

# 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

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 *AlloCiné.fr* and *ii*) news reaction from *Lemonde.fr*. They collect 211 texts, annotated at EDU and document level. At the EDU level, subjectivity is annotated while at the document level, subjectivity and discourse relations are annotated. Results in subjectivity show that, at EDU level, Cohen's Kappa varies between 0.69 and 0.44 depending on the corpus and, at the document level, Kappa is between 0.73 and 0.58, respectively. They do not give results regarding the annotation of discourse relations.

Asher et al. (2009) create a corpus with discourse and subjectivity annotation. They categorize opinions in four groups (REPORTING, JUDGMENT, ADVISE and SENTIMENT), using SDRT as the annotation framework for discourse. Exactly, they use five types of rhetorical relations (CONTRAST/CORRECTION, EXPLANATION, RESULT and CONTINUATION). They collect three corpora (movie reviews, letters and news reports) in English and French. 150 texts are in French and 186 texts in English. According to Kappa measure, in opinion categorization, the inter-annotator agreement is 95% while in discourse segmentation it is 82%.

Mittal et al. (2013) follow a similar methodology. By the annotation of negation and discourse relations in a corpus, they measure the improvement made in sentiment classification. They collect 662 reviews in Hindi from review websites (380 with a positive opinion and 282 with a negative one). Regarding discourse, they annotate violating expectation conjunctions that oppose or refute the current discourse segment. According to their results, after implementing negation and discourse information to HindiSentiWordNet (HSWN), the accuracy of the tool increases from 50.45 to 80.21. They do not mention the inter-annotating agreement of violating expectation conjunctions.

To sum up, this section gives us a general overview about discourse-based annotated corpora for sentiment analysis. Corpora have been made for specific aims, annotating only some characteristics or features related to discourse and discourse relations. This situation differs from our work, because our work describes the annotation process of the relational discourse structure and how the function in the rhetorical relation affect to the analysis in the semantic orientation.

### 3 Theoretical framework and methodology

#### 3.1 Theoretical framework: Rhetorical Structure Theory

We have annotated the opinion text corpus using the principles of *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1988; Taboada and Mann, 2006), as it is the most used framework in the annotation of discourse structure and coherence relations in Basque where there are some tools (Iruskieta et al., 2013, 2015b) to study rhetorical relations. According to this framework, a text is coherent when it can be represented in one discourse-tree (RS-tree). In a discourse-tree, there are elementary discourse units (EDU) that are interrelated. The relations are called *coherence relations* and the sum of these coherence relations forms a discourse-tree. Moreover, the text spans present in a discourse relation may enter into new relations, so relations can form compound and recursive structures.

Elementary discourse units are text spans that usually contain a verb, except in some specific situations. The union of two or more EDUs creates a coherence relation. There are initially 25 types of coherence relations in RST. In some cases, one EDU is more important than other one and, in this case, the most important EDU in the relation is called *nucleus-unit* (basic information) while the less important or the auxiliary EDU is called *satellite-unit* (additional information). Coherence relations of this type are called *hypotactic relations*. In contrast, in other relations, EDUs have the same importance and, consequently, all of them are nucleus. The relations with EDUs of same rank are called *paratactic relations*. The task that selects the nucleus in a relation is called *nuclearity*.

Hypotactic relations are also divided into two groups according to their effect on the reader. Some relations are *subject matter* and they are re-

lated to the content of text spans. For example, CAUSE, CONDITION or SUMMARY are subject matter relations. On the other hand, the aim of other relations is to create some effect on the reader. They are more rhetorical in their way of functioning. EVIDENCE, ANTITHESIS or MOTIVATION belong to this group.

Figure 1 presents a partial discourse-tree of an opinion text (tagged with the code LIB29). The text is segmented and each text span is a discourse unit (EDU). The discourse units are linked by different types of rhetorical relations. For instance, the EDUs numbered with 15 and 16 are linked by an ELABORATION relation and the EDUs ranging from 15 to 20 are linked by LIST (multinuclear relation). On the other hand, the EDU numbered 2 is the central unit of this text because other relations in the text are linked to it and this text span is not attached to another one (with the exception of multinuclear relations).

According to Taboada and Stede (2009), there are three steps in RST-based text annotation:

- 1- Segmentation of the text in text spans. Spans are usually clauses.
- 2- Examination of clear relations between the units. If there is a clear relation, then mark it. If not, the unit belongs to a higher-level relation. In other words, the text span is part of a larger unit.
- 3- Continue linking the relations until all the EDUs belong to one relation.

