eman ta zabal zazu

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

eman ta zabal zazu

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 *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!

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

INTRODUCTION 1

METHODOLOGY 9

RESULTS **40**

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 [like$_+$]$_+$ that movie.

(2) I [like a lot$_+$]$_{++}$ that movie.

(3) I do not [like$_+$]$_-$ that movie.

(4) I [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 assigned 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.

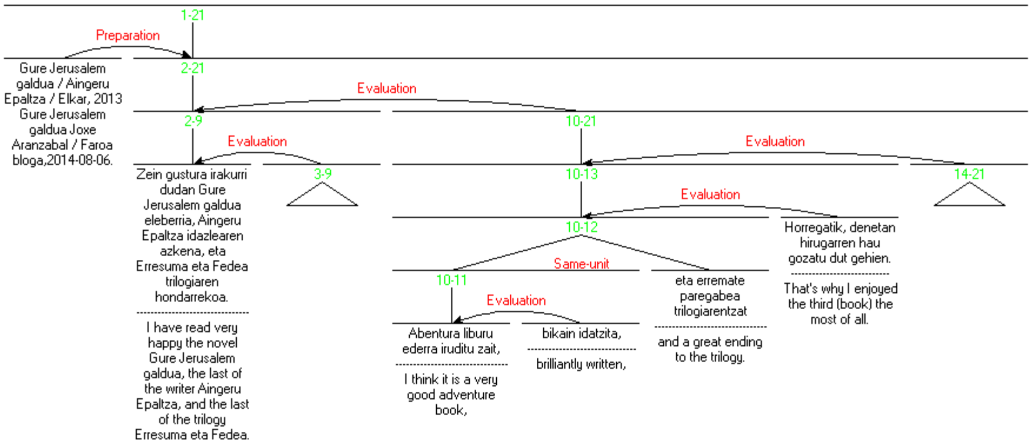

1-21
Preparation
Gure Jerusalem galdua / Aingeru Epaltza / Elkar, 2013 Gure Jerusalem galdua Joxe Aranzabal / Faroa bloga,2014-08-06.
2-21
Evaluation
2-9
10-21
Evaluation
Evaluation
Zein gustura irakurri dudan Gure Jerusalem galdua eleberria, Aingeru Epaltza idazlearen azkena, eta Erresuma eta Fedea trilogiaren hondarrekoa.
I have read very happy the novel Gure Jerusalem galdua, the last of the writer Aingeru Epaltza, and the last of the trilogy Erresuma eta Fedea.
3-9
10-13
14-21
Evaluation
10-12
Same-unit
Horregatik, denetan hirugarren hau gozatu dut gehien.
That's why I enjoyed the third (book) the most of all.
10-11
Evaluation
eta erremate paregabea trilogiarentzat
and a great ending to the trilogy.
Abentura liburu ederra iruditu zait,
I think it is a very good adventure book,
bikain idatzita,
brilliantly written,

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.

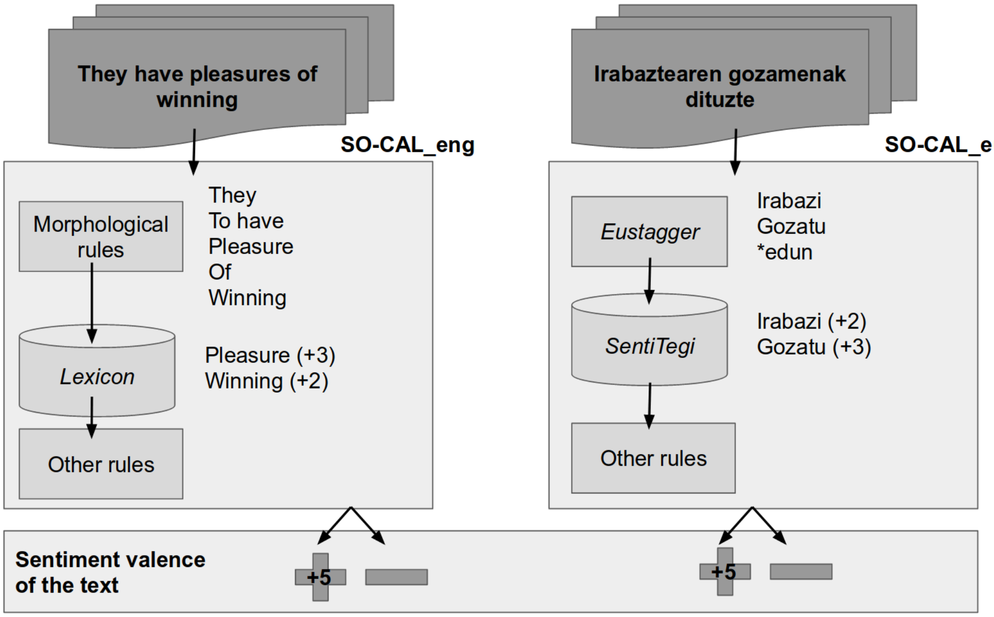


Figure 2.6 – Comparison of the structures of the English and Basque versions of the SO-CAL tool.

