

# The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations

**Mikel Iruskieta**  
Dept. Language and  
Literature Didactics  
mikel.iruskieta@ehu.es

**Arantza Díaz de Ilarraza**  
Dept. Computer Languages  
and Systems  
a.diazdeilarraza@ehu.es

**Mikel Lersundi**  
Dept. Basque Language  
and Communication  
mikel.lersundi@ehu.es

IXA NLP Group, Manuel Lardizabal 1, 48014 Donostia

## Abstract

This article aims to analyze how agreement regarding the central unit (macrostructure) influences agreement when establishing rhetorical relations (microstructure). To do so, the authors conducted an empirical study of abstracts from research articles in three domains (medicine, terminology, and science) in the framework of Rhetorical Structure Theory (RST). The results help to establish a new criteria to be used in RST-based annotation methodology of rhetorical relations. Furthermore, a set of verbs which can be utilized to detect the central unit of abstracts was identified and analyzed with the aim of designing a preliminary study of an automatic system for identifying the central unit in rhetorical structures.

## 1 Credits

This study was carried out within the framework of the following projects: IXA group, Research Group of type A (2010-2015): IT344-10 (Basque Government); SKaTeR: Scenario Knowledge Acquisition by Textual Reading: TIN2012-38584-C06-02 (Spanish Ministry of Economy and Competitiveness); Hibrido Sint: Rule-based and Statistical-based syntactic analyzers. Corpus management in an XML standard based framework: TIN2010-20218 (Spanish Ministry of Science and Innovation); TACARDI: Context-aware Machine Translation Augmented using Dynamic Resources from Internet: TIN2012-38523-C02-01 (Spanish Ministry of Science and Innovation).

## 2 Introduction

One of the biggest challenges in annotating the rhetorical structure of discourse has to do with the reliability of annotation. When two or more individuals annotate a text, discrepancies generally arise as a result of the way each human annotator interprets the text (Taboada and Mann, 2006). Besides, markers specifying the rhetorical relations between discourse units do not always exist (Taboada, 2006). Even if they appear in the text, these markers do not always establish rhetorical relations unequivocally (van Dijk, 1998; Mann and Thompson, 1987). Despite this ambiguity, discourse markers are considered to be a form of linguistic evidence which are used to signal coherence relations and which are useful in detecting certain rhetorical relations (Georg et al., 2009; Iruskieta et al., 2009; Pardo and Nunes, 2004).

In searching for linguistic evidence to determine the rhetorical structure of texts, scholars have analyzed not only discourse markers but also verbs. For example, Pardo and Nunes (2004) first rhetorically annotated their Corpus TCC (a Portuguese corpus containing scientific texts in the computational domain) and then analyzed verbs related to certain rhetorical relations, finding that verbs such as *buscar* ‘search, look for’, *objetivar* ‘objectify, intend’, *pretender* ‘intend, mean’, *procurar* ‘search, look for’, *servir* ‘serve, meet the requirements of’, and *visar* ‘aim, drive’ are related to the PURPOSE relation.

They also found that other rhetorical relations such as CAUSE, EVIDENCE and RESULT are indicated by other types of verbs.

This paper aims to answer the following research questions:

- (i) Does agreement about the central unit affect inter-annotator reliability when annotating rhetorical relations?
- (ii) Are there some types of verbs that can be used as “indicators” (Paice, 1980) to identify the central unit of a rhetorical structure?

Besides we focus on how to identify the unit associated with the main node in the rhetorical structure tree or, in other words, the “central unit” (Stede, 2008), the “central proposition” (Pardo et al., 2003), the “central subconstituent” (Egg and Redeker, 2010) or the “salient unit of the root node” (Marcu, 1999). To our knowledge, no other research has attempted to identify this unit, the central unit of a rhetorical structure tree, by semantically studying the verb within the framework of RST. This topic, however, could have both theoretical and methodological implications.

The structure of the paper is as follows: Section 3 describes the theoretical framework, corpus and methodology utilized in this study. Section 4 lays out the results obtained. Section 5 presents a preliminary study on the semantic classes of the verbs belonging to central unit. The final section presents conclusions and suggests directions for future research.