Following Iruskieta et al. (2014) we think that it is recommendable, after segmenting the corpus, to identify first the central unit, and then mark the relations between different text spans.

### 3.2 The Basque Opinion Corpus

The corpus used for this study is the *Basque Opinion Corpus* (Alkorta et al., 2016). This corpus has been created with 240 opinion texts collected from different websites. Some of them are newspapers (for instance, Berria and Argia) while others are specialized websites (for example, Zinea for movies and Kritiken Hemeroteka for literature).

The corpus is multidomain and, in total, there are opinion texts of six different domains: sports, politics, music, movies, literature books and weather. The corpus is doubly balanced. That

is, each domain has the same quantity of opinion texts (40 per domain) and each semantic orientation (positive or negative subjectivity) has the same quantity of opinion texts per each domain (20 positive and 20 negative texts per domain). We extract preliminary corpus information using the morphosyntactical analysis tool Anahitza (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
<b>Movie</b>	21 + 9	9	30
<b>Weather</b>	10 + 5	5	15
<b>Literature</b>	5	20 + 5	25
<b>Total</b>	50	39	70

Table 1: Number of texts annotated by two annotators. The number after the sum sign indicates the quantity of texts with double annotation.

**1- Limiting the annotating work.** Annotating 240 texts needs a lot of work and time. For that reason, we have thought to annotate some part of the corpus initially and, if the results of the annotation are acceptable, continue with the work. Taking into account the previously described data, both annotators have worked with 70 texts (29.16%) of three different domains. 21 texts from the movie domain have been annotated by one annotator and other 9 texts have been annotated by the two annotators. 10 texts from weather have been annotated once and other 5 texts of the same domain by two annotators. Finally, 25 texts

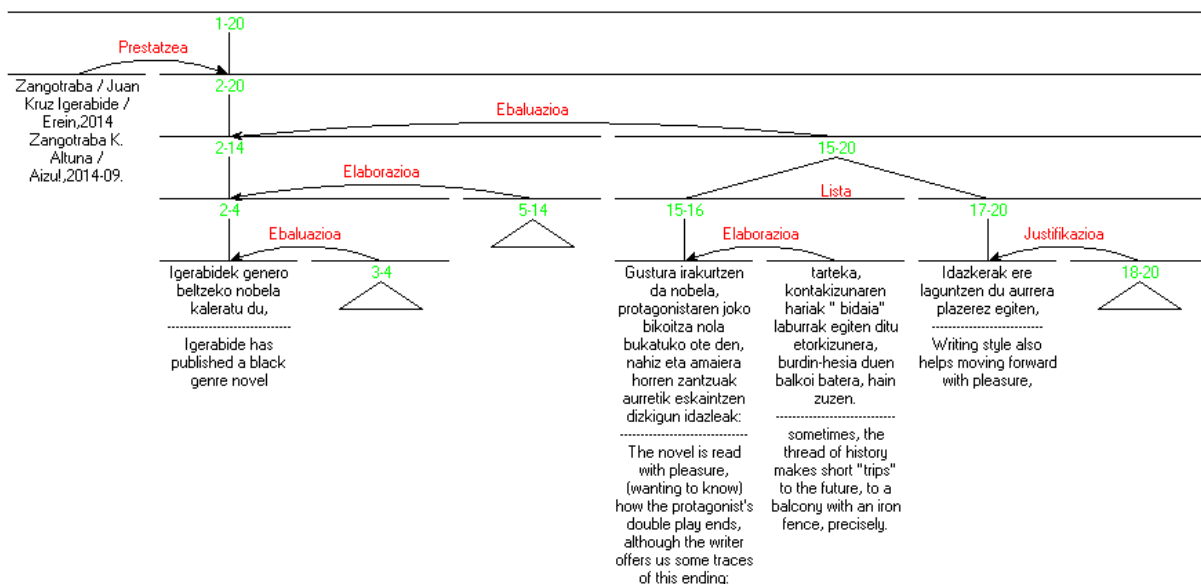


Figure 1: Part of a discourse-tree of the LIB29 review annotated with the RST framework.

of literature reviews have been annotated by one annotator and other 5 texts from the same domain by two. In total, 19 texts from 70 (27.14%) have been annotated by two annotators.

