similarities in several aspects with Bhatia et al. (2015) and Heerschop et al. (2011) in the sense that, like us, they use discourse information to detect the most important text spans and apply different treatments to them. The main difference lies in the main objectives of each work. Bhatia et al. (2015); Heerschop et al. (2011) use this approximation to check if their results improve upon previous work. In contrast, our aim is to analyze the interaction between rhetorical relations and sentiment analysis. In other words, we want to check whether some discourse structures affect the sentiment valence or semantic orientation of discourse relations or what the effect of nuclearity is in different text spans. This information can, naturally, be deployed in sentiment analysis, but our first goal is to explore this relationship, before we can decide how best to implement this knowledge.

Our work is unique in that it explores the interaction of sentiment and discourse for Basque, a relatively understudied language. Existing work has approached Basque from the point of view of RST (Iruskieta, 2014) and, separately, from the point of view of sentiment analysis (San Vicente, Saralegi, & Agerri, 2015; Alkorta et al., 2018). Some of our previous work has combined the RST framework and sentiment analysis (Alkorta, Gojenola, Iruskieta, & Pérez, 2015; Alkorta, Gojenola, Iruskieta, & Taboada, 2017). In this paper, we pursue this relationship more rigorously.

Data and methodology

In this section, we will present: 1) a description of the corpus and the annotation of discourse relations, 2) a parameter analysis, 3) the method for comparison of parameters and 4) the reliability and evaluation measures.

1) **The corpus and annotation of discourse relations.** The study of discourse relations and sentiment in Basque is based on 28 reviews about literature. These reviews are part of the Basque Opinion Corpus (Alkorta et al., 2016) and they are also available in the Basque RST Treebank[2] (Iruskieta et al., 2013).

Discourse relation	**Instances**
ENABLEMENT/MOTIVATION	6
CONDITIONAL subgroup	18
CONTRAST	26
EVIDENCE/JUSTIFY	53
CONCESSION/ANTITHESIS	70
CAUSE subgroup	71
EVALUATION/INTERPRETATION	140
Total	**384**

Table 2
Discourse relations of the study.

[2] http://ixa2.si.ehu.es/diskurtsoa/index.php.

We first extracted the discourse relations in Table 2 from the reviews. In total, 384 discourse relations of seven different types were extracted. According to the Classical Mann and Thompson extended classification (Mann & Thompson, 1988), there are 12 types of discourse relations[3] but we decided to take seven of them into account. The rationale behind this decision was that the selected discourse relations were more likely to have subjective content in comparison with others. Moreover, Trnavac et al. (2016) show that CONCESSION, ELABORATION, EVALUATION, EVIDENCE and RESTATEMENT most frequently intensify the polarity of the opinion words they contain. After the extraction, two linguists assigned the semantic orientation of these reviews, the text spans of the discourse relations and the discourse relations.

2) **Measuring the inter-annotator agreement of the semantic orientation.** The semantic orientation of discourse relations and their spans were annotated by one person. We did, however, take a subset of the corpus and calculated Cohen's kappa coefficient between two linguists, to test the level of reliability that could be achieved in such annotation.

For the reliability study, two linguists assigned semantic orientation to discourse relations and their spans. This involved three different types of semantic orientation per discourse relation: i) semantic orientation of the discourse relation as a whole, ii) semantic orientation of the nucleus, and iii) semantic orientation of the satellite (except in multinuclear relations). In total, 40% of the corpus was annotated by both linguists. The level of inter-annotator agreement, measured using Cohen's kappa, was 0.58, which is considered as 'moderate agreement' according to Landis and Koch (1977).

		A2			
		NEG	**NEU**	**POS**	**Grand Total**
A1	**NEG**	64	**27**	7	98
	NEU	17	65	16	98
	POS	11	**28**	158	197
	Grand Total	92	120	181	393

Table 3
Contingency table of two annotators regarding the semantic orientation annotation of discourse relations.

In Table 3 we show a contingency table of confusion across the two annotators. We can see that the largest differences are related to the neutral semantic orientation. When one annotator (A2) assigned a neutral semantic orientation, the other annotator (A1) assigned a positive or negative one. This occurs in 120 instances. The opposite case also occurred, but with less frequency (98 instances). This is somewhat expected, as we have a tendency to conflate subjectivity (i.e., semantic orientation) with either positive or negative polarity. Neutral semantic orientation is more difficult to identify accurately (see Example (5)).

(5) [Idazleak berak esan bezala, "gertaera xumeak" dira,]$_1$
[baina hala ere, arazoak euren onetik ateratzen ditu pertsonaiak.]$_2$
[As the writer himself says, these are "humble facts."]$_1$
[but even so, problems drive people up the wall.]$_2$

[3] The 12 types of discourse relations according to Rhetorical Structure Theory (RST) are the following: CIRCUMSTANCE, SOLUTIONHOOD, ELABORATION, BACKGROUND, ENABLEMENT and MOTIVATION group, EVIDENCE and JUSTIFY group, relations of CAUSE, ANTITHESIS and CONCESSION, CONDITION and OTHERWISE, INTERPRETATION and EVALUATION, RESTATEMENT and SUMMARY, and SEQUENCE and CONTRAST.