### 3 Theory, corpus and methodology

#### 3.1 Theory

Various theories describe the relational structure of a text (Asher and Lascarides, 2003; Grosz and Sidner, 1986; Mann and Thompson, 1987). This study is based on Mann and Thompson’s (1987) Rhetorical Structure Theory (RST), an applied, language-independent theory that describes coherence between text fragments. It combines the idea of nuclearity –that is, the importance of an individual fragment from within the discourse– with the presence of rhetorical relations (RR) (hypotactic and paratactic relations) between these fragments. Mann and Thompson (1987) argue that nuclear units play a more important role for text coherence than satellites.

This has significant implications for automatic text summarization. Ono et al. (1994) and Rino and Scott (1996) suggest that the summary of a text can be obtained by deleting optional satellites, an argument based on the property of nuclearity in hypotactic relations. Da Cunha (2008) describes rules based on nuclearity which can be used to summarize medical texts. For a more in-depth, critical explanation of nuclearity, see Stede (2008) and for additional information on RST, see Taboada and Mann (2006) and Mann and Taboada (2010).

According to RST, hypotactic and paratactic relations connect elementary discourse units (EDUs) or groups of discourse units (span). Elementary units cannot be divided into simpler units. In this paper, a “central unit” is defined as the clause which best expresses the topic or main idea of the text. The central unit of a rhetorical structure tree is the elementary unit or group of elementary units which comprise the nucleus of its main node. Hypotactic units have a single nucleus in the central unit, while paratactic units contain multiple nuclei.

For example,<sup>1</sup> in the rhetorical structure tree presented in Figure 1, unit 7 is the central unit of the elementary units that are numbered from 1 to 7, since it is the nuclear unit of the root node which and has the relation PREPARATION associated to it. The root node covers the entire structure of the text, and since it is not linked to any other unit, no other associated nuclei have the same degree of central importance (Marcu, 1999). The central unit indicates the most important unit in the structure, which is indicated in Figure 1 by the verb *analizatzzen* ‘analyze’.

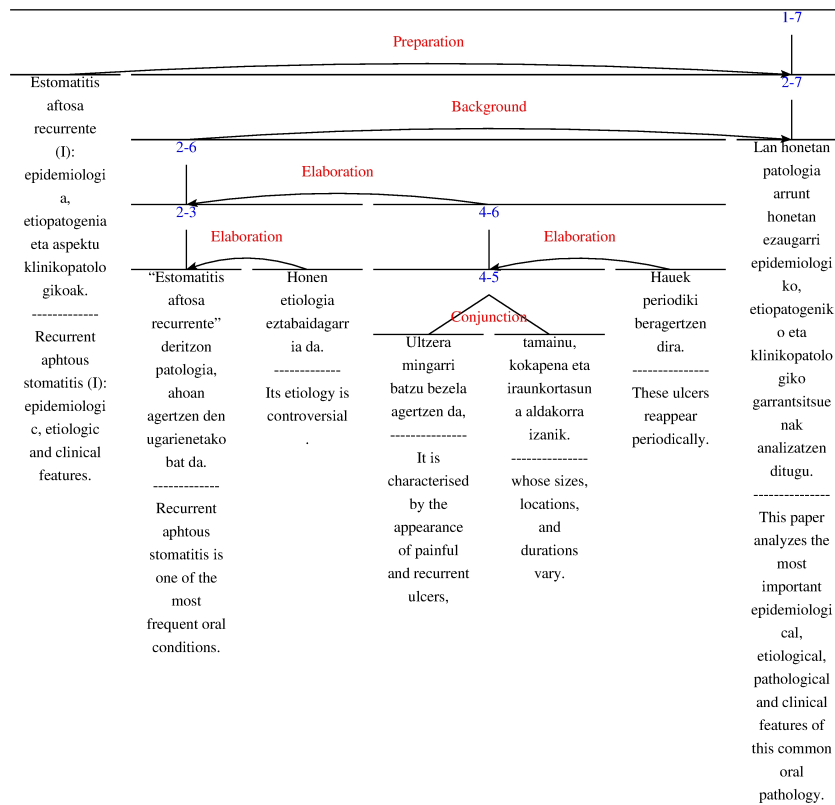
Determining nuclearity (that is, deciding which of the two associated spans has a more central role based on the intentions of the writer) is key in assigning rhetorical relations. In fact, Stede (2008) has questioned the way in which rhetorical structure is represented in RST based on several reasons:

- i) It is not clear what grounds are used to make the decision: is it because of nuclearity or because of the effect of a rhetorical relation?