**2- Annotation procedure and process.** We decided to follow the annotation guidelines proposed by Das and Taboada (2018). Each person annotated four or five texts per day during two or three weeks. The time to annotate documents varied according to the domain. The texts corresponding to the weather domain are shorter and, consequently, easier to annotate while texts about movies as well as those of the literature domain are more difficult because their writing style is more implicit (less indicators and relation signals) and complex (longer at least). Approximately, each weather text was annotated in 20 minutes while movie and literature texts were annotated in one hour.

**3- Measurement of inter-annotator agreement.** In order to check the quality of the annotation process, inter-annotator agreement was measured. This was calculated manually following the qualitative evaluation method (Iruskieta et al., 2015a) using F-measure. In this measurement, in contrast with the automatic tool, the central subconstituent factor was not taken into account.

**4- Semantic orientation extraction.** Using the Basque version of the SO-CAL tool (Taboada et al., 2011), we have extracted the subjective information of rhetorical relations in the three domains of the corpus in order to check how the type

of rhetorical relation affects their sentiment valence. SO-CAL needs a sentiment lexicon where words have a sentiment valence between  $-5$  and  $+5$ . The Basque version of the sentiment lexicon contains 1,237 entries.

We have extracted the sentiment valence of 75 instances of 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
<b>Total</b>	<b>39.81</b>	<b>170 of 427</b>

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<sup>1</sup> in three domains. The results<sup>2</sup> 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<sup>3</sup>.

This information contrast between discourse

<sup>1</sup>We decide to choose these rhetorical relations, because we think they are more related to opinions and emotions.

<sup>2</sup>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.

<sup>3</sup>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.

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,]<sub>-1.5</sub> N[film erakargarri eta berezia da Victoria.]]<sub>+6</sub>]<sub>+4.5</sub> (ZIN19)  
*[S[Although it has weak points,]<sub>-1.5</sub> N[Victoria is an entertaining and special movie.]]<sub>+6</sub>]<sub>+4.5</sub>*
- (2) [N[Joxek emaztea eta lagunak dauzka,]<sub>-1.5</sub> S[gaizki tratatzen baditu ere.]]<sub>-4.5</sub>]<sub>-2.5</sub> (SENTAIZ02)  
*[N[Joxe has a wife and friends,]<sub>+2</sub> S[although he treats them badly]]<sub>-4.5</sub>]<sub>-2.5</sub>*
- (3) [S[Eta Redmaynen lana oso ona bada ere,]<sub>+1</sub> N[Vikanderrena bikaina da.]]<sub>+5</sub>]<sub>+6</sub> (ZIN15)  
*[S[Although Redmayn's work is very good]]<sub>+1</sub>, N[Vikander's is excellent.]]<sub>+5</sub>]<sub>+6</sub>*

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.

Sum of sentiment valences	CONCESSION		EVALUATION	
	Nucleus	Satellite	Nucleus	Satellite
<b>Weather</b>	39.41	39.75	49.86	33.35
<b>Literature</b>	61.02	68.73	53.13	80.30
<b>Movies</b>	13.98	19.45	26.01	45.58
<b>Total</b>	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,]<sub>0</sub> S[eta apustu ausarta egin du bertan.]]<sub>+3</sub>]<sub>+3</sub> (SENTBER04)  
*[N[Arrate Mardaras has published her first book recently, Pendrive,]<sub>0</sub> S[and she has made a daring bet there.]]<sub>+3</sub>]<sub>+3</sub>*
- (5) [N[Bada, erraz ikusten den filma da “The danish girl”].]<sub>+1</sub> S[Atsegina da, hunkigarria, entretenigarria]]<sub>+6</sub>]<sub>+7</sub> (ZIN15).  
*[N[So, “The danish girl” is a film easy to watch.]]<sub>+1</sub> S[It is nice, touching, entertaining.]]<sub>+6</sub>]<sub>+7</sub>*
- (6) [N[Talde lana izatetik pertsonaia bakar-raren epika izatera pasako da erdialdetik aurrera]]<sub>+0.5</sub> S[eta horretan asko galduko du filmak.]]<sub>-3.9</sub>]<sub>-3.4</sub> (ZIN39)  
*[N[It is going to pass from being team work to epic of one person]]<sub>+0.5</sub> S[and in that, the film will lose a lot.]]<sub>-3.9</sub>]<sub>-3.4</sub>*

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  $\pm 3$  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<sup>4</sup> 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

<sup>4</sup>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.