In Example (5), one annotator has considered all the rhetorical relation and its spans neutral while the other annotator does not consider it. According to one annotator, the semantic orientation is positive and negative for the first and second discourse units, respectively. Besides, the semantic orientation of all the rhetorical relations as a whole is negative. In contrast, the second annotator has considered the semantic orientation of the first and second discourse units and all the rhetorical relation neutral.

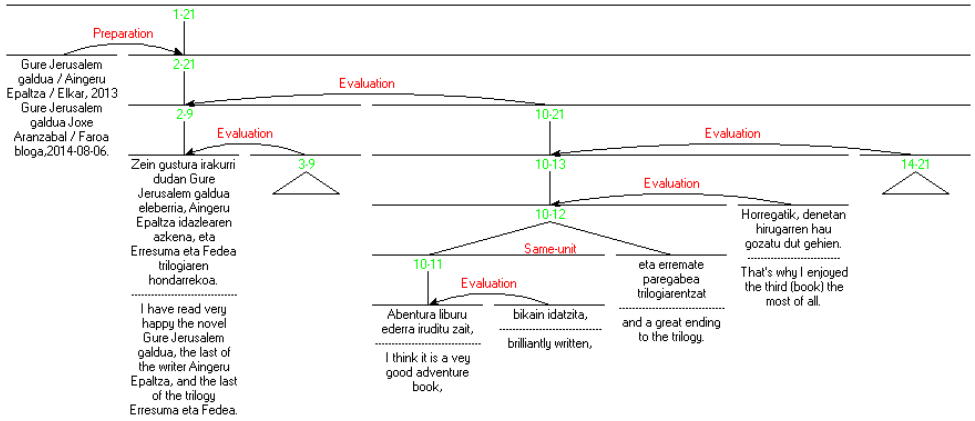


Figure 1. RST-tree for text SENTFAR-01.

3) **Analysis of parameters.** We use Figure 1 and the EVALUATION relation (EDU 10-13) as an example for illustration purposes. After the extraction process, discourse relations were analyzed with the following parameters:

   i) Nuclearity. The discourse relation can be mononuclear or multinuclear. In the first case, one of the units of the discourse relation is more important than the other, in the sense that it contributes more to the effect that the relation intends to achieve on the listener/reader. The smallest units of discourse that enter into a discourse relation are usually referred to as Elementary Discourse Units (EDUs). The most important EDU is labelled as nucleus, with the other unit receiving the satellite label. Relations that are composed of a nucleus and satellite can appear in two different patterns: Nucleus-Satellite or Satellite-Nucleus. In addition, some relations present two or more EDUs as having the same importance; those are referred to as multinuclear relations. In Figure 1, the span 11 is linked with a EVALUATION relation to the span 10 and follows the N(ucleus)-S(atellite) pattern. The span 13 follows the same structure and it is linked to span 10-12 with an EVALUATION relation.

   ii) Semantic orientation to text spans and discourse relations. We assigned a semantic orientation[4] to different text spans and discourse relations. First, we assigned the semantic orientation to text spans. In the example, the text spans 10-12 and 13 both have a positive semantic orientation. This means that the EVALUATION relation links two positive discourse units

[4] We assigned three types of semantic orientation to discourse units and discourse relations: POS, positive; NEG, negative and NEU, neutral.

(10-12 and 13). Then, we assigned the semantic orientation to the entire discourse relation (positive).

iii) Distance to central unit. The central unit is the most important nucleus in the text, one that most directly contributes to the intention or effect that the text intends to achieve (Iruskieta, da Cunha, & Taboada, 2015; Marcu, 2000). We measured the distance between discourse relations and central unit counting the number and types (NN or NS) of discourse relations between the central unit and the current discourse relation (horizontal distance). As far as the span 13 is concerned, the distance in discourse relations is +2 in Figure 1, because there are 2 NS discourse relations (2 NS EVALUATIONs) between span 13 and span 2, which is the central unit, we give a distance of +2.

iv) Type of discourse relation. The last parameter used in our study was the type of relation. As Table 2 shows, in total seven types of discourse relations were considered.

The EVALUATION discourse relation joins the 10-12 and 13 EDUs. That means that the 10-12 EDU explains a situation and EDU 13 makes an evaluative comment about it.

Results

In this section, we present the results trying to address our main research questions. We first explore our conclusions regarding nuclearity and sentiment analysis, to then approach the semantic orientation in texts. We end this section with a discussion of the position of discourse relations in opinion texts.

Semantic orientation, nuclearity and position in discourse relations

Tables 4 and 5 present the breakdown of semantic orientation with regard to nuclearity and position. Our first question is whether the semantic orientation of the entire relation is more often that of the nucleus or the satellite. For instance, if the relation is positive overall, is it more often the case that the nucleus is also positive, or the satellite? Table 4 shows that, compared to satellites, the nuclei coincides in more cases in semantic orientation with the whole discourse relation they belong to (0.71 in the case of nuclei and 0.64 in the case of satellites). However, this tendency does not happen in all types of discourse relations. In the case of EVALUATION, the semantic orientation of the rhetorical relation is more often the same as the semantic orientation of the satellite (0.82 for the satellite vs. 0.69 for the nucleus).