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.

doi: 10.13053/CyS-22-4-3075

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.

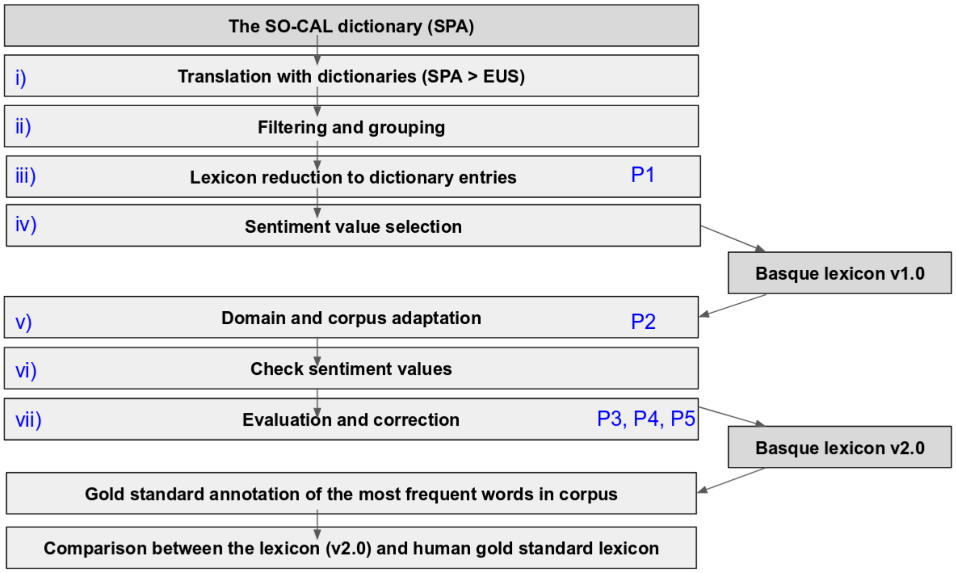

The SO-CAL dictionary (SPA)
i) Translation with dictionaries (SPA > EUS)
ii) Filtering and grouping
iii) Lexicon reduction to dictionary entries
P1
iv) Sentiment value selection
Basque lexicon v1.0
v) Domain and corpus adaptation
P2
vi) Check sentiment values
vii) Evaluation and correction
P3, P4, P5
Basque lexicon v2.0
Gold standard annotation of the most frequent words in corpus
Comparison between the lexicon (v2.0) and human gold standard lexicon

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.

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

[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.

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.

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.

[4] The semantic oriented Basque lexicon is available at: `http://ixa.si.ehu.es/node/11438`

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*$_{+2}$ <u>*ezinik*</u> *jarraitzen du Eibarrek*]$^{+2}$. (KIR17)
[(The soccer team) Eibar continues <u>without</u> 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.

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

[1] The semantic lexicon is available on the web at: http://ixa.si.ehu.es/node/11438

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 85–90
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
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.

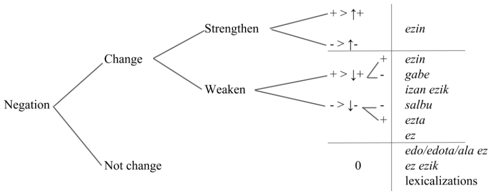


Figure 1: The effects of negation on semantic orientation according to negation markers.

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
	gabe	[NP/VP/clause +] *gabe*	41
7	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*$_{(+)}$ <u>*ez*</u> *dira nahikoak*$_{(+)}$ *izaten literatura ona*$_{(+)}$ *sortu ahal izateko.* BER01
good$_{(+)}$ goals$_{(+)}$ are <u>not</u> 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).

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-

[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.