---

<sup>1</sup>Examples are extracted from the Basque corpus used in this study (Iruskieta, 2014).

Figure 1: A rhetorical structure tree for text GMB0301 (Annotator 1)



ii) Nuclearity poses challenges for annotation. This led Carlson et al. (2001) to present multi-nuclear versions of some nuclear relations from the classic extended classification.

We also identified the same problems. Examples (1) and (2) demonstrate how different choices of nuclearity affect agreement in rhetorical relations.

- (1) [Emaizta:]<sub>1</sub> [Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (% 51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (% 50,5).]<sub>2</sub> GMB0401  
 [Results:]<sub>1</sub> [The average user is as follows: male (51.4%), middle-aged (43.2 years old), and treated for trauma (50.5%).]<sub>2</sub>

Annotator 1 (A1) decides that the second unit in Example (1) is more important than the first unit. The second annotator (A2), however, makes the exact opposite decision. Both annotators arrive reach their conclusions based on structural reasons. Disagreements about the importance of each text fragment influence the rhetorical relation: A1 annotates the relation as PREPARATION while A2 chooses to label the relation as ELABORATION.

Example (2) demonstrates how different interpretations of nuclearity affect agreement with regard to the rhetorical relation.

- (2) [Erabiltzaileen % 80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea]<sub>1</sub> [eta kontsulta hauen % 70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek].<sub>2</sub> GMB0401  
 [It is calculated that about 80% of users come to emergency services on their own initiative]<sub>1</sub>  
 [and that 70% of visits are considered minor by health care personnel].<sub>2</sub>

A1 believes that the second unit in Example (2) provides more detailed characteristics about the users (e.g. the second unit is a satellite of the first unit) and therefore annotates the relation as hypotactic

(ELABORATION). A2, on the other hand, annotates the same discourse segment as a paratactic relation (CONJUNCTION), considering the marker *eta* ‘and’ to be the most significant element, indicating that she or he believes that two different elements of emergency services are being discussed.

According to Bateman and Rondhuis (1997), when determining nuclearity at the higher levels of a tree structure, RST clearly establishes a global view of a text, since an analysis is by definition incomplete until all units in the text have a function which is depicted by a single structure. It is logical that if nuclearity plays a role in determining rhetorical relations at the lower levels of a rhetorical structure, it will also affect the structure’s higher levels. If two annotators have a different global point of view (e.g. they annotate different central units), they will also annotate different rhetorical relations. Therefore, our hypothesis is that trees which have the same global interpretation of text structure will have greater agreement in the annotation process; i.e., in the labeling of rhetorical relations, while those with differing global structures will have lower agreement. This hypothesis underpins the methodology used to answer the first research question of this study.

The next subsection describes the corpus used for this study.

### 3.2 Corpus

This study sought to analyze short but well structured texts written in Basque in order to determine linguistic evidence which could be used to indicate the central unit of rhetorical structure. The corpus utilized in this study consists of three corpora from the same genre (abstracts) from three different specialized domains: medicine, terminology and science. The communicative goal of these texts is to present specialized knowledge, since both the writer and readers are experts. Medical texts include the abstracts of all medical articles written in Basque in the *Gaceta Médica de Bilbao* (GMB) ‘Medical Journal of Bilbao’ between 2000 and 2008. Terminology texts are abstracts from the proceedings of the *Congreso Internacional de Terminología* (TERM) ‘International Conference on Terminology’ organized by UZEI –the Basque Centre for Terminology– in 1997, while scientific articles are abstracts of papers from the University of the Basque Country’s *Jornadas de Investigación de la Facultad de Ciencia y Tecnología* (ZTF) ‘Research Conference of the Faculty of Science and Technology’, which took place in 2008.

After the annotation process (central unit and rhetorical relations among others), the annotated corpus was evaluated (Iruskieta et al., Forthcoming) and harmonized by a judge (Iruskieta, 2014). The harmonized corpus can be consulted in the RST Basque TreeBank<sup>2</sup> (Iruskieta et al., 2013a).

### 3.3 Methodology

Before presenting the process followed to get our goals, let us explain that, when we began this research, the GMB corpus had previously been annotated manually (Iruskieta et al., 2013b) by two linguists using the extended classification of RST (Mann and Taboada, 2010) while the other two corpora (TERM and ZTF) were not tagged. The results of the comparison done about the relationship of agreement between the annotation of the central unit and the annotation of the rhetorical structure in GMB led us to redefine the annotation strategy for TERM and ZTF in the sense that we asked annotators to identify the central unit (one or more) before tagging the rhetorical structure.