(6) [[Igerabideren estiloa aberatsa baita,]$_N$ → [[POS]
[egoerari eta pertsonaiei ederki egokitua,]$_S$] → [POS]POS] (SENTAIZ02)
[[The style of Igerabide is rich,]$_N$ →[[POS]
[beautifully adapted to the situation and the characters.]$_S$ → [POS]POS]

In Example (6), there is an annotation of the EVALUATION rhetorical relation in the text SENTAIZ02. In the first line of the example, there is a nucleus-unit with a positive semantic orientation. In the second line, the semantic orientation is also positive in satellite-unit and consequently the semantic orientation of the rhetorical relation is positive.

By contrast, the CONCESSION group shows the largest difference between nuclei and satellites, with a 0.70 match for nuclei, 0.33 for satellites. Another aspect is where the highest SO match between discourse relations and nuclei or satellites appears, the results showing that nuclei in CAUSE and the EVIDENCE group match most (0.83). In the case of the CONTRAST relation, the semantic orientation of the discourse relation with one nuclei is high (0.73) but much less frequent with both nuclei (0.31).

Relations	Instances	N	S
CAUSE	71	0.83	0.66
CONCESSION group	70	0.70	0.33
CONDITIONAL	18	0.61	0.56
ENABLEMENT group	6	0.60	0.5
EVALUATION group	140	0.69	0.82
EVIDENCE group	53	0.83	0.64
CONTRAST*	26	0.73	0.31
Total	384	0.73	0.65

Table 4
Agreement of the semantic orientation of discourse relations with nucleus and satellites.

Relations	Instances	First span	Last span
CAUSE	71	0.66	0.83
CONCESSION group	70	0.30	0.73
CONDITIONAL	18	0.28	0.89
ENABLEMENT group	6	0.67	0.50
EVALUATION group	140	0.68	0.83
EVIDENCE group	53	0.79	0.68
CONTRAST	26	0.35	0.69
Total	384	0.58	0.78

Table 5
Agreement of the semantic orientation of discourse relations and position (with the first or last span).

Whereas our first comparison relates to the hierarchical structure of discourse, in terms of nuclei and satellites, we then consider linear structure. We are interested in whether the semantic orientation of the discourse relation as a whole tends to match that of the EDU to its most immediate right or the EDU to the immediate left.

When we compare the semantic orientation of the discourse relation and position (first or last span), the results are clearer. Table 5 shows that the last span fits more (0.78) with the discourse relation in comparison with the first span (0.58). However, not all the discourse relations show the same tendency. ENABLEMENT and EVIDENCE have the same semantic orientation as that of the first span more often. In the case of ENABLEMENT, the score with the first and last spans is 0.67 and 0.50, respectively, while the scores are 0.79 and 0.68 for the EVIDENCE relation. The highest difference in score appears in the CONDITIONAL relation, because the first span has a low agreement in semantic orientation (0.28) and a high one (0.89) with the last span. Finally, the last spans of EVALUATION and CAUSE matches more often (0.83) the semantic orientation of the discourse relation as a whole.

In summary, regarding the semantic orientation of discourse relations, the results show that the semantic orientation of the entire relation is more often that of its nucleus. This is to be expected because, according to the RST framework, the nucleus is the most important EDU of the relation, so it makes sense that the semantic orientation of the nucleus contributes the most to the semantic orientation of the relation as a whole. However, this is not the case for all types of relations. For example, in the EVALUATION group, the semantic orientation of the satellite fits in more occasions. This is because of the characteristics of

this type of relation. In the EVALUATION group, the nucleus presents a situation and the satellite makes an evaluative comment about the situation.

Comparison of the semantic orientation of the text and discourse relation's distance from the central unit

Another aspect we wanted to investigate is the match of semantic orientation of the text and the semantic orientation of a discourse relation, according to the distance of the discourse relation from the central unit in the RST tree. The question is, given the semantic orientation of the entire text, does the semantic orientation of relations in the text change as those relations are farther away from the central unit? Figure 2 shows that the match of semantic orientation between opinion texts and their discourse relations is higher when the discourse relation is closer to the central unit. Distances −2, −1, 0, 1 and 2 present the highest match between all the distances between from −6 to +6[5].

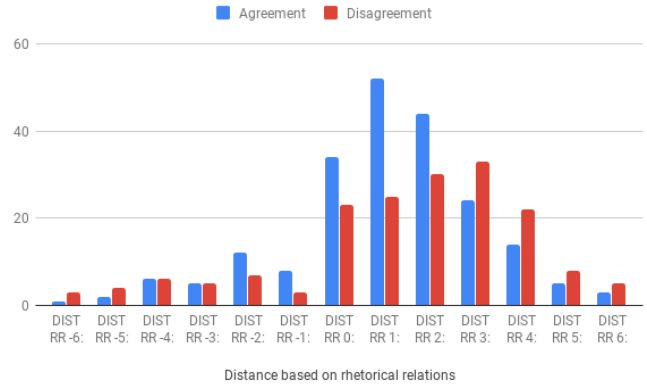


Figure 2. Semantic orientation of the text and discourse relations according to distance from the central unit.