[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

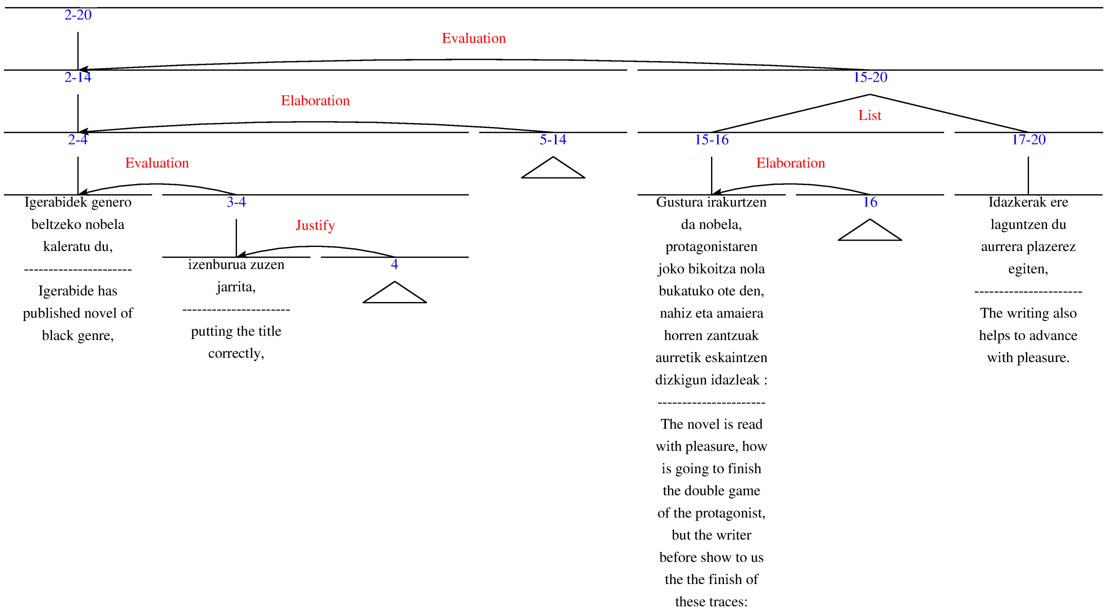
[12] All the texts are available in *Kritiken Hemeroteka* at http://kritikak.armiarma.eus/

[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


2-20
Evaluation
2-14
15-20
Elaboration
List
2-4
5-14
15-16
17-20
Evaluation
Elaboration
3-4
16
Justify
4
Igerabidek genero beltzeko nobela kaleratu du,
Igerabide has published novel of black genre,
izenburua zuzen jarrita,
putting the title correctly,
Gustura irakurtzen da nobela, protagonistaren joko bikoitza nola bukatuko ote den, nahiz eta amaiera horren zantzuak aurretik eskaintzen dizkigun idazleak :
The novel is read with pleasure, how is going to finish the double game of the protagonist, but the writer before show to us the the finish of these traces:
Idazkerak ere laguntzen du aurrera plazerez egiten,
The writing also helps to advance with pleasure.

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.

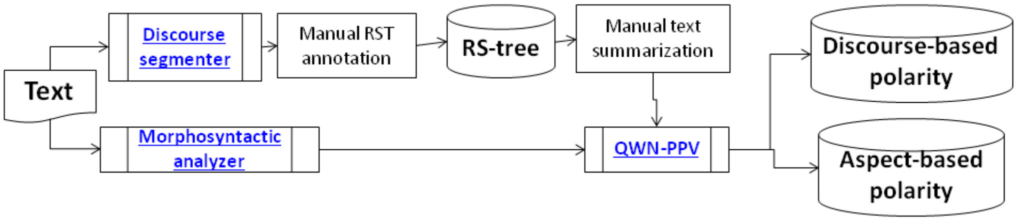
4 Results

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 intrasentence 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. Finegrained 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. ©2017 Association for Computational Linguistics

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.