The steps carried out for the annotation of the corpora were the following:

- A. Elementary Discourse Units segmentation. The corpus was segmented at intra-sentential level using a minimal set of criteria (Iruskieta et al., 2011a) by each annotator using the RSTTool program (O’Donnell, 1997)
- B. Central unit identification (TERM and ZTF). Both annotators determined the central unit<sup>3</sup> and the verbs present in the central unit of a scientific abstract in TERM and ZTF domains.<sup>4</sup>

<sup>2</sup>The RST Basque TreeBank is available at <http://ixa2.si.ehu.es/diskurtsoa/en/fitxategiak.php>.

<sup>3</sup>We calculate a baseline to illustrate the complexity of the central unit selection reporting the average number of EDUs: average number of 22.58 EDUs per central unit candidates per text. The average was calculated based on the number of EDUs, over the number of texts.

<sup>4</sup>The central units (CU) can be consulted also in RST Basque TreeBank.

- C. Rhetorical tree structure annotation. Rhetorical relations were annotated by each annotator using the RSTTool program with the extended classification (Mann and Taboada, 2010) of RST.
- D. Evaluation. Agreement in rhetorical tree structures were manually evaluated following the qualitative methodology proposed in Iruskietia et al. (Forthcoming), but taking into account the structures with the same central unit and distinguishing between the rhetorical relations linked or not to central unit.
- E. Interpretation. We compared the results of central unit agreement and disagreements to check for possible correlations using a t-test formula at 99.5% confidence.

## 4 Results

Our main hypothesis is that an agreement on central unit leads us to a higher agreement on rhetorical relations; in other words, identifying the main idea of the text helps the human annotator in the identification of the structure of the text and, therefore, the agreement between annotators is higher.<sup>5</sup>

### 4.1 Correlation between agreement on rhetorical relations and agreement on central unit

The observation made about the GMB, where we argued that annotators agree more on rhetorical relations when they annotated the same central unit, remained after considering results of a more extended corpus with two new corpora (TERM and ZTF) and two additional annotators.

Results confirm this fact even when the difference has been substantially reduced from 0.1497 to 0.0426 when more data (all the corpus) were considered. Table 1 presents the global results of the comparison between the agreement on central unit ('= CU')<sup>6</sup> and mean agreement on rhetorical relations for the corpus as a whole.

	GMB			Corpus		
	= CU	≠ CU	Diff.	= CU	≠ CU	Diff.
<b>Mean</b>	0.7456	0.5959	<b>0.1497</b>	0.5915	0.5489	<b>0.0426</b>
<b>SD</b>	0.1833	0.1749		0.1429	0.1125	

Table 1: Mean agreement (and standard deviation) of the central unit and rhetorical relations

We perform a significant test for the differences. We confirmed that the populations being compared have a normal distribution following the Kolmogorov-Smirnov test (p-value of K-S test was 0.913) and have the same variance (p-value of F-test was 0.063). Therefore, two tail independent samples t-test was used with a 0.013 p-value, denying the null hypothesis.

Other hypothesis and combinations were analyzed with positive results: a significant agreement was observed when we compared agreement in rhetorical relation linked to central unit when annotators tagged the same central unit and when they tagged different central units. It is very difficult to establish which rhetorical relation are linked to central unit when annotators do not tag the same central unit.

### 4.2 Correlation between agreement on rhetorical relations linked or not to central unit

After our main hypothesis was confirmed, we went ahead in the tree structure and we checked whether there is higher agreement in rhetorical relations linked to the central unit (considering the structures where there was agreement in central unit), than in the other relations of the tree structure. For example, in the rhetorical structure tree presented in Figure 1, we consider two relations linked to central unit PREPARATION (1>2-7) and BACKGROUND (2-6>7), while the other four relations are not linked to central unit (ELABORATION (2<3), ELABORATION (2-3<4-6), ELABORATION (4-5<6) and CONJUNCTION (4=5)). Table 2 presents the results of relations linked to central unit with relation not linked to central unit:

In structures with the same central unit we compare between the agreement in rhetorical relations linked to the central unit and all the other relations. Percent agreement is substantially higher when we

<sup>5</sup>The results of all corpora considered indicate that the change in methodology improved central unit agreement between annotators slightly in TERM and ZTF. This highlights the benefits of a first step followed in TERM and ZTF which entails detecting the central unit.

<sup>6</sup>And '≠ CU' for disagreement on central unit.

	GMB			Corpus		
	Linked	Not	Diff.	Linked	Not	Diff.
Mean	0.7454	0.5881	<b>0.1573</b>	0.7179	0.5449	<b>0.1730</b>
SD	0.2695	0.3344		0.2107	0.1850	

Table 2: Comparison between rhetorical relations linked and no-linked to central unit in structures with the same central unit

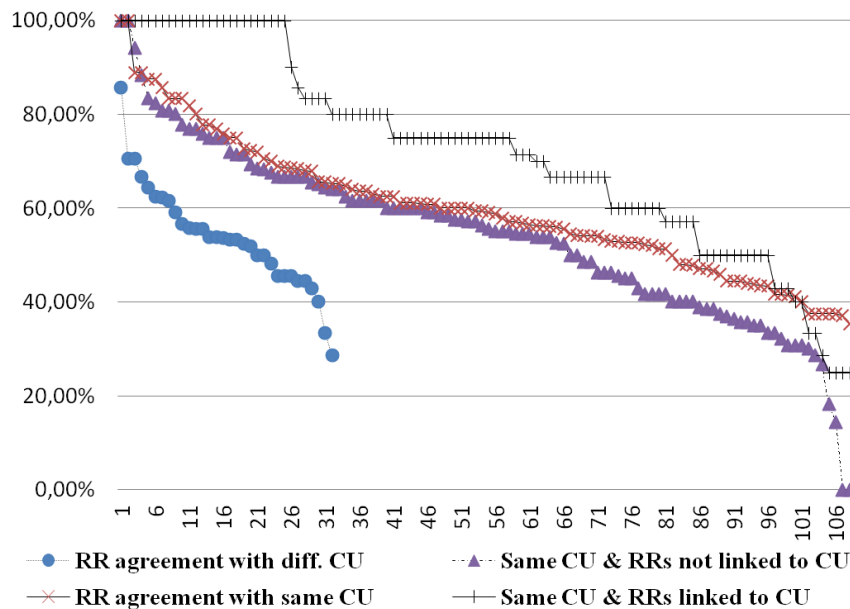
observe the relations linked to the central unit: 17.3% higher than the agreement on the relations that are not linked to the central unit. Populations being compared follow a normal distribution (p-value of K-S test was 0.93) but they do not have the same variance (p-value of F-test is 0.09). The result of the null hypothesis was rejected (p-value of t-test was smaller than 0.001), so we can establish a correlation. The average rhetorical relation agreement on a text according to the central unit, is no different to the average percentage of agreement in the rhetorical relations linked to the UC to those not linked.

### 4.3 Discussion of results

To illustrate the results on agreement (or not) on central unit and average agreement on rhetorical relations linked (or not) from Tables 1 and 2, we present comparisons of the populations in Figure 2:

- When the central unit was the same, the average agreement on relations is represented with red crosses.
- When the central unit was different, the average agreement on relations is represented with blue circles.
- When the central unit was the same and the relations are linked to central unit with black crosses.
- When the central unit was the same and the relations are not linked to central unit with violet triangles.

Figure 2: Representation of mean agreement between RR (vertical) and the number of relations considered in a structure (horizontal) according to the central unit.



These results help to answer the first research question of this study and seem to indicate that there is a correlation between these two kinds of agreement: *i*) greater agreement on detecting the central unit correlates with greater agreement on the annotation of rhetorical relations (results from Table 1 are illustrated in Figure 2 comparing the distance of the red crosses [a] with blue circles [b]), *ii*) also on those which are linked to the central unit (results from Table 2 are illustrated in Figure 2 comparing the distance

of the black crosses [c] with the violet triangles [d]).

This analysis leads to two conclusions:

- i) When considering the methodology for labeling rhetorical structure, annotating the central unit is an important first step before labeling rhetorical relations at least in short texts such as abstracts.
- ii) In Computational Linguistics, a process which helps to automatically identify the central unit is important for determining some restrictions in rhetorical structure mainly determined by the genre/domain structure.

In order to discuss these results, first of all we have to consider that the central unit is a nuclear unit and that relations are linked at various levels (intra-sentential level and inter-sentential level); there are more relations linked at inter-sentential level. For example, in Figure 1 two relations linked to central unit are only at inter-sentential level. This seems to show that these results (rhetorical relations linked to central unit) are not so trivial, since the degree of agreement expected at higher level tree structures is lower. In other words, the agreement at lower levels is higher than in the high level. For example, Marcu and Echihab (2002) argue that automatic annotation of certain rhetorical relations should be addressed first at intra-sentential level because they are less ambiguous. Soricut and Marcu (2003) mention that some of the rhetorical relations are derived from syntactic structures. These results (11.50% higher agreement at intra-sentential level, than at inter-sentential level in the GMB corpus) were confirmed in Basque by Irukieta et al. (2011b).

## 5 Identifying the semantic class of verbs in the central unit

Our final goal is the automatic detection of central unit. To this end, we wanted to find lexical-semantic markers in the central unit<sup>7</sup> in each domain in greater detail. The meanings of the main verbs were analyzed and their semantic class determined as per the SUMO ontology (Niles, 2003). The relation between meaning and semantic class was obtained by means of the MCR semantic database, which includes various lexical-semantic and ontological databases. Data from the GMB, TERM, and ZTF corpora are grouped in Table 3 by semantic classes at the most general level, e.g. “Intentional Psychological Process” (IPP), “Social Interaction” (SI), “Internal Change” (IC) and “Predicate”.

SUMO	SUMO	MCR synset	GMB	TERM	ZTF
IP-IPP	Reasoning	analyze <sub>1</sub> , show <sub>2</sub> , base <sub>1</sub>	<b>0.4615</b>	0.2273	0.0870
	Comparing	value <sub>2</sub> , compare <sub>1</sub>	<b>0.2692</b>		
	Classifying	classify <sub>1</sub>			0.0870
	Learning	review <sub>1</sub>	0.0385		
	Guiding	take <sub>3</sub>		0.0455	
	Process	gain <sub>4</sub>			0.1739
IP-IPP		recognize <sub>2</sub> , determine <sub>8</sub> , hold <sub>6</sub> , focus <sub>1</sub>	0.0385	0.0909	0.0435
IP-SI	Communication	present <sub>2</sub> , address <sub>9</sub> , recount <sub>1</sub> , propose <sub>1</sub>	0.0385	<b>0.4545</b>	0.0435
IP		perform <sub>1</sub> , target <sub>1</sub> , set-up <sub>15</sub> , work <sub>1</sub> , make <sub>3</sub> , use <sub>1</sub>	0.1154	0.0909	0.0870
IP	Searching-Investigating	investigate <sub>1</sub>			0.0435
IP	Organizational Process	serve <sub>2</sub>			0.0435
IC		palliate <sub>2</sub>		0.0455	
Predicate		be <sub>1</sub> , develop <sub>5</sub> , constitute <sub>1</sub> , hold <sub>4</sub>	0.0385	0.0455	0.3913

Table 3: Summary comparison of verbs by domain

The results of this empirical study indicate that each domain tends to use verbs from the same semantic class. For example, in the GMB corpus, the central unit was usually marked with verbs from the IPP category. On the other hand, in the TERM corpus, verbs from the IPP and SI category. Verbs in the central unit of the ZTF corpus are marked with IPP and Predicate class.

Therefore, the results demonstrate that:

<sup>7</sup>Results show that there are multiple EDUs functioning as the central unit of the text in the three corpora: 9 multiple EDU functioning as central unit in GMB, 2 multiple EDUs in TERM and 3 multiple EDUs in ZTF.

- i) A study is needed to identify the SUMO class of the verbs used in a specific domain. For example in our corpus the central unit is indicated with verbs that belong to the IPP class for all three domains. However, other classes also have to be considered, SI for TERM and Predicate for ZTF.
- ii) In the case of weak verbs, other indicators<sup>8</sup> help to identify the central unit. The TERM and ZTF corpora are more marked by noun class indicators than the GMB corpus (Iruskieta, 2014). Another reason is that the direct observation of the central unit makes the central unit selection more consistent. An evidence of that is that all the verbs in central unit are from the same SUMO class in TERM and ZTF corpora by both annotators. Furthermore, it could also be argued that the use of different verbs has to do not only with the field but also with the medium: the GMB corpus derives from texts published in a periodical while the TERM and ZTF corpora include texts published in Conference proceedings. In other words, it could be argued that the medium influences the writing style, and consequently, impacts the verb classes used in the texts. This is in line with the main argument of this study, since different verbs are used to indicate the central unit in the TERM and ZTF corpora, which share the same medium but belong to different fields.

So far, this paper has provided a partial answer to the second research question. However, to automatically detect the central unit by means of verbs (with the help of other types of signals) it is necessary to consider these three issues:

- i) The verb form which is used in the central unit might also be used in non-central units in the rhetorical structure tree.
- ii) Tools which disambiguate the sense of analyzed verbs are necessary in order to know what SUMO class they belong to.<sup>9</sup>
- iii) The central unit is not always indicated with a verb and, therefore, other types of signals (or combinations) can help in the automatic identification of the central unit.

The next phase of this research considered whether verb forms which appear in the central unit unequivocally indicate this unit or whether they can also appear in other types of units. This entailed calculating the frequency with which each studied verb appeared and counting the percentage of appearances which correspond to the central unit.

From the results obtained so far we can't establish any clear tendency but rather some preliminary conclusions that must be ratified with the analysis of more data.

Phenomena related to the central unit appeared in this study of ambiguity:

- i) In GMB corpus verbs that indicate the central unit with a high enough frequency are from IPP category *baloratu* 'value<sub>2</sub>'; there exist other verbs that can be considered but they are not so frequent, e.g. *alderatu* 'compare<sub>1</sub>', *gainbegiratu* 'review<sub>1</sub>', *aztertu* and *analizatu* 'analyze<sub>1</sub>', and *ezagutu* 'recognize<sub>2</sub>'.
- ii) In TERM corpus, the second sense of the verb present in MCR, 'present<sub>2</sub>' (its equivalents in Basque are the verbs *plazaratu*, *aurkeztu*, *aipatu*, *berri eman* and *jardun*), has a high frequency but a high degree of ambiguity. We can't identify the central unit on the basis of its occurrence.
- iii) In the ZTF corpus, the central unit was not always indicated with a verb.

## 6 Conclusions and future research

After considering the relationship between identifying the central unit in a text and annotating its rhetorical structure, it has been demonstrated that a correlation exists between these two tasks, since a greater degree of agreement with regard to the central unit leads to a greater degree of agreement in rhetorical. Besides there is more agreement in rhetorical relations linked to the central units than in relations that are not linked.

This study has investigated verbs which mark the central unit of a rhetorical structure and the correlation of the agreement in central unit with the agreement in rhetorical relations. Its goal has been

<sup>8</sup>According to Paice (1980) indicators can be nouns ('paper', 'method', 'result'), determiners ('this', 'the', 'a') and pronouns ('we', 'I'), among others.

<sup>9</sup>In attempting to automatically detect coherence relations which are not indicated or vaguely indicated using WordNet (Miller et al., 1990) Sporleder and Lascarides (2007) obtained better results using morphological strategies than using semantic generalization strategies. This is due to the fact that, as far as we know, NLP has yet to focus on disambiguating words.



to consider aspects which are relevant for establishing a methodology to help set general criteria for identifying the central unit of texts.

This study also considered which verbs appear in the central units, their semantic classes (according to SUMO categories), and how they identify the central unit. Verbs used to indicate the central units vary in different domains: in the GMB corpus, the central unit was more frequently and the least ambiguously indicated with verbs from the IPP category (SUMO), while in the TERM, SI verbs were most frequent and the least ambiguous.

Testing these results in a larger corpus (and different domains and text structures) could lead to applications for automatic text summarization tasks (classifying clauses), since the central unit is the most important unit in the text.

Furthermore, this study has explained the steps to automatically detect the central unit based on the ambiguity of the verb which marks the central unit. More studies about other indicators (and their combinations) are necessary to automatically detect the central unit.

## References

- [Asher and Lascarides2003] Asher, Nicholas and Alex Lascarides. 2003. *Logics of conversation*. Cambridge Univ Pr, Cambridge.
- [Bateman and Rondhuis1997] Bateman, John A. and Klaas Jan Rondhuis. 1997. Coherence relations: Towards a general specification. *Discourse Processes*, 24(1):3–49.
- [Carlson et al.2001] Carlson, Lynn, Daniel Marcu, and Mary Ellen Okunowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, Denmark, 1-2 September. Association for Computational Linguistics.
- [da Cunha2008] da Cunha, Iria. 2008. Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Phd-thesis, IULA, Universitat Pompeu Fabra.
- [Egg and Redeker2010] Egg, Markus and Gisela Redeker. 2010. How complex is discourse structure? In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, page 16191623, Valletta, Malta, 19-21 May.
- [Georg et al.2009] Georg, Georg, Hugo Hernault, Marc Cavazza, Helmut Prendinger, and Mitsuru Ishizuka. 2009. From rhetorical structures to document structure: shallow pragmatic analysis for document engineering. In *9th ACM symposium on Document engineering*, pages 185–192, Munich, Germany, 16-18 September. ACM.
- [Grosz and Sidner1986] Grosz, Barbara J. and Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.
- [Iruskieta et al.2009] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2009. Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso [Correlations between rhetorical relations and discourse markers]. In *27th AESLA Conference*, pages 963–971, Ciudad Real, Spain.
- [Iruskieta et al.2011a] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2011a. Bases para la implementación de un segmentador discursivo para el euskera [Bases for an Implementation of a Discourse Parser for Basque]. In *Workshop A RST e os Estudos do Texto*, Mato Grosso, Brazil, 24-26 October.
- [Iruskieta et al.2011b] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2011b. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. *Procesamiento del Lenguaje Natural*, 47:144.
- [Iruskieta et al.2013a] Iruskieta, Mikel, Mara Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de la Calle. 2013a. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.
- [Iruskieta et al.2013b] Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2013b. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 0(0):132.
- [Iruskieta et al.Forthcoming] Iruskieta, Mikel, Iria da Cunha, and Maite Taboada. Forthcoming. A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*.

- [Iruskieta2014] Iruskieta, Mikel. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalan (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). Phd-thesis, Euskal Herriko Unibertsitatea, Donostia. [http://ixa2.si.ehu.es/~jibquirm/tesia/tesi\\_txostena.pdf](http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena.pdf).
- [Mann and Taboada2010] Mann, Willian C. and Maite Taboada. 2010. RST web-site. <http://www.sfu.ca/rst/>.
- [Mann and Thompson1987] Mann, Willian C. and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. *Text*, 8(3):243–281.
- [Marcu and Echihiabi2002] Marcu, Daniel and Abdessamad Echihiabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics.
- [Marcu1999] Marcu, Daniel, 1999. *Discourse trees are good indicators of importance in text*, pages 123–136. Advances in Automatic Text Summarization. MIT, Cambridge.
- [Miller et al.1990] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of lexicography*, 3(4):235–244.
- [Niles2003] Niles, Ian. 2003. Mapping WordNet to the SUMO ontology. In *Proceedings of the IEEE International Knowledge Engineering conference*, pages 23–26.
- [O’Donnell1997] O’Donnell, Michael. 1997. RSTTool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, Duisburg, Germany.
- [Ono et al.1994] Ono, Kenjl, Kazuo Sumita, and Seijl Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 344–348. Association for Computational Linguistics.
- [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.
- [Pardo and Nunes2004] Pardo, Thiago A. S. and Maria G. V. Nunes. 2004. Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil [Rhetorical relations and its surface markers: an analysis of scientific texts corpus in Portuguese of Brazil]. Technical Report NILC-TR-04-03.
- [Pardo et al.2003] Pardo, Thiago A. S., Lucia H. M. Rino, and Maria G. V. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, pages 196–196.
- [Rino and Scott1996] Rino, Lucia H. M. and Donia R. Scott. 1996. A discourse model for gist preservation. *Advances in Artificial Intelligence*, pages 131–140.
- [Soricut and Marcu2003] Soricut, R. and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156. Association for Computational Linguistics.
- [Sporleder and Lascarides2007] Sporleder, Caroline and Alex Lascarides. 2007. Exploiting linguistic cues to classify rhetorical relations. In *Recent Advances in Natural Language Processing*, pages 532–539, Borovets, Bulgaria, 27-29 September.
- [Stede2008] Stede, Manfred, 2008. *RST revisited: Disentangling nuclearity*, pages 33–57. ‘Subordination’ versus ‘coordination’ in sentence and text. John Benjamins, Amsterdam and Philadelphia.
- [Taboada and Mann2006] Taboada, Maite and Willian C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.
- [Taboada2006] Taboada, Maite. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592.
- [van Dijk1998] van Dijk, Teun A. 1998. *Texto y contexto: semántica y pragmática del discurso*. Cátedra.