If the results are analyzed for each discourse relation, there are some differences. In the CAUSE discourse relation, the instances that have the same semantic orientation prevail only in distances 0 and 1. In contrast with the overall tendency, the CONCESSION/ANTITHESIS group shows a discontinuous pattern regarding the same semantic orientation of the text. Relations with the same semantic orientation of the text are more frequent in two different non contiguous distance portions. Instances with the same semantic orientation happen more in distances −2, +1, +2, while a different semantic orientation is more frequent in distances 0 and +3.

Our interpretation about the semantic orientation between discourse relations and texts based on position is that when a discourse relation is closer to the central unit, there is typically a match between the semantic orientation of that discourse relation and the semantic orientation of the text overall. In our view, the reason for this is that the topic and the effect of discourse relations that are closer to the central unit are more closely aligned with the global topic and intended effect of the text. As a consequence, the semantic orientations will also be closely aligned.

[5] Distances with the signal – are situated to the left of the central unit. Distances with a positive sign '+' are situated to the right of the central unit.

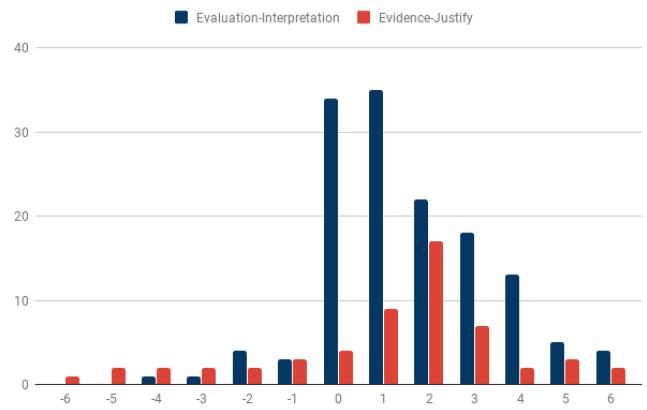


Figure 3. The most usual position of EVALUATION-INTERPRETATION and EVIDENCE-JUSTIFY discourse relations in opinion texts.

Discourse relations in opinion texts

An interesting aspect that deserves attention is the most usual position of different discourse relations in opinion texts. We have analyzed the distribution of instances of discourse relations with respect to the central unit. Figure 3 presents the INTERPRETATION-EVALUATION and EVIDENCE-JUSTIFY discourse relations, showing that for INTERPRETATION-EVALUATION, there are few instances before the central unit and, after it there is a high number of instances in distances 0 and 1. Finally, the number of instances decreases when moving away from the central unit. The CONTRAST multinuclear relation also shows similar characteristics, in fact, there are many instances of this discourse relation in distances -1 and $+3$ but between those distances, there is almost no instance. The situation is different in the case of EVIDENCE-JUSTIFY where, in contrast to the previous relations, it shows a greater degree between distances -1 to $+3$ but a regular distribution in the other distances because its instances are more uniform before and after the central unit.

Additionally, Figure 4 offers the results of the ANTITHESIS-CONCESSION and CAUSE-RESULT-PURPOSE discourse relations. These discourse relations share some similarities regarding their distance from the central unit. As it happened with the EVIDENCE-JUSTIFY group, these groups are distributed in all the distances in greater or lesser degree. But in contrast to other discourse groups, they present two peaks. In both relations, one peak is situated before the central unit and another one after it. In the case of ANTITHESIS-CONCESSION, the peaks are situated in distances -2 and $+3$. In a similar way, the peaks appear in distances -4 and $+1$ when it comes to the CAUSE-RESULTS-PURPOSE group.

Finally, Figure 5 shows the position of types of discourse relations based on our corpus. The results show big differences between instances of types of discourse relations. From distance -1, EVALUATION-INTERPRETATION is the most usual relation. It presents more than 40 instances in distances $+1$ and $+2$.

Although EVALUATION-INTERPRETATION is by far the most usual relation from distance -1, the

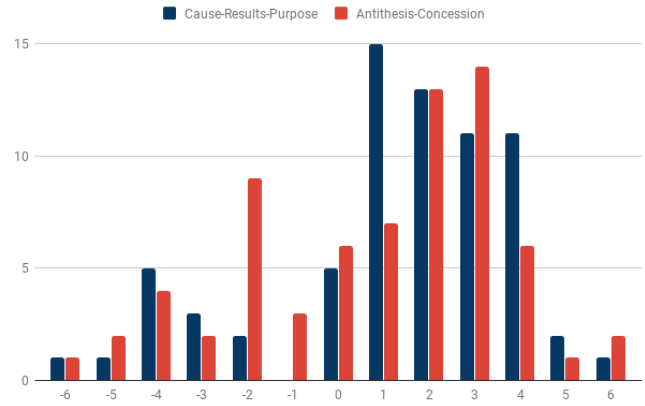


Figure 4. The most usual position of ANTITHESIS-CONCESSION and CAUSE-RESULT-PURPOSE discourse relations in opinion texts.

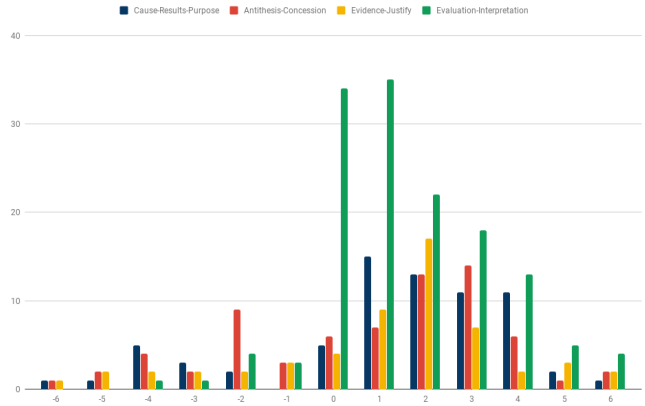


Figure 5. Instances of discourse relations based on distance from central unit.