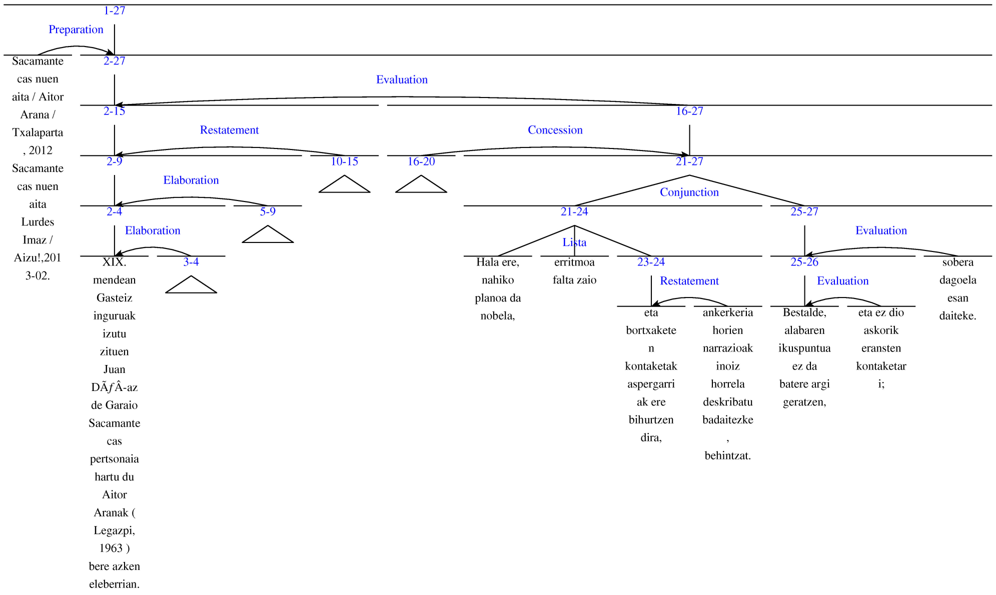

1-27
Preparation
Sacamante cas nuen aita / Aitor Arana / Txalaparta , 2012 Sacamante cas nuen aita Lurdes Imaz / Aizu!,201 3-02.
2-27
Evaluation
2-15
Restatement
2-9
Elaboration
2-4
Elaboration
XIX. mendean Gasteiz inguruak izutu zituen Juan DÃƒÂ-az de Garaio Sacamante cas pertsonaia hartu du Aitor Aranak (Legazpi, 1963) bere azken eleberrian.
3-4
5-9
10-15
16-27
Concession
16-20
21-27
Conjunction
21-24
Lista
Hala ere, nahiko planoa da nobela,
erritmoa falta zaio
23-24
Restatement
eta bortxaketen kontaketak aspergarriak ere bihurtzen dira,
ankerkeria horien narrazioak inoiz horrela deskribatu badaitezke , behintzat.
25-27
Evaluation
25-26
Evaluation
Bestalde, alabaren ikuspuntua ez da batere argi geratzen,
eta ez dio askorik eransten kontaketari;
sobera dagoela esan daiteke.

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 <u>ez</u> du arretarik bereganatzen$_{(+4)}$ (...) (LIB18_EVA).
English: (...) the narration also does <u>not</u> 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 relational 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 related 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

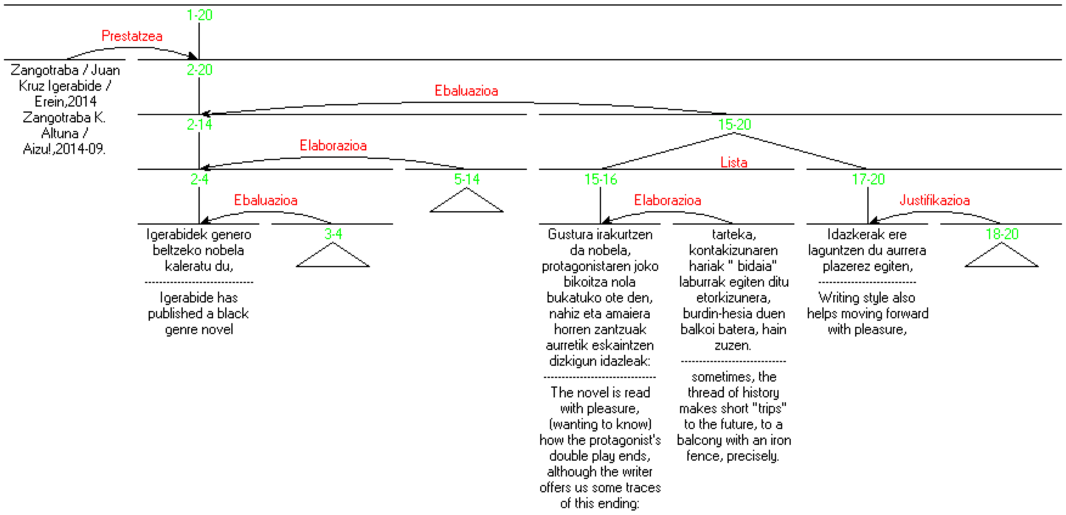

1-20
Prestatzea
Zangotraba / Juan Kruz Igerabide / Erein,2014 Zangotraba K. Altuna / Aizul,2014-09.
2-20
Ebaluazioa
2-14
15-20
Elaborazioa
Lista
2-4
5-14
15-16
17-20
Ebaluazioa
Elaborazioa
Justifikazioa
Igerabidek genero beltzeko nobela kaleratu du,
Igerabide has published a black genre novel
3-4
Gustura irakurtzen da nobela, protagonistaren joko bikoitza nola bukatuko ote den, nahiz eta amaiera horren zantzuak aurretik eskaintzen dizkigun idazleak:
The novel is read with pleasure, (wanting to know) how the protagonist's double play ends, although the writer offers us some traces of this ending:
tarteka, kontakizunaren hariak " bidaia" laburrak egiten ditu etorkizunera, burdin-hesia duen balkoi batera, hain zuzen.
sometimes, the thread of history makes short "trips" to the future, to a balcony with an iron fence, precisely.
Idazkerak ere laguntzen du aurrera plazerez egiten,
Writing style also helps moving forward with pleasure,
18-20

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 annotate 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 PROSAMED (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