Domain	Constituent		Attachment		N-S		Relation	
	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
<b>Total</b>	<b>216 of 413</b>	<b>0.52</b>	<b>164 of 413</b>	<b>0.40</b>	<b>274 of 413</b>	<b>0.66</b>	<b>131 of 413</b>	<b>0.32</b>

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	#	Total
RRs			
ELABORATION	MOTIVATION	9	19
ELABORATION	INTERPRETATION	6	
RESULT	ELABORATION	4	
INTERPRETATION	JUSTIFICATION	4	4
CONCESSION	CONTRAST	6	14
EVALUATION	CONTRAST	4	
LIST	CONJUNCTION	4	

Table 5: Disagreement in rhetorical relations.

According to Table 5, ELABORATION has been used by one annotator whereas the other has employed a more informative relation. In two cases, the first annotator (A1) has annotated an EVALUATION relation while the other annotator (A2) has annotated MOTIVATION and INTERPRETATION. In other case, A2 has annotated ELABORATION whereas A1 has tagged RESULT. In total, there are 19 instances in which ELABORATION has been annotated by one of the annotators. Moreover, there are 4 instances of disagreement between INTERPRETATION and JUSTIFICATION. Finally, there are also disagreements in multinuclear relations. While A2 has annotated CONTRAST in 10 relations, A1 has employed CONCESSION and EVALUATION. There are also 4 instances of disagreement between LIST and CONJUNCTION.

Our interpretation of this results is that one annotator (A1) tends to annotate more general rhetorical relations (e. g. ELABORATION) while other annotator (A2) annotates more precise relations. When it comes to multinuclear relations, it seems that A1 annotator has a tendency to not an-

notate multinuclear relations.

## 5.2 Checking the usefulness of the corpus for sentiment analysis

The second aim of this work has been to check the usefulness of the corpus for sentiment analysis. Firstly, the results have shown that in some cases the Basque version of SO-CAL does not assign a suitable semantic orientation to all the rhetorical relations, even when the semantic orientation of EDUs of the relation is correct. This means that the information of rhetorical relations would be needed in order to make a lexicon-based sentiment classification. In other words, this suggests that it would be recommendable to assign weights to EDUs of rhetorical relations to model their effect on sentiment analysis. Each type of rhetorical relation has different characteristics and, consequently, the way to assign weights to EDUs in each relation must be different.

For that reason, we have made a preliminary study with the purpose of checking how different types of rhetorical relations present a semantic orientation and what is the distribution of words with sentiment valence in rhetorical relations. The study of CONCESSION has shown that *i*) the probability of sentiment words appearing in nuclei as well as satellites is similar, and that *ii*) nucleus always prevails over the satellite and, consequently, the semantic orientation of nucleus must be the semantic orientation of all the rhetorical relation. However, the semantic orientation of the satellite must be also taken into consideration in the semantic orientation of all the rhetorical relation. Although comparing with nucleus, satellite has to be less important.

The opposite situation happens in EVALUATION. Here, we can see that words with sentiment valence concentrate more on the satellite while there are fewer words with sentiment valence in the nucleus. That means that the weight must be assigned to the satellite because that part of the relation is more important from the point of view of sentiment analysis.

This interpretation of the results suggests that the Basque Opinion Corpus annotated using RST can be useful for different tasks of sentiment analysis, in fact, the preliminary analysis made with rhetorical relations shows some characteristics and differences that are related to rhetorical relations.

## 6 Conclusion and Future Work

In this work, we have annotated a part of the Basque Opinion Corpus using Rhetorical Structure Theory. Then, we have measured inter-annotator agreement. The manual evaluation of the results shows that the inter-annotator agreement of the type of rhetorical relations is 39.81%. On the other hand, using an automatic tool we have obtained more fine-grained results regarding aspects of relations and attachment, as well as nuclearity, with an inter-annotator agreement higher than 0.5. We have also identified that ELABORATION, EVALUATION and some multinuclear relations show the highest disagreement.

On the other hand, we have also checked the usefulness of this annotated corpus for sentiment analysis and the first results show that it is useful to extract subjectivity information of different rhetorical relations. In CONCESSION relations, the semantic orientation of the nucleus always prevails but the valence of the satellite must also be taken into consideration. In EVALUATION relations, words with sentiment valence concentrate on satellite.

In future, firstly, we plan to build extended annotation guidelines to annotate the corpus with more reliability. This would be the previous step before annotating the entire corpus. On the other hand, we would like to continue analyzing how the subjective information is distributed in relations.

## Acknowledgments

This research is partially supported by a Basque Government scholarship (PRE\_2018\_2\_0033), the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE) project 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. *Linguisticae 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. ANAL-HITZA: 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.