CONCESSION relation in distance 0 and the CAUSE relation in distance +1 appear with high frequency. In distance +2, CAUSE and JUSTIFY show the same quantity of instances and in distance +3, the CONCESSION relation prevails over other discourse relations. Finally, in distance +4, the RESULT and PURPOSE relations are the most usual.

As far as the position of all the discourse relations of the corpus in texts is concerned, our interpretation is that there is a pattern even when the number of instances in modest to give a clear evidence. Taking into account the results of Figure 5, it seems that the order of discourse relations in texts is the following:

PURPOSE (distance −4), CONCESSION (−3 and −2), and EVALUATION (other distances). Without the EVALUATION relation, the order would be the following in other distances: CONCESSION (0), CAUSE (+1), CAUSE/JUSTIFY (+2), CONCESSION (+3), and RESULTS/PURPOSE (+4).

Conclusion and Future Work

This work presents an empirical study on semantic orientation and discourse relations. Our goal was to explore the interaction of semantic orientation (in the form of positive/negative or neutral sentiment) and discourse relations, in particular certain aspects of discourse relations such as nuclearity, position in the text or type of relation.

We started out by creating an annotated corpus of Basque opinion texts. We relied on a corpus that had already been annotated with discourse relations and assigned semantic orientation to all the discourse relations in the texts and their spans. In order to ensure the quality of the annotations, we performed an inter-annotator agreement study, which showed moderate agreement.

Using this corpus, we calculated different measures of the interplay between sentiment and discourse structure. We examined the semantic orientation of nuclei and satellites, and its match (or not) with the semantic orientation of the entire relation. We also considered any differences across relations.

Secondly, we calculated the distance of each discourse relation from the central unit, and whether the distance entails any difference in semantic orientation matching between the semantic orientation of the entire text and that of the relation under consideration.

The results of all those measures lead to the following conclusions:

1- Discourse relations coincide more in semantic orientation with their nuclei than with their satellites.

2- Discourse relations coincide more in semantic orientation with their last spans (right) than with their first spans (left).

3- The semantic orientation of the relation coincides much more when the satellite appears in the last span (left).

4- When it comes to distance from the central unit, the semantic orientation of relations which are at distances situated between −2 and +2 coincides in more occasions.

This work has had some limitations. Although the number of instances is considerable (in total, 384 instances), their distribution is not uniform across different types of relations. For example, CONDITIONAL (18 instances) and ENABLEMENT/MOTIVATION (6 instances) relations have few instances, making broad generalizations problematic. On the other hand, the inter-annotator agreement as for the semantic orientation of discourse relations and its spans is κ 0.58 which is considered 'moderate agreement' according to Landis and Koch (1977).

In the near future, we would like to pursue further study. First of all, we want to increase the number of discourse relations in the corpus, in order to achieve more precise results. We also plan to rewrite the guidelines to annotate the semantic orientation of discourse relations to achieve a higher inter-annotator agreement. Extending the analysis to others discourse relations is the other objective. Finally, after meeting the previous objectives, we would like to refine the discourse structure of opinion texts proposed in this study. For the future work we plan to automate these findings and add different pipelines to analyze the semantic orientation taking into account these discourse phenomena, following the subsequent tasks:

i) determine the EDUs and the central unit of the text with an existing central unit detector (Atutxa, Bengoetxea, Diaz de Ilarraza, & Iruskieta, 2019), *ii*) after that, add discourse relations with EusDisParser, the Basque RST parser developed by Iruskieta and Braud (2019) and *iii*) finally add the semantic orientation with Sentitegi (Alkorta et al., 2018).

Acknowledgements

This research is funded by the Basque Government PRE_2018_2_0033 grant. We would like to add a special thank also to the Simon Fraser University for given infrastructure to develop this study.

References

Alkorta, J., Gojenola, K., & Iruskieta, M. (2016). Creating and evaluating a polarity-balanced corpus for basque sentiment analysis. In *Iwoda16 fourth international workshop on discourse analysis. santiago de compostela, september* (Vol. 29).

Alkorta, J., Gojenola, K., & Iruskieta, M. (2018). Sentitegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, *22*(4).

Alkorta, J., Gojenola, K., Iruskieta, M., & Pérez, A. (2015). Using relational discourse structure information in basque sentiment analysis. In *Sepln 5th workshop rst and discourse studies.*

Alkorta, J., Gojenola, K., Iruskieta, M., & Taboada, M. (2017). Using lexical level information in discourse structures for basque sentiment analysis. In *Proceedings of the 6th workshop on recent advances in rst and related formalisms* (pp. 39–47).