Nowadays, there are plenty of tools and dictionaries to analyze the semantic orientation of texts. Most of them follow a bag-of-words approach, and, until now, syntactic or discourse constraints have been out of their interests because it has been assumed that this information was not necessary or it was too complex to process adequately. In order to add contextual discourse information, we need to know how relational discourse structure phenomena affects or modifies the semantic orientation value of text spans, and then add this information to a semantic orientation calculator. The main aim of this work is to study the most relevant features of discourse relations regarding sentiment analysis. We want specifically to identify the discourse structures that coincide more with the semantic orientation of discourse relations and opinion texts. To do so, we have annotated a corpus in two ways: a) the relational structure of texts following Rhetorical Structure Theory (RST) and b) the semantic orientation of the annotated discourse relations. We analyze if the semantic orientation of such discourse relations affects to different features of discourse structure such as semantic orientation, nuclearity, position, relation types and discourse relation distance from central unit. The results suggest that the semantic orientation of the rhetorical relations match more instances with semantic orientation of the nuclei than with the semantic orientation of the satellites and match more cases with the semantic orientation of the last spans (right) of discourse relations than the first spans (left). Moreover, discourse relations near the central unit coincides more with the semantic orientation of texts than those that are far (regardless right or left span). Finally, the results suggest that the discourse relations are not evenly distributed, but rather they tend to appear in specific places in discourse trees and are sensitive to the semantics of coherence relations.

Keywords: corpus, opinion texts, discourse relation, central unit, sentiment analysis, Basque

Introduction

Sentiment analysis aims at extracting subjectivity from language, often expressed as polarity, that is, whether the sentiment 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', that is, unordered sequences of words, without syntactic or discursive constraints. This technique is applied to various types of texts, such as tweets, headlines, news articles, on-line comments, reviews or performance evaluations, to mention some of them. In Example (1), a Basque sentiment lexicon (Alkorta, Gojenola, & Iruskieta, 2018) was used to determine the polarity of a sentence.[1]

(1) Irabazi ezinik jarraitzen du Eibarrek. (KIR17)
(The soccer team) Eibar continues without winning.

Using a bag-of-words approach, a sentiment lexicon would assign a sentiment value as presented in Example (2).

(2) Ezin$_0$; Eibar$_0$; Irabazi$_{+2}$; Jarraitu$_0$
Can not$_0$; Eibar$_0$; Win$_{+2}$; Continue$_0$

If we sum up all the values of Example (2), the sentiment value of the sentence would be 0.5 (+2/4 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 more contextual valence shifters in Basque, apart from negation markers that

[1] In the examples, the code in parentheses means the domain (KIR means sport) and the number of opinion text. The opinion texts belongs to (Alkorta, Gojenola, & Iruskieta, 2016). Moreover, the underlined element stands for the semantic orientation shifter, that is to say, the linguistic element that can change the semantic orientation of the words.

goes far from the aim of this work in both directions: towards higher structures (discourse) or towards lower constructions (morphology or even phonology phenomena signaled with some morphemes).

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. We have formulated the following questions considering the most significant discourse structure phenomena in relation with sentiment analysis:

1- In discourse relations, which spans coincides with the semantic orientation of discourse relations? The spans could correspond to either the nucleus, the satellite, the first or the last span.

2- In opinion texts, do the discourse relations coincides more with the semantic orientation of the text when they are situated in some specific positions in the discourse tree? Is there any discourse structure which can be considered as a valence shifter?

3- Regarding hierarchical structure, do the studied discourse relations appear at any distance from the central unit? Is there any pattern of distribution taking into account the type of discourse relations and the semantic orientation?

The work is organized as follows: after presenting related work, we describe data and methodology After that, we presents the results and discussion and, finally, we conclude this study, also proposing directions for future work.

Related work

In recent years, the use of discourse structure in sentiment analysis has increased considerably. Different frameworks have been used to study different elements of discourse structure in sentiment analysis. Moreover, the annotation of subjectivity as well as its evaluation have been different in each of the works.

In discourse structure based sentiment analysis, different types of discourse theories have been used. For example, we could mention Rhetorical Structure Theory (Bhatia et al., 2015; Heerschop et al., 2011; Zirn et al., 2011), and Segmented Discourse Representation Theory (SDRT) (Chardon et al., 2013; Asher et al., 2008).

The corpora used in those works present a broad range of different aspects. The most common type of text corresponds to reviews, although there are other types of texts, such as tweets (Mukherjee & Bhattacharyya, 2012) or news. 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, like (Heerschop et al., 2011) and (Zirn et al., 2011).

With respect to methodology, there is a great amount of variation. The annotation of the corpora and creation of the gold standard have been one of the most common part of the methodology (Chardon et al., 2013). On the other hand, (Heerschop et al., 2011) and (Zirn et al., 2011) use sentiment lexicons to assign the sentiment valence to different features. Several types of discourse information have been used, for instance, discourse reweightening 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).

For evaluation, different aspects have been taken into account. There are works where the evaluation is made measuring the effects of linguistic phenomena. For instance, (Chardon et al., 2013) use a bag-of-segments and discourse information, (Mukherjee & Bhattacharyya, 2012) evaluates a lexicon-based classification with and without discourse information, (Heerschop et al., 2011) evalutes the effect of word position and, finally, (Zirn et al., 2011) evaluates 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).

As far as discourse relations and sentiment in Basque is concerned, there are works about the RST framework (Iruskieta, 2014) and sentiment analysis (San Vicente, Saralegi, & Agerri, 2015; Alkorta et al., 2018), but there are few and preliminary works combining the RST framework and sentiment analysis (Alkorta, Gojenola, Iruskieta, & Pérez, 2015; Alkorta, Gojenola, Iruskieta, & Taboada, 2017).

This work presents similarities in several aspects with (Bhatia et al., 2015) and (Heerschop et al., 2011) in the sense that 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 the results have been improved in comparison with previous works. 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 to sentiment valence or semantic orientation of discourse relations or what the effect of nuclearity is in different text spans.

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.

Authors	Approach	Discourse information	Sentiment information
(Mukherjee & 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, Benamara, Mathieu, Popescu, & Asher, 2013)	SDRT	Partial discourse information, Full discourse information	Semantic orientation at discourse unit and document level
(Bhatia, Ji, & Eisenstein, 2015)	RST	Discourse unit position, Recursive neural networks, RST parser	Semantic orientation at discourse unit and document level
(Asher, Benamara, & Mathieu, 2008)	SDRT	Relations: CONTRAST, CORRECTION, SUPPORT, RESULTS and CONTINUATION	Four groups: reporting, judgment, advise and sentiment
(Somasundaran, Namata, Wiebe, & Getoor, 2009)	Dialog Act	Types of frame relations (reinforcing and non-reinforcing relations)	Polarity (positive, negative, neutral) of dialogue units
(Taboada, Voll, & Brooke, 2008)	RST	Nuclei, Topic sentences	Semantic orientation of nuclei and texts
(Zirn, Niepert, Stuckenschmidt, & Strube, 2011)	RST	Relations: CONTRAST (CONCESSION and CONTRAST), NON-CONTRAST.	Semantic orientation of discourse relations
(Zhou, Li, Gao, Wei, & Wong, 2011)	-	A set of cue-phrase-based patterns	-

Table 1
Previous work on discourse-based sentiment analysis.

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.

Firstly, we decided to extract the discourse relations of 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 motive behind this decision was that the selected discourse relations were more likely to have subjective content in comparison with others. Moreover, (Trnavac, Das, & Taboada, 2016) shows that CONCESSION, ELABORATION, EVALUATION, EVIDENCE and RESTATEMENT most frequently intensify the polarity of the opinion words. 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.** Even though all the semantic orientation of discourse relations and their spans has been annotated by one person; Cohen's kappa coefficient between two linguists was calculated taking a subset of the corpus.

[2] `http://ixa2.si.ehu.es/diskurtsoa/index.php`.

[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.

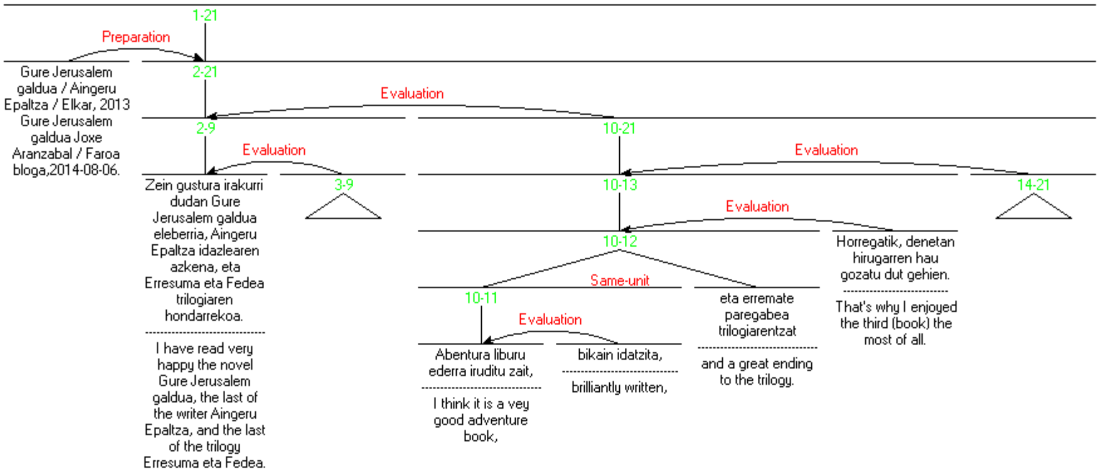

1-21
Preparation
Gure Jerusalem galdua / Aingeru Epaltza / Elkar, 2013 Gure Jerusalem galdua Joxe Aranzabal / Faroa bloga,2014-08-06.
2-21
Evaluation
2-9
10-21
Evaluation
Evaluation
Zein gustura irakurri dudan Gure Jerusalem galdua eleberria, Aingeru Epaltza idazlearen azkena, eta Erresuma eta Fedea trilogiaren hondarrekoa.
I have read very happy the novel Gure Jerusalem galdua, the last of the writer Aingeru Epaltza, and the last of the trilogy Erresuma eta Fedea.
3-9
10-13
14-21
Evaluation
10-12
Horregatik, denetan hirugarren hau gozatu dut gehien.
That's why I enjoyed the third (book) the most of all.
Same-unit
10-11
eta erremate paregabea trilogiarentzat
and a great ending to the trilogy.
Evaluation
Abentura liburu ederra iruditu zait,
I think it is a vey good adventure book,
bikain idatzita,
brilliantly written,

Figure 1. RST-tree of the text SENTFAR-01.

Two linguists had to assign semantic orientation to discourse relations and their spans. So, each annotator had to assign three semantic orientation for each discourse relation: i) the discourse relation as a whole, ii) nucleus, and iii) satellite (except in multinuclear relations). In total, 40% of the corpus has been annotated by both linguists. The measure of inter-annotator agreement can be considered satisfactory, giving a Cohen's kappa coefficient of 0.58, that is considered as "moderate agreement" according to (Landis & 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, the contingency table shows that the biggest differences are related to the neutral semantic orientation. When one annotator (A2) has assigned a neutral semantic orientation, the other annotator (A1) has assigned a positive or negative one. This occurs in 120 instances. The opposite case has also happened but with less frequency (98 instances).

3) **Analysis of parameters.** We use Figure 1 and the EVALUATION relation (EDU 10-13) as 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 EDU of discourse relations is more important than the other. The most important EDU is the nucleus and the satellite EDU is less relevant. They show two different patterns: Nucleus-Satellite or Satellite-Nucleus. On the other hand, there are multinuclear relations where two EDUs have the same importance. In Figure 1, the span 11 is linked with a EVALUATION relation to the span 10 and follows the N(ucleus)-S(satellite) 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. Firstly, 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 (10-12 and 13). Then, we assigned the semantic orientation to the discourse relation (positive).

iii) Distance to central unit. We measured the distance between discourse relations and central

[4] We assigned three types of semantic orientation to discourse units and discourse relations: POS, positive; NEG, negative and NEU, neutral.

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 the span 13 and the 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 had been studied.

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 will present the results trying to address our main research questions. Firstly, we give the results concerning nuclearity and sentiment analysis, and, then, we approach the semantic orientation in texts and, finally, we discuss the position of discourse relations in opinion texts.

Semantic orientation and nuclearity in discourse relations

Table 4 presents the results obtained from the comparison of the semantic orientation in various discourse phenomena: *i*) Relation and nuclearity: to know if the semantic orientation of the rhetorical relation fits more with satellites or with nuclei. *ii*) Relation and position: to know if the semantic orientation of the rhetorical relation fits more with the semantic orientation of the EDU at the right or with the EDU at the left. 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 match of the rhetorical relation is bigger in satellites than in nuclei. Here, the semantic orientation in nucleus and the overall SO of the rhetorical relation match in 0.69 while for the satellite it is 0.82. On the other hand, the CONCESSION group shows the biggest difference between nuclei and satellites regarding score, with the match of 0.70 for nuclei, and with the match of 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
Comparison of the semantic orientation of discourse relations and nuclearity.

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
The semantic orientation of discourse relations and position (first or last span).

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 in more occasion with the first span. 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. On the other hand, the highest difference in score appears in the CONDITIONAL relation, because the first span has the same semantic orientation in 0.28 cases and 0.89 with the last span. Finally, the last spans of EVALUATION and CAUSE fits in more occasions (0.83) with the semantic orientation of the discourse relations.

From our point of view, regarding the match of semantic orientation in discourse relations, the results show that the semantic orientation of nuclei fits more with the semantic orientation of all the discourse relations. This is understandable because, according to RST framework, nucleus is the most important EDU of the relation and, consequently, it must be in accordance with what is said in the text. However, this does not happen with all types of relations. For example, in the EVALUATION group, the semantic orientation of satellite fits in more occasions. This happens 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. Therefore, this results is also understandable.

Comparison of the semantic orientation of the text and discourse relation's distance from the central unit

Another aspect to take into consideration 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. Figure 2 shows that the match of semantic orientation between opinion texts and its 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].

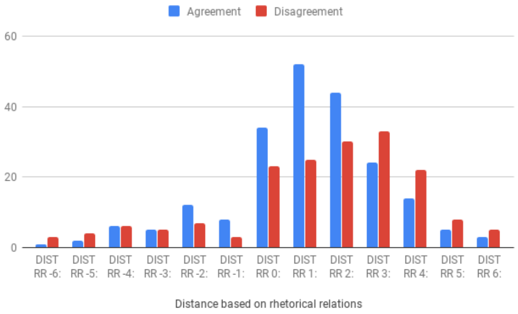


Figure 2. Semantic orientation of the text and discourse relations according to distance from the central unit.

If the results are analyzed for each discourse relation, there are some differences. In the CAUSE discourse relation, the instances that has 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 same semantic orientation between discourse relations and texts based on position is that when one discourse relation is closer to the central unit, the same semantic orientation between that discourse relation and all the text is higher. In our opinion, the reason for this situation is the following: the topic of discourse relations that are closer to the central unit has more similarity with the global topic of the text. For that reason, the point of view (opinion) about the same topic is more similar and consequently, the same semantic orientation between the discourse relations that are closer to the central unit and the global semantic orientation of the text is higher.

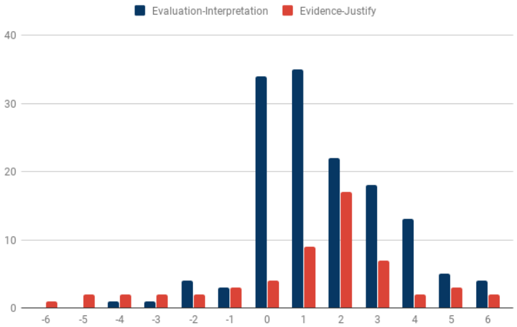


Figure 3. The most usual position of EVALUATION-INTERPRETATION and EVIDENCE-JUSTIFY discourse relations in opinion texts.

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-RESULTS-PURPOSE discourse relations. These discourse relations share some similarities regarding their distance from the central unit. As it happened with the EVIDENCE-JUSTIFY

[5] Distances with the signal − are situated to the left of the central unit. In contrast, distances with the signal + are situated to the right of the central unit.

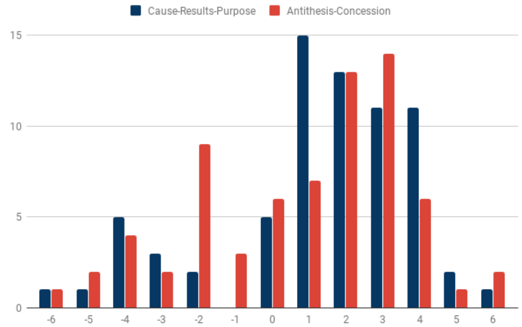


Figure 4. The most usual position of ANTITHESIS-CONCESSION and CAUSE-RESULTS-PURPOSE discourse relations in opinion texts.

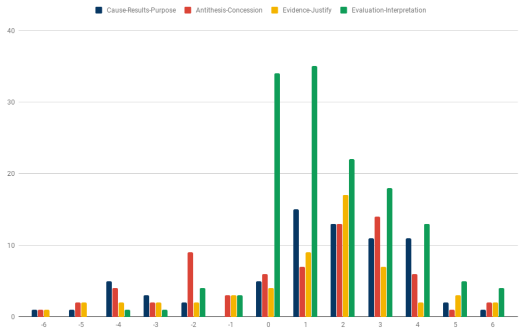


Figure 5. Instances of discourse relations based on distance from central unit.

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 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. The aim is to identify how the semantic orientation is affected by discourse relations and concepts related to it.

Firstly, we have created a subcorpus with some of the discourse relations. Then, we have assigned the semantic orientation to all the discourse relations and their spans. The inter-annotator agreement has been calculated in order to certify the quality of the annotation. Moreover, the distance of each discourse relation from the central unit has been annotated. After that, several features have been measured in order to check the hypothesis proposed. Finally, the results have been analyzed in order to search for the effects of discourse structure and discourse relations on semantic orientation.

The study has allowed to reach those findings:

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 in different type of discourse relations has been irregular. Due to this situation, CONDITIONAL (18 instances) and ENABLEMENT/MOTIVATION (6 instances) relations have few instances, difficulting a better analysis as far as these type of relations are concerned. On the other hand, the f-measure of the inter-annotator agreement as for the semantic orientation of discourse relations and its spans is 0.58 which is considered "moderate agreement" according to (Landis & Koch, 1977).

In the near future, we we want to meet several objectives. Firstly, we want to increase the instances of discourse relations in order to achieve more precise results. On the other hand, we want 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 automatize these findings and add different pipe lines to analyze the semantic orientation taking into account these discourse phenomena, following the subsequent tasks: *i*) determine the EDUs and the CU of the text with the CU detector (Atutxa, Bengoetxea, Diaz de Ilarraza, & Iruskieta, 2019), *ii*) after that, add discourse relations with EusDisParser, the Basque RST parser developed by (Iruskieta & 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

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).

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.

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., 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.

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: Semi-manually 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). Euskal 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.