UNIVERSITY OF THE BASQUE COUNTRY
EUSKAL HERRIKO UNIBERTSITATEA
DEPT. BASQUE LANGUAGE AND COMMUNICATION
EUSKAL HIZKUNTZA ETA KOMUNIKAZIO SAILA

The relational discourse structure in pragmatics: description and evaluation in Computational Linguistics

Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa Hizkuntzalaritza Konputazionalean

---

Mikel Iruskieta Quintian

The abbreviated translation of the PhD THESIS

Donostia, February 02, 2014

Abbreviations

A:	Attachment node
AIMRaD:	Abstract, Introduction, Method, Results and Discussion
C:	Composition unit
$\mathbf{C}_{CU}$:	The most important unit in composition unit
D-LATG:	Discourse Lexicalized Tree Adjoining Grammar
DM:	Discourse marker
DRT:	Discourse Representation Theory
DU:	Discourse unit
EDU:	Elementary discourse unit
GMB:	Medical subcorpus
IMRaD:	Introduction, Method, Results and Discussion
MCR:	Multilingual Central Repository
CG:	Constraint Grammar
N:	Nucleus unit
N-N:	Multinuclear relation
N-S:	Nuclear relation
PDTB:	Penn Discourse TreeBank
R:	Rhetorical relation
RST:	Rhetorical Structure Theory
S:	Satellite unit
SDRT:	Segmented Discourse Representation Theory
SUMO:	Suggested Upper Merged Ontology
TERM:	Terminology subcorpus
GS:	Gold standard
UVI:	Unified Verb Index
CU:	Central unit
ZTF:	Scientific subcorpus

Abstract

Written human communications usually consist of more than one sentence, and the coherence relations that exist between these sentences cannot be explained in terms of a successive sequence of phrases (van Dijk 1997). Normally, coherent texts have a structure that is much more complex than mere juxtaposition, providing, of course, that the author wishes to explain him or herself clearly and take into account all the different sides (even the opposing ones) of the issue at hand. This structure is called relational discourse structure, and its description is located within the field of pragmatics known as discourse analysis.

Upon reading works focusing on relational discourse structure, we realize that although a concerted effort has been made by the scientific community to describe the two main phenomena of the main discourse theory (hierarchical structure and the rhetorical relations between text segments), hardly any work has been carried out in this field in relation to the Basque language, and implicit coherence relations have not been taken into account. This thesis-report describes how we annotated scientific abstracts from different domains with the relational discourse structures found in them. It also describes how we overcame the most important problem encountered when annotating texts at this level, namely inter-annotator subjectivity. To this end, we used Rhetorical Structure Theory (RST) (Mann eta Thompson 1987), the most widely accepted theory for describing relational discourse structure phenomena in the field of computational linguistics.

As stated above, for the Basque language, coherence relations have only been partially analyzed to date, with almost all focus being firmly placed on explicit coherence relations. This thesis seeks to redress this situation by describing coherence relations (both explicit and implicit) at different levels (micro-structure and macro-structure), and based on semantic-pragmatic criteria. Moreover, thanks to an innovative annotation method that will also be presented here, the paper's main claim is that inter-annotator subjecti-

vity is not always present to the same degree in the backbone of hierarchical structures, at the different levels of the discourse structure tree or indeed in certain coherence relations between different text segments. To demonstrate this, we propose an innovative qualitative-quantitative relational discourse structure evaluation system. Although we have used this system here to evaluate the reliability of an annotated text in the Basque language, we will also demonstrate that it can be used to compare structures in parallel corpora. Moreover, in order both to avoid circularity problems between rhetorical relations and their signals that may arise as the result of a training phase designed to increase inter-annotator agreement, and to enhance the reliability of discourse structures, we first established the criteria to be followed by the super annotator within RST. The principal outcome of this proposal is a set of characteristics of the first reference corpus in the Basque language annotated with relational discourse structure. We will also outline some innovative search tools to consult the contents of the tagged corpus and will describe the work carried out to disseminate the corpus and make it available to the scientific community at large. The files of the corpus annotated at different levels have been made available to any interested party, in the hope that they will prove useful to certain tasks involved in the processing of the Basque language, including: automatic segmentation, information retrieval, automatic summarization and machine translation, among others.

The addresses of the corpus annotated with relational discourse structure, the electronic version of the thesis in Basque, and the abbreviated translation of the thesis are as follows:

— Annotated corpus:
`http://ixa2.si.ehu.es/diskurtsoa/`

— Thesis-report in Basque:
`http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena.pdf`

— Abbreviated translation of the thesis-report:
`http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena_itzulita.pdf`

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1

Introduction

According to van Dijk (1997), in the mid 1960s, the interaction between new disciplines triggered a major paradigm change in the field of human science. These disciplines included Semiotics, Psycholinguistics, Sociolinguistics, Pragmatics and Discourse Analysis. In relation to Linguistics,

> Linguists were not lagging far behind during the late 1960s, when some of them realized that the use of language obviously was not reduced to the structures of isolated, abstract, invented sentences — as was the case in structural and generative grammars — but needed analyses of structures 'beyond the sentence' and of whole 'texts', for instance to account for anaphora and coherence. Whereas initially still largely within the formal paradigm of 'text grammars', also this linguistic approach soon merged with the other approaches to a more empirical analysis of actual language use. The names associated with these early attempts at text and discourse grammars are János Petöfi (1971), Wolfgang Dressler (1972), and Teun A. van Dijk (1972, 1977), in Europe, and Joseph Grimes (1975), Tom Givón (1979), Sandra Thompson and Bill Mann in the USA, the latter two under the label of *Rhetorical Structure Theory* (Mann & Thompson, 1988).
>
> (van Dijk 1997:6)

Discourse analysis only really started to flourish following the spread of these disciplines. But so often happens in moments such as these, the search for a more precise description of language uncovered a number of gaps and shortcomings in relation to discourse analysis and linguistics. This is simply part of scientific progress. Nevertheless, those working today in the field of linguistics, particularly Computational Linguistics, face a larger problem, namely that posed by the fact that, in certain tasks, more success is obtained using statistical methods than using advanced linguistic theory. According

to Hovy (2011), the main problem in Computational Linguistics today is the need to find ways of processing the large amount of data used.

Initially, Computational Linguistics focused mainly on machine translation, providing word-for-word translations based on literal meanings. The failure of this undertaking prompted Computational Linguistics to explore and compute other areas, such as, for example, the phenomena located outside the sentential field,[1] which are related to the area of pragmatics.

Although there are some theoretical works in Computational Linguistics which focus on the discourse level, one of the distinguishing characters of computation linguistics at this particular level is its operational nature, which ensures successful applications. Indeed, it is having application as its objective that confers upon discourse analysis its operational nature, turning it into a discipline in which theories must be applicable and ongoing evaluation is an absolute necessity. As a result of this operationality, the problems or topics related to the theoretical framework of Computational Linguistics are analyzed from an eminently practical perspective.[2] Since the aim of Computational Linguistics at the discourse level is to describe language phenomena and the relationships that emerge during their use, linguists and information technologists working in this field do so with two objectives in mind:

1) To represent texts (by describing phenomena found at the discourse level).
2) To create texts (by creating text segments containing more than one sentence, based on specific information).

In relation to representation, since the key topic studied is inference,[3] all approaches to discourse analyze inference. The two main phenomena upon which inference is based are:

- The reference structure of the discourse (anaphora and co-reference).
- The relational structure or rhetorical structure of the discourse (coherence relations).

[1] A clause is understood as a collection of words containing a verb and, sometimes, a series of other components which are governed or modified by said verb. Clauses can be combined to form a complex conjunction of clauses. A sentence is understood as a collection of words running from one full stop to the next; or, to be more precise, from one terminal punctuation mark (full stop, question mark, exclamation mark) to another. Thus, a sentence can be made up by a simple or complex clause, but also by a series of words that do not contain a verb.

[2] For further information about Computational Linguistics at the discourse level, see Bunt eta Black (2000) and Jurafsky eta Martin (2000).

[3] According to The Concise Oxford Dictionary of Linguistics inference is: Any conclusion drawn from a set of propositions, from something someone has said, and so on. This includes things that follow logically: cf. implication, entailment. It also includes things that, while not following logically, are implied, in an ordinary sense, e.g. in a specific context: cf. e.g. conversational implicature.

Relational discourse structure (or alternatively rhetorical structure)[4] is the name given to the structure that makes up all the coherence relations of a text. In this thesis, our aim is to establish a methodology for representing rhetorical structure in real texts, in order to obtain a reference corpus annotated with rhetorical structure.

In order to offer an initial approach to rhetorical structure, we will adapt, translate and explain a number of examples from the work of van Dijk (1980b):

(1) John gaixorik dago. Gripea dauka.
John is sick. He has the flu.

(2) Johnek ezin du etorri. Gaixorik dago.
John can't come. He is sick.

(3) Tiketa erosi dut eta nire aulkira joan naiz.
I bought a ticket and went to my seat.

(4) Tiketa erosi dut eta uretara buruz salto egin dut.
I bought a ticket and dived into the water.

(5) Peter zinemara joan zen. Berak begi urdinak ditu.
Peter went to the cinema. He has blue eyes.

In general, we can say that the second sentence of Example (1) provides more detail regarding the first; that in Example (2), the second sentence provides an explanation of what is stated in the first sentence; and that in Example (3), there is a sequence of events between the two clauses.

van Dijk (1980b) clearly highlights the special effort we have to make to understand the rhetorical structure or context for the events outlined in examples (3) and (4). However, to understand the local coherence in these two examples, or in other words the relationship between their clauses: the relationship between buying a ticket and going to one's seat in Example (3) and the relationship between buying a ticket and diving into the water in Example (4), one first needs to understand macro-structure or subject being discussed. In Example (3) the subject under discussion is going to the cinema, whereas in Example (4), it is going to the swimming pool. Thus, there must be coherence between the macro-structure and local coherence (rhetorical relations).

In Example (5), on the other hand, the sentences are not coherent at a local level, even though in both cases we are talking about the same person. Moreover, it is difficult to find a link between the macro-structure and micro-structure.

[4]In the Basque language, the term "proposizio-egitura" (proposition structure) has sometimes been used from another perspective to refer to "erlazio-egitura" (rhetorical structure).

van Dijk (1980b) distinguishes between two different kinds of coherence relations[5] or discourse relations in relational discourse structure: *i*) semantic (Example (1))[6] and *ii*) pragmatic (Example (2)).[7]

However, as with so many other phenomena also, there is no widespread agreement regarding the exact definitions of the different types of discourse relations. For example, Hovy (1993) identifies three types of relation: *i*) ideational or semantic, *ii*) interpersonal or rhetorical and *iii*) textual.

Discourse relations have been analyzed from different perspectives within the theoretical frameworks that focus on the subject of hierarchical discourse structure (Wolf eta Gibson 2004; Asher eta Lascarides 2003; Forbes *et al.* 2003; Moser *et al.* 1996; Polanyi 1988; Mann eta Thompson 1987; Litman eta Allen 1987; Cohen 1987; Grosz eta Sidner 1986; Hobbs 1979). Stedek (2008a) mentions the trends that can be identified in different theories:

a. Those that apply syntactic or semantic theories to the discourse level: i.e. theories that are based on sentences but which offer specific formalization. Only a small number of these analyze real texts. Segmented Discourse Representation Theory (SDRT) (Asher eta Lascarides 2003); Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) (Forbes *et al.* 2003); Linguistic Discourse Model (LDM) (Polanyi 1988).

b. Works based on real data and those that take as many language phenomena as possible into account. These generally tend to have shortcomings in their formalization or somewhat vague relation definitions. The following approaches fall into this category: Rhetorical Structure Theory (RST) (Mann eta Thompson 1987; Carlson eta Marcu 2001), Wolf eta Gibson (2004), and *Penn Discourse TreeBank* (PDTB) (Miltsakaki *et al.* 2004). We will analyze the most important theories of both trends in Chapter 2 of this work, namely SDRT and D-LTAG from the first category and RST from the second.

This thesis forms part of the language processing-related research areas currently being investigated by the IXA group. Our group's ultimate aim is to develop the automatic or semi-automatic systems necessary for the Basque language.

Our research into morphology is now almost fully developed and implemented, but although much work has been carried out in the fields of syntax

[5] According to Wolf and Gibson Wolf eta Gibson (2005), in the approach to representing the information structure of discourse, coherence relations show how the meaning of one discourse segment is related to that of another.

[6] In Example (1), we say the relations are semantic because there is a link (elaboration) between the two situations (being sick and having the flu) expressed.

[7] In Example (2), if the relation between the two situations (not being able to come and being sick) were semantic, then we would merely assume that being sick was the cause for his not coming. But we all know that in our society, being sick (i.e. having the flu) is an accepted justification for not attending a meeting or going to work. Thus, since in this context the coherence relation between the two sentences was established in order to provoke a specific effect in the listener, we say that the relation is pragmatic.

and semantics, there is still a lot more to be done in order to obtain a solid, reliable tool. The field of pragmatics, however, is, in the words of Alegria *et al.* (2011), still a "wild, undeveloped jungle". Thus,

> Hizkuntzaren erabateko ulerkuntza automatikoa oraindik urruti dago. Oraingo ezagutza mugatua da, baina azken urteetan argi frogatu da teknologia ez-oso (ala partzial) hori gauza dela aplikazio praktikoak sortzen. Eta helburu horrekin jarduten dugu Ixa taldean.
>
> (Alegria *et al.* 2011:30)

The goal of gaining a total automatic understanding of a language is still a long way off. However, while current knowledge is limited, in recent years it has been clearly proven that incomplete (or partial) technology is capable of creating practical applications. And it is towards this goal that the IXA group works.

Nevertheless, this "wild, undeveloped jungle" does contain some works of note in the Basque language, including some focusing on Language Processing (Barrutieta *et al.* (2001)) and some on relational discourse structure, as well as others such as those carried out by Euskaltzaindia[8] (1990, 1994, 1999, 2005), Larringan (1995), Salaburu (2012), or Alberdi eta Garcia (2012). We will discuss these works in chapter 2.

In the field of pragmatics, solid tools at different language levels are required for carrying out certain tasks. For example, in this thesis, we used the following tools at the following levels:

- Morphosyntax:
    1. 1. Morphosyntactic analysis was carried out using the MORPHEUS tool.[9]

       In this process, linguistic information is automatically added to all known tokens (both words and punctuation marks). This information is annotated with data relating to category, subcategory, case and other linguistic aspects. The main problem with this process is that some words have more than one analysis; in other words, at this level, analyses are conducted without taking context into account.
    2. We used EUSTAGGER[10] for lemmatization and for identifying syntactical functions (Aduriz *et al.* 2003). Using Constraint Grammar (Karlsson *et al.* 1995), the single, most appropriate option is selected from each set of analyses established in the previous phase for each word, using the information provided by nearby words;

[8]The Basque Language Academy: `http://www.euskaltzaindia.net/index.php?lang=en`.

[9]To try MORPHEUS: `http://ixa2.si.ehu.es/demo/analisianali.jsp`.

[10]To try EUSTAGGER: `http://ixa2.si.ehu.es/demo/analisimorf.jsp`.

and during the same process, syntactical functions are identified using the rules derived from language-based knowledge.

3. The reason for identifying lexical units or collocations (Urizar 2012) of diverse words is to enable analysts to determine those units made up by two or more words, providing of course that they have a fixed composition.

– Syntax:

4. In order to obtain information regarding surface syntax we used IXAti, [11] an instrument which was developed using the rules of Constraint Grammar (Aduriz *et al.* 2004).
5. To annotate texts with syntactic dependency, we based our discourse segmentation process on texts analysed using MALTIXA[12] (Diaz de Ilarraza *et al.* 2005).
6. We adapted an automatic clause-segmentation tool (Arrieta 2010) and used it to segment the discourse.

– Semantics:

7. To identify and classify named entities, we used the EIHERA[13] tool (Alegria *et al.* 2003).
8. For the automatic disambiguation[14] of the meaning of certain words, we used an automatic tool.

Thus, using the tools developed to date by the Ixa group, this thesis aims to redress some of the shortcomings identified in the field of pragmatics, and in doing so, it also aims to lay the groundwork for adding another level of linguistic information to the analysis chain, namely that of rhetorical structure.

As part of this process, our aim is to lay the foundations for the manual annotation of a corpus. Sophisticated language processing tools generally tend to be based on an annotated corpus, and indeed, annotated corpora are vital to the successful completion of many complex linguistic tasks, including: automatic text creation (Bouayad-Agha 2000), automatic text summarization (Marcu 2000b), machine translation (Ghorbel *et al.* 2001), assessment of written texts (Burstein *et al.* 2003), and information retrieval (Haouam eta Marir 2003), among others. In order to run the aforementioned applications, it is first necessary to have a corpus annotated with linguistic information at different levels (including the discourse level). The criteria used for selecting the texts to make up the corpus were as follows: they had to be well-structured, brief and written in more than one language. Consequently, we chose a selection of scientific abstracts from three different domains: me-

[11] To try IXAti: `http://ixa2.si.ehu.es/demo/zatiak.jsp`.

[12] To try MALTIXA: `http://ixa2.si.ehu.es/maltixa/index.jsp`.

[13] To try EIHERA: `http://ixa2.si.ehu.es/demo/entitateak.jsp`.

[14] To see the nouns on which work has been carried out in Euskal WordNet, and the demo for disambiguating nouns manually disambiguated in *EuSemcor*: `http://ixa3.si.ehu.es/wsd-demo/`.

dicine, terminology and science and technology. The corpus is described in more detail in section 2.1.1. In response to the urgent need to develop the field of discourse and as a result of the work carried out so far in this sense, we developed the Basque RST TreeBank, the first Basque corpus annotated with relational discourse structure.

1.1 The aims of the thesis

This thesis has three main aims: *i*) to describe the relational discourse structure of a Basque corpus; *ii*) to establish an annotation method; and *iii*) to provide a linguistic description of the main phenomena that may emerge at the discourse level as a consequence of the analysis of different cases during this process. To this end, we will first analyze the annotation phases and evaluation methods most commonly used today. Secondly, we will identify the shortcomings or problems related to annotation and evaluation and finally, we will propose a series of possible solutions to help avoid these pitfalls. In order to fulfill our aims, we decided to divide and organize the specific areas of our research as follows:

1. General decisions regarding the relational discourse structure annotation process:
    - *a.* to measure the influence of the macro-structure (the main ideas expressed in the text) on the micro-structure (inter-annotator agreement regarding coherence relations) and, if necessary, to propose an annotation phase which takes the macro-structure into account.
    - *b.* to establish the characteristics of the texts required to complete the corpus.
    - *c.* to specify the work to be carried out by the annotators and to prepare the corpus annotation tools in order to avoid circularity between the phases.
    - *d.* to describe the main shortcomings detected in the methods used to measure inter-annotator agreement and to propose a qualitative methodology to redress these problems.
    - *e.* to propose criteria to resolve disagreements between annotators working on the same text, in order to increase the reliability of the relational discourse structure.

2. In relation to discourse segmentation:
    - *a.* to analyze the segmentation proposals made within the theoretical framework and to merge the Basque language clause linkage categories.
    - *b.* following corpus segmentation and the calculation of inter-annotator agreement, to assess the quality of segment annotation in accordance with standard measurements.

   c. to establish a register of problems encountered by annotators during segmentation and points upon which they agreed, making decisions and proposing a set of criteria for Basque language text segmentation.
   d. to create and assess a segmentation tool for segmenting Basque language texts at the discourse level.

3. In relation to macro-structure:

   a. to analyze the characteristics of the central unit that best expresses the core idea of the macro-structure.
   b. to obtain a corpus with a harmonized central unit.
   c. to analyze the correlations which exist between the central unit and the rhetorical relations.
   d. to develop a methodology for selecting the central unit.
   e. to analyze the characteristics of "indicators" (Paice 1980) in order to develop an automatic system for detecting the central unit.

4. As regards the relational discourse structure:

   a. to analyze rhetorical relations in the theoretical framework, and to present definitions and examples for texts written in the Basque language.
   b. following corpus annotation and the calculation of inter-annotator agreement, to measure the quality of the annotations in accordance with standard measurements.
   c. to describe the problems and disagreements experienced by annotators during rhetorical relation annotation, and to propose a methodology to resolve them.
   d. to propose a rhetorical relation signal classification method for the Basque language, to enable "signal" (Taboada eta Das 2013) annotation.
   e. to provide a detailed description of signals in order to automatically detect certain rhetorical relations.
   f. to identify the signals for all rhetorical relations in the corpus.

5. Disseminating the results. The Basque RST TreeBank was established in order to make all tools and resources developed available to the general public.

1.2 Outline of the thesis report

In order to outline how we will fulfill these aims Hovy (2010), we will follow the five-phase scheme described above (theoretical phase, preparation phase, annotation phase, evaluation phase and delivery phase). Each phase

is described in a separate chapter, except for the annotation and the evaluation phases. Since we divided the annotation phase into three sub-phases, three chapters are dedicated to this theme, with the annotation and evaluation phase sub-sections being clearly indicated. Thus, the thesis report is organized as follows:

- Chapter 1 - Introduction. In the first chapter we provide a general introduction to the research theme, explaining what relational discourse structure is and outlining both our motivation and our aims. We also offer an overview of what is presented in each chapter. And finally, we list those publications linked to our research and specify the different sub-sections to which they correspond.

- Chapter 2 - Theoretical phase: This chapter summarizes the principal theories for annotating relational discourse structure at a pragmatic level.

  Firstly, we select and define the specific annotation phenomenon upon which we will focus: relational discourse structure. Next, we present the most important computational theories that describe relational discourse structure, namely: Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory (SDRT) and Discourse Lexicalised Tree Adjoining Grammar (D-LTAG). All theories outlined in this chapter are described at the same level; then, in subsequent phases, we explore the justification for the theoretical framework and the theoretical concepts related to certain annotation themes in more detail. Finally, we finish with a summary of the chapter contents.

- Chapter 3 – Preparation phase: This chapter presents the methodology used for annotating and evaluating relational discourse structure.

  Firstly, we argue our reasons for selecting the theory (RST) used to annotate the corpus with discourse relations, and outline its limits and advantages. Secondly, we describe both the corpus itself (consisting of 60 scientific abstracts) and the annotators (4 linguists with no prior training). Thirdly, we outline the main criteria for resolving problems of circularity between the annotation phases (segmentation, macro-structure, rhetorical structure and signals for rhetorical relations). Circularity problems mainly occur between segmentation and rhetorical relations, and between rhetorical relations and their signals. Fourthly, we describe the main characteristics of the delivery phase, before finishing with a summary of the chapter contents.

- Chapter 4 - Text segmentation (annotation, evaluation and harmonization - phase I).

  Firstly, we describe the basic concepts of discourse segmentation and how we avoided problems of circularity (Matthiessen eta Thompson

1987). Secondly, we outline how we adapted the basics of RST to the Basque language, and we present the record of discourse segmentation cases, ordered in accordance with clause linkage hierarchical downgrading. Thirdly, we explain how we evaluated the segmentation carried out by our annotators. Next, we present the results of the inter-annotator agreement measurements and then describe the work carried out to date in the field of automatic discourse segmentation, before finishing with a summary of the chapter contents.

- Chapter 5 - Identification of the macro-structure (annotation, evaluation and harmonization - phase II).

  Firstly, we explain how we defined and annotated the "central unit", or the most important discourse unit of the tree structure that determines the macro-structure. Secondly, we report inter-annotator agreement in relation to the central unit, and describe the harmonization criteria used. Thirdly, we describe the elements of the verb and noun categories that were used as indicators of the central unit, and calculate their level of ambiguity, before finishing with a summary of the chapter contents.

  In RST, this step in the annotation phase is an innovative proposal. In both Iruskieta *et al.* (Forthcomingb) and this thesis, we outline three advantages for inter-annotator agreement that are offered by the inclusion of this step: *i*) since the macro-structure is selected directly, the inter-annotator agreement in relation to the central unit increases, even though in texts from TERM: the sub-corpus of terminology-related abstracts and ZTF: the sub-corpus of science and technology-related abstracts, the probability of selecting the central unit is smaller than in GMB: the sub-corpus of medicine-related abstracts; *ii*) in rhetorical structures with the same central unit, the inter-annotator agreement for rhetorical relations is higher and statistically significant; *iii*) the average inter-annotator agreement for rhetorical relations linked to the central unit is higher than that for other relations, and is statistically significant.[15]

- Chapter 6 - Relational discourse structure (annotation, evaluation and harmonization - phase III).

  Firstly, we explain what rhetorical relations are and outline some of the problems connected with them. Secondly, we outline the two principal methods for evaluating the rhetorical structures built by annotators: *i*) the quantitative evaluation method proposed by Marcuk (2000a); and *ii*) the qualitative-quantitative evaluation method developed during the work carried out in relation to this thesis (da Cunha eta Iruskieta 2010;

[15]Since the central unit and rhetorical relations are linked phenomena, this question will be presented in its corresponding subsection after we describe rhetorical relations.

Iruskieta *et al.* Forthcominga). We outline the drawbacks of the former and the advantages of the latter. Next, we report the results for inter-annotator agreement in relation to rhetorical relations, and we specify the areas of the tree structure and the relations in which problems were encountered, along with the criteria followed by the super-annotator to resolve disagreements and ensure a harmonized corpus. Before finishing, we present a record of rhetorical relation signals and measure inter-annotator agreement for these signals within the cause subgroup, in order to specify the reliability of the work carried out. We also outline the criteria used by the super-annotator to harmonize the cause subgroup signals. Finally, we provide a summary of the chapter contents.

It is important to note in relation to this phase that the method we propose for evaluating RST rhetorical structures is an innovative one, and that there is no circularity (Spenader eta Lobanova 2009) in the annotation method between rhetorical relations and their signals. Moreover, we propose here, for the first time, the criteria for harmonizing disagreements regarding rhetorical relations.

- Chapter 7 - Delivery phase (The Basque RST TreeBank):

  In order to ensure that the corpus annotated with rhetorical relations at the discourse level proves useful to the scientific community in general, we have applied a number of automatic processes developed by the IXA group and have annotated and disambiguated the corpus at various different levels (morphological, syntactical and discourse) through the web service. Consequently, the searches that can be carried out in the corpus are described, with the aim of enabling contributions and/or criticism from other researchers.

  It is worth mentioning that the work carried out in the delivery phase has moved beyond the limits of other works carried out in other languages within the field of RST.

- Chapter 8 - Conclusions and future work:

  This chapter presents the conclusions of the thesis and lists and highlights its contributions to the field. It also identifies some of the future areas of research opened up by this work.

1.3 Publications

Firstly, Table 1.1 lists the works upon which the drafting of this thesis report was based.

Secondly, we would like to highlight the importance of some of the works stemming from the thesis. For example, da Cunha eta Iruskieta (2010) has

Papers	Section	Theme
Iruskieta (2012)	2.1	Explanation of RST
Iruskieta *et al.* (2011a)	4	Automatic segmentation
Iruskieta *et al.* (Forthcomingb)	5	Central unit
Iruskieta *et al.* (2013b)	6.2.2	The drawbacks of quantitative evaluation
Iruskieta *et al.* (2011b)	6.3.1	Relation and segmentation levels
da Cunha eta Iruskieta (2010)	6.2.3	Qualitative evaluation of relations
Iruskieta *et al.* (Forthcominga)	6.2.3	Qualitative evaluation of relations
Iruskieta *et al.* (2009)	6.5	Discourse markers for rhetorical signals
Iruskieta eta da Cunha (2010b)	6.5	Discourse markers for signals (Spanish and Basque)
Iruskieta *et al.* (2013a)	7	The RST Basque RST *TreeBank*

1.1 taula – Publications linked to the various thesis sections

enjoyed a notable degree of success. In addition to being cited sixteen times, it has also had an impact at an international level, being mentioned in the text presented at the international conference entitled Genre- and Register-related Text and Discourse Features in Multilingual Corpora, organized from 11 to 12 January 2013 by the Linguistic Society of Belgium and the Institut libre Marie Haps in Brussels.[16]

The papers written as the result of this doctoral thesis and those focusing on related themes published in scientific journals are as follows:

- Iruskieta M., Díaz de Ilarraza A., Lersundi M. (Aldizkariren batera bidaltzeko). "Detecting the central unit in rhetorical structure trees: A key step in annotating rhetorical relations". (Iruskieta *et al.* Forthcomingb)

- Iruskieta, M.; Da Cunha, I.; Taboada, M. (LRE aldizkarira bidalita 2013ko ekainean). "A Qualitative Evaluation Method for Rhetorical Relations: An Application to Analyses in English, Spanish and Basque". (Iruskieta *et al.* 2013b)

- Iruskieta M., Díaz de Ilarraza A., Lersundi M. 2013. "RST-based Discourse Annotation for Specialized Medical Texts in Basque". CLLT 0.0: 1–32. (Iruskieta *et al.* 2013b)

- Iruskieta M., Díaz de Ilarraza A., Lersundi M. 2011. "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: 139-143. (Iruskieta *et al.* 2011b)

- Iruskieta, M.; Da Cunha, I. 2010. "El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera-español". Calidoscópio 8(3): 181-202. (Iruskieta eta da Cunha 2010a)

[16] `http://www.mariehaps.be/?id=622`.

– da Cunha, I.; Iruskieta, M. 2010. "Comparing rhetorical structures in different languages: The influence of translation strategies". Discourse Studies 12 (5): 1-36. (da Cunha eta Iruskieta 2010)

The papers written as the result of this doctoral thesis and those focusing on related themes presented at conferences are as follows:

– Iruskieta M., Aranzabe, M.J.; Díaz de Ilarraza A.; Gonzalez, I.; Lersundi M.; Lopez de la Calle, O. 2013. "The RST Basque TreeBank: an online search interface to check rhetorical relations". IV Workshop RST and Discourse Studies. Fortaleza, Brasil, Outubro 21-23. (Iruskieta *et al.* 2013a)

– Iruskieta M., Díaz de Ilarraza A., Lersundi M. 2011. "Bases para la implementación de un segmentador discursivo para el euskera". Anais do III Workshop A RST e os Estudos do Texto: 18-29. Cuiaba, Brasil, Outubro 24-26. (Iruskieta *et al.* 2011a)

– Iruskieta, M. da Cunha, I. 2010. "Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera". Bueno Alonso, J.L., et al. (Eds). 2010: Analizar datos > Describir variación. Vigo: Universidade de Vigo. 146-159. (Iruskieta eta da Cunha 2010b)

– da Cunha I., Iruskieta M. 2009. "La influencia del anotador y las técnicas de traducción en el desarrollo de árboles retóricos. Un estudio en español y euskera". In Proocedings of the 7th Brazilian Symposium in Information and Human Language Technology. pp. 1-21. Sao Carlos, Brasil, September 8-11. (da Cunha eta Iruskieta 2009)

– Iruskieta M., Díaz de Ilarraza A., Lersundi M. 2009. "Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso". Modos y Formas de la Comunicación Humana. In Caballero, R & Pinar, M.J.(Eds). Ediciones de la Universidad de Castilla-La Mancha. Cuenca. 963-972. (Iruskieta *et al.* 2009)

– Iruskieta, M.; Díaz de Ilarraza, A.; Lersundi, M. 2008. "Análisis de los marcadores del discurso para el euskera: Denominación, clases, relaciones semánticas y tipos de ambigüedad". In Bretones, M. C. et al. (Eds). Applied Linguistics Now: Understanding Language and Mind. Almería: Universidad de Almería. 1271-1282. (Iruskieta *et al.* 2008)

Despite not being the direct result of this doctoral thesis, the following publications are nevertheless works carried out in the field of language processing, and the majority of them are vital to understanding many of the tools used to create the Basque RST TreeBank.

- Morphological level:

  - Aldezabal I., Ceberio K., Esparza I., Estarrona A., Etxeberria J., Izagirre E., Quintian Iruskieta M., Uria L. "EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio mailan etiketatzeko eskuliburua". UPV/EHU / LSI / TR 11-2007. (Aldezabal *et al.* 2007b)

- Syntactical level:

  - Aldezabal I., Aranzabe M.J., Arriola J., Díaz de Ilarraza A., Estarrona A., Fernandez K., Iturria L., Quintian Iruskieta M. 2007. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua. UPV/EHU / LSI / TR 12-2007. (Aldezabal *et al.* 2007a)
  - Uria L., Estarrona A., Aldezabal I., Aranzabe M.J., Díaz de ILarraza A., Iruskieta M. 2009. Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque. Lecture Notes in Computer Science (LNCS) 5449: 72-85. Springer. (Uria *et al.* 2009)

- Semantical level:

  - Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Pociello E., Quintian Iruskieta M. 2006. "Improving the Basque WordNet by corpus annotation". In Proceedings of Third International WordNet Conference. pp. 287-290. Jeju Island, Korea. (Agirre *et al.* 2006b)
  - Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Pociello E., Quintian Iruskieta M. 2006. "A methodology for the joint development of the Basque WordNet and Semcor". In Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC). Genoa, Italy. (Agirre *et al.* 2006a)
  - Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Pociello E., Quintian Iruskieta M. 2005. "EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa". Euskalingua-7: 212-219. (Agirre *et al.* 2005a)
  - Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Pociello E., Quintian Iruskieta M. 2005. EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editzatze-, etiketatze- eta epaitze-lanak. UPV/EHU/LSI/TR 23-2005. (Agirre *et al.* 2005b)

- Discourse level:

  - Garcia J., Iruskieta M. 2013. Birformulatzaile zuzentzaileak testu idatzietan. Gomez, Ricardo & Ezeizabarrena, Maria Jose (arg.).

Eridenen du zerzaz kontenta. Sailkideen omenaldia Henrike Knörr irakasleari (1947-2008). Bilbo: EHU. (Garcia eta Iruskieta 2013)

1.4 Outline of the translation

The abbreviated translation of the thesis report differs from the original in a number of different ways. The original thesis is more detailed than the abbreviated translation and contains details of the research work carried out to link the findings of and conclusions drawn in the various papers in a coherent way. Thus, the structure of the abbreviated translation of the thesis report is as follows:

- Chapter 1 – Introduction.

  In the first chapter we provide a general introduction to the research theme, explaining what relational discourse structure is and outlining both our motivation and our aims. We also offer an overview of what is presented in each chapter. And finally, we list those publications linked to our research and specify the different sub-sections to which they correspond.

- Chapter 2 — Preparation phase:

  This chapter presents the methodology used for annotating and evaluating relational discourse structure. Firstly, we argue our reasons for selecting the theory (RST) used to annotate the corpus with discourse relations, and outline its limits and advantages. Secondly, we describe both the corpus itself (consisting of 60 scientific abstracts) and the annotators (4 linguists with no prior training). Thirdly, we outline the main criteria for resolving problems of circularity between the annotation phases (segmentation, macro-structure, rhetorical structure and rhetorical relation signals). Circularity problems mainly occur between segmentation and rhetorical relations, and between rhetorical relations and their signals. Fourthly, we describe the main characteristics of the delivery phase, before finishing with a summary of the chapter contents.

- Chapter 3 – Conclusions and future work:

  This chapter presents the conclusions of the thesis and lists and highlights its contributions to the field. It also identifies some of the future areas of research opened up by this work.

  Below are the references related to the translation of the abbreviated thesis report. Nevertheless, the subsequent sections are made up by scientific papers, and the references corresponding to each are given at the end of each paper.

- Paper 4 Iruskieta *et al.* (2011a) – The paper which describes the first prototype for automatic discourse segmentation in the Basque language.

  The paper describes how we reused and adapted the tool Iruskieta *et al.* (2011a) used for the Basque language in order to design the automatic segmentation instrument described in (Arrieta 2010). The clause segmentation tool proposed by (Arrieta 2010) identifies phrases through grammar rules based on linguistic information (Constraint Grammar) and automatic learning techniques (Carreras 2005). The rules are used to identify phrase endings, and automatic learning techniques based on the linguistic information of each word are used to identify their beginnings also.[17]

- Paper 5 (Iruskieta *et al.* Forthcomingb:) – A paper on the correlations between the central unit and rhetorical relations. If the macro-structure (central unit) determines the micro-structure (rhetorical relations), then it is logical to assume that if we harmonize the central unit, the inter-annotator agreement for certain relations should increase. If agreement regarding the principal idea increases, then this may have a positive effect on agreement regarding rhetorical relations. In order to determine whether or not this effect is significant, this paper aims to identify any possible correlation between the central unit and rhetorical relations. If a correlation were indeed to exist, then it would produce changes in the rhetorical structure annotation phase,[18] as well as in the evaluation of rhetorical relations.[19]

- Paper 6 (Iruskieta *et al.* 2011b:) – A paper on the correlations between segmentation levels and rhetorical relations. This paper demonstrates

[17] In order to improve the aforementioned results, in the Basque version of the thesis report we explain how we used another two syntax-based tools developed by the IXA group for automatic discourse segmentation:

- We annotated the texts with morphosyntactic information using the IXAti tool (Aduriz *et al.* 2004) and added end markers to the morphosyntactic information using Constrained Grammar-based rules.
- We annotated the texts with syntactic dependency using the MALTIXA tool (Diaz de Ilarraza *et al.* 2005), and added end markers using some dependency-based rules.

Even though the results obtained were better, we believe there is still room for improvement; therefore, these results are provisional. The following are the results for end marker identification (F_1): 66.94% based on Arrieta (2010) 69.69% based on Aduriz *et al.* (2004), and 80.68% based on Diaz de Ilarraza *et al.* (2005).

[18] Generally, in RST, there are two annotation phases: *i*) segmentation and *ii*) the building of the rhetorical structure. If there were a correlation between the central unit and rhetorical relations, then there would be three annotation phases: *i*) segmentation, *ii*) identification of the central unit, and *iii*) the building of the rhetorical structure.

[19] Generally, when measuring inter-annotator agreement for rhetorical relations, all relations carry the same weight regardless of whether they are low on the tree (easiest) or higher up the tree structure (most difficult).

that agreement is higher and stronger at the intra-sentential level; in other words, according to the corpus data, agreement is based on the composition point and the attachment point (RCA), rather than on other partial agreements (RA, RC and R). The results demonstrate that an incremental annotation method is a suitable strategy, among other reasons because inter-annotator agreement tends to be higher at lower levels.

- Paper 7 (Iruskieta *et al.* 2013b:) – An in-depth analysis of the quantitative method for evaluating rhetorical relations. In this paper Marcu (2000a) we analyze, among other things, the drawbacks of the tree structure evaluation method.[20]

- Paper 8 (da Cunha eta Iruskieta 2010:) – The first proposal for comparing rhetorical relations from a qualitative and quantitative perspective. In this paper, we establish the basics of the innovative qualitative-quantitative methodology for comparing rhetorical relations, and compare tree structure for both Basque and Spanish.

- Paper 9 (Iruskieta *et al.* Forthcominga:) – The latest proposal for comparing rhetorical relations from a qualitative and quantitative perspective. In this paper, we systematize the innovative qualitative-quantitative methodology for comparing rhetorical relations, and compare tree structures for Basque, English and Spanish. We also describe the impact of strategic translations and different annotator interpretations on the rhetorical structure.

- Paper 10 (Iruskieta *et al.* 2013a:) – The paper corresponding to the website from which the Basque RST TreeBank can be consulted. In this paper we describe the different consultations that can be made using the web service that was established specially to enable consultations regarding the factors analyzed in the thesis report.[21]

[20] Some of these drawbacks are mentioned in the following works: van der Vliet (2010a), da Cunha eta Iruskieta (2010), and Iruskieta *et al.* (2013b).

[21] Due to space constraints in the thesis report, in this paper we describe in more detail some interesting phenomena that can be analyzed in more depth using the website. Examples include: how rhetorical relations