Asher, N., Benamara, F., & Mathieu, Y. Y. (2008, August). Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters* (pp. 7–10). Manchester, UK: Coling 2008 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C08-2002

Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., & Iruskieta, M. (2019, 09). Towards a top-down approach for an automatic discourse analysis for basque: Segmentation and central unit detection tool. *PLOS ONE*, *14*(9), 1-25. Retrieved from https://doi.org/10.1371/journal.pone.0221639 doi: 10.1371/journal.pone.0221639

Benamara, F., Taboada, M., & Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, *43*(1), 201-264.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015, September). Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2212–2218). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D15-1263 doi: 10.18653/v1/D15-1263

Chardon, B., Benamara, F., Mathieu, Y., Popescu, V., & Asher, N. (2013). Measuring the effect of discourse structure on sentiment analysis. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 25–37). Berlin, Heidelberg: Springer Berlin Heidelberg.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). Polarity analysis of texts using discourse structure. In *Proceedings of the 20th acm international conference on information and knowledge management* (pp. 1061–1070). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2063576.2063730 doi: 10.1145/2063576.2063730

Iruskieta, M. (2014). Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea*.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., & de Lacalle, O. L. (2013). The rst basque treebank: an online search interface to check rhetorical relations. In *4th workshop rst and discourse studies* (pp. 40–49).

Iruskieta, M., & Braud, C. (2019). Eusdisparser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the workshop on discourse relation parsing and treebanking 2019* (pp. 62–71).

Iruskieta, M., da Cunha, I., & Taboada, M. (2015, June). A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, *49*(2), 263–309. Retrieved from http://dx.doi.org/10.1007/s10579-014-9271-6 doi: 10.1007/s10579-014-9271-6

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Mann, W. C., & 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.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: MIT Press.

Mukherjee, S., & Bhattacharyya, P. (2012, December). Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of COLING 2012* (pp. 1847–1864). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C12-1113

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications* (pp. 1–10). Springer.

San Vicente, I., Saralegi, X., & Agerri, R. (2015, June). EliXa: A modular and flexible ABSA platform. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 748–752). Denver, Colorado: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S15-2127 doi: 10.18653/v1/S15-2127

Somasundaran, S., Namata, G., Wiebe, J., & Getoor, L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1* (pp. 170–179). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1699510.1699533

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, *2*, 325-347.

Taboada, M., Voll, K., & Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*.

Trnavac, R., Das, D., & Taboada, M. (2016). Discourse relations and evaluation. *Corpora*, *11*(2), 169–190.

Trnavac, R., & Taboada, M. (2012). The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, *34*(3), 301-318.

Zhou, L., Li, B., Gao, W., Wei, Z., & Wong, K.-F. (2011, July). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 162–171). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D11-1015

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011, November). Fine-grained sentiment analysis with structural features. In *Proceedings of 5th international joint conference on natural language processing* (pp. 336–344). Chiang Mai, Thailand: Asian Federation of Natural Language Processing. Retrieved from https://www.aclweb.org/anthology/I11-1038

CONCLUSION

5

Contributions, conclusions, and future work

As we mentioned at the beginning of this thesis work, nowadays there are numerous opinions on the Internet and extracting subjective information from these texts and, specifically, knowing if these opinion texts have a positive or negative opinion is a challenge. In this direction, we have developed tools and resources for extracting subjective information from Basque opinion texts. Moreover, we have studied different linguistic phenomena that change the sentiment valence of words, known as contextual valence shifters.

In this section, we will present the contributions and conclusions drawn from the sentiment analysis regarding Basque, as well as the challenges facing the future.

5.1 Contributions

This thesis is one of the first works in sentiment analysis of written texts in Basque from an applied linguistics perspective. After building a corpus of opinion texts and developing a sentiment lexicon, the phenomena affecting sentiment valence and semantic orientation of the words have been studied. In the future, these aspects will need to be considered when developing a document-level sentiment classification based on sentiment lexicon. These resources and tools we have developed as well as the study of linguistic phenomena are the result of the methodology we have followed.

5.1.1 Tools and resources related to sentiment analysis

Concerning sentiment analysis, we have developed an annotated corpus of opinion texts in Basque, a sentiment lexicon in Basque called *Sentitegi* and a document-level sentiment classifier in Basque.

- Basque Opinion Corpus (Alkorta et al., 2016). This corpus includes 240 opinion texts from six domains. It has annotated from discourse structure and subjectivity. The *RST* approach (Mann and Thompson, 1988) has been used for annotating discourse structure and, in total, 70 opinion texts have been annotated. 192 central units of opinion texts have also been identified. From sentiment analysis, we have determined whether the text is positive or negative. Furthermore, in 28 opinion texts, for certain types of discourse relations and their constituents (nucleus and satellite), we have assigned the semantic orientation (positive or negative) as well as sentiment value (numerical sentiment value from -5 to $+5$).

- *Sentitegi*, the sentiment lexicon in Basque (Alkorta et al., 2018). This sentiment lexicon includes 1,237 words of 4 grammar categories. It is based on the Spanish lexicon of the SO-CAL tool (Brooke et al., 2009) and it is also enriched with the English lexicon (Taboada et al., 2011). This lexicon has been adapted to six domains: weather, sport, politics, music, film and literature and the Basque Opinion Corpus has been used for that purpose.

- Document-level sentiment classification tool in Basque. Another contribution of this thesis has been the sentiment classifier based on sentiment lexicon. We have based our work on the SO-CAL tool (Taboada et al., 2011) to develop the Basque version. However, because of the typological differences between English and Basque, there are some differences between the tools of both languages.

  In the English sentiment classifier, there are sentiment lexicons, morphology-related rules, and general rules. In contrast, in the Basque version, we have removed the rules related to morphology and we have integrated the *Eustagger* lemmatizer (Alegria et al., 2002). The Basque classifier, firstly, lemmatizes the text and, then, looks up whether the lemmatized word is contained in the lexicon and in this case the classifier assigns a sentiment valence to the word.

5.1.2 Analysis of valence shifters in Basque

To work on sentiment analysis, we have analyzed contextual valence shifters of different grammatical levels that affect to words and phrases. We have focused on (Polanyi and Zaenen, 2006):

- We have studied morphological valence shifters. As we have seen, the phonological phenomenon of expressive palatalization, based on the instances of the corpus, reinforces the sentiment valence of the word. In morphology, we have found prefixes and suffixes that can strengthen or weaken the valence of the word. Besides, we have studied several affixes related to negation.

- We have also worked on the syntactic level and specifically on the negation markers and their scope. Our study shows that negation markers can strengthen or weaken the sentiment valence of words or phrases but also that in some cases they do not change the valence or semantic orientation.

  It is also worth mentioning that the negation marker ("ezin") has a double behavior depending on the grammatical category of the word around it. The results have also shown that some negation markers appear more frequently with other words. We call these lexicalized structures. In terms of CG rules to identify elements related to negation, we have found that it is more difficult to identify scope than negation markers and lexicalized structures because of their irregular structure and variable length.

- Finally, we have also studied the discourse level and, for that purpose, we followed the *RST* approach (Mann and Thompson, 1988). In this work, we have focused on discourse relations and central unit. In discourse relations, we have seen that the nuclei or the final text span of the discourse relations strengthen the agreement of the semantic orientation between these constituents and the discourse relations. Besides, when a discourse relation is closer to the central unit, the semantic agreement of the relation and the whole text is strengthened. Consequently, according to the results, the nucleus and the last text span in discourse relations and central unit in the texts are strengthening valence shifters.

Regarding the central unit, we have also studied the words with sentiment valence. According to our results, in the central unit, 12.40% of the words have sentiment valence and their most usual grammatical categories are adjectives, nouns, and verbs. In the central unit, we have also studied words with sentiment valence from the domain and the results show differences between the domains. For example, in sports, the most common grammatical category with sentiment valence is the verb. In contrast, in weather, the most common grammatical category with sentiment valence is the adjective.

5.2 Conclusions

In this work, we established some goals before starting the thesis and have fulfilled those goals. After completing the thesis, we come to the following conclusions.

- Translation to create a sentiment lexicon in Basque can be a good option for sentiment analysis. In fact, according to Pearson correlation results, the quality of the vocabulary of the Basque language created following our methodology is good. The correlation coefficient is 0.76 for the words annotated by the lexicon and the gold standard. But in some cases, the gold standard, unlike the lexicon we have developed, assigns a certain value of sentiment to the words, and as a result, the coefficient drops to 0.54. That means that the quality of the lexicon is high but there is a possibility to expand the lexicon incorporating more words with sentiment valence.

- To develop a document-level sentiment classifier based on lexicon, it is necessary to take into account contextual valence shifters. As the work has shown, at different levels of grammar, from phonology to discourse, there are valence shifters and they can have three effects on the sentiment valence or semantic orientation of the words and phrases: strengthening, weakening, or no change.

- The rich morphology of Basque also influences sentiment analysis. The results of the thesis work show that Basque prefixes and suffixes can affect the sentiment valence of words, strengthening or weakening its

value and, in some cases, changing the sign of the valence. From semantics, the morphemes in Basque are diverse and they also influence valences. The thesis analysis also shows that expressive palatalization strengths the sentiment valence of words.

- In syntax, we have found that negation markers strengthen, weaken, or do not change the sentiment valence of words or phrases. Besides, we have identified lexicalized structures. They are negation markers that appear with specific words with great frequency and they also affect the sentiment valence of phrases.

- Discourse structure also plays an important role in sentiment classification based on the lexicon. Nucleus and the last text span in discourse relations and central units in texts affect to agreement of semantic orientation between the mentioned elements, increasing the agreement.

Taking into account the above points, from the linguistic point of view, it can be concluded that we have taken the first steps in the sentiment analysis in Basque. However, it is still necessary to begin to work on other aspects or to deepen them, in particular, the contextual valence shifters that are unexplored.

5.3 Future work

In the study of sentiment analysis and its computational processing, we predict the following lines of work for the future:

- To have an even more diverse Basque Opinion Corpus. Nowadays, the corpus includes opinion texts from specialized newspapers and websites and it is composed of 6 domains. The aim for the future is to compile opinion texts. As mentioned in the development of the methodology,we have found few opinion texts with good quality to study discourse. The aim is to compile texts of this type to study other phenomena that have not yet been explored. For example, when ordinary people write texts it is possible to stretch words by repeating a letter as well as use emoticons, which can also affect semantic orientation.

- To improve and expand the annotation of the discourse structure of the corpus. Currently, 70 of the 240 texts in the corpus are annotated and

we want to increase that number. However, before further annotation, it is necessary to reach a better inter-annotator agreement.

- To extend the sentiment lexicon and to deal with different events of the methodology. As we have seen, the quality of the vocabulary we have created is good but it is not yet possible to identify all the words and/or expressions with subjectivity. As a result, we want to increase its size and make it useful for more domains. Finally, we also want to solve the situation of polysemous words, that is to say, decide what to do with words that have two opposing semantic orientations in different contexts. Words like "ikaragarri" (frightening/enormous) have the opposite semantic orientation when it comes to a movie or an accident.

- To continue identifying and studying further contextual valence shifters in the Basque language. In this thesis, we have studied one type of valence shifter in each grammatical category, but there are still some valence shifters that have not yet been studied. We can mention, for example, conditionals, interrogative sentences or, even more complex, the irony. Some of the valence shifters may have specificities in Basque and others may be similar in different languages.

- To deeper study the discourse level. In this work, we focus on discourse relations and central unity. In the future, however, we want to go further and explore other components of discourse structure. In other words, we want to work on other elements of discourse trees, such as the central subconstituent. We also want to make a deeper study on types of discourse relations and find differences that may exist between them.

- To further develop the document level sentiment classifier. The tool is currently made up of Basque lexicon, lemmatizer, and general rules, but we want to integrate more modules into it. For example, we would like to set up a module related to the contextual valence shifters used in this thesis and check if the quality of the tool improves.

Bibliography

Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., and Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6.

Alkorta, J., Gojenola, K., and Iruskieta, M. (2016). Creating and evaluating a polarity-balanced corpus for Basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela, September*, volume 29.

Alkorta, J., Gojenola, K., and Iruskieta, M. (2018). SentiTegi: Semimanually Created Semantic Oriented Basque Lexicon for Sentiment Analysis. *Computación y Sistemas*, 22(4).

Altuna, B., Aranzabe, M. J., and de Ilarraza, A. D. (2017). Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In *Iñaki Alegria, Ainhoa Latatu, Miren Josu Ormaetxebarria eta Patxi Salaberri (ed.), II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea, 127-134, Udako Euskal Unibertsitatea (UEU), Bilbo.*

Altuna, P., Salaburu, P., Goenaga, P., Lasarte, M. P., Akesolo, L., Azkarate, M., Charriton, P., Eguskitza, A., Haritschelhar, J., King, A., Larrarte, J. M., Mujika, J. A., Oyharçabal, B. n., and Rotaetxe, K. (1985). Eu-

skal Gramatika Lehen urratsak (EGLU) II. *Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo*.

Barnes, J., Lambert, P., and Badia, T. (2018). MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. *arXiv preprint arXiv:1803.08614*.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From English to Spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 383–389.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Das, D. and Taboada, M. (2018). RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Lang. Resour. Eval.*, 52(1):149–184.

Euskara Institutua, EHU (2011). Sareko euskal gramatika, www.ehu.eus/seg.

Iruskieta, M., Da Cunha, I., and Taboada, M. (2015). A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Karlsson, F., Voutilainen, A., Heikkilae, J., and Anttila, A. (2011). *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Mann, W. C. and 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.

Oñederra, L. (1990). *Euskal fonologia: palatalizazioa: asimilazioa eta hots sinbolismoa*. Servicio Editorial de la Universidad del País Vasco = Euskal Herriko Unibertsitatea.

O'Donnell, M. (2000). RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 253–256. Association for Computational Linguistics.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., and Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, (58):77–84.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

San Vicente, I. and Saralegi, X. (2016). Polarity Lexicon Building: to what Extent Is the Manual Effort Worth? In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

San Vicente, I., Saralegi, X., and Agerri, R. (2017). Elixa: A modular and flexible absa platform. *arXiv preprint arXiv:1702.01944*.

Saralegi, X., San Vicente, I., and Ugarteburu, I. (2013). Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-resourced Languages. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 96–108, Berlin, Heidelberg. Springer-Verlag.

Sarasola, I. (2005). *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

Stone, P. J. and Hunt, E. B. (1963). A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Urdangarin, L. M. M. (1982). Morfología de la composición lexical euskérica. *Fontes linguae vasconum: Studia et documenta*, 14(39):233–272.

Zerbitzuak, E. H. (2013). Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. Usurbil: Elhuyar.

Terminology and abbreviations

Sentimenduen analisi (*Sentiment analysis*)
Orientazio semantiko (*Semantic orientation*)
Sentimendu-balentzi (*Sentiment valence*)
(Testuinguruko) balentzia-aldatzaile (*(Contextual) valence shifter*)
Sentimenduen sailkatzaile (*Sentiment classifier*)
Sentimenduen lexikoi (*Sentiment lexicon*)
Murriztapen Gramatika, MG (*Constraint Grammar, CG*)
Ezeztapen-marka (*Negation mark*)
(Ezeztapenaren) irismena (*Scope*)
Egitura lexikalizatu (*Lexicalized structure*)
Egitura Erretorikoaren Teoria (*Rhetorical Structure Theory, RST*)
Unitate zentrala, UZ (*Central Unit, CU*)
Erlaziozko diskurtso-egitura (*Relational discourse structure*)

The writing of this thesis
was ended on 15th October, 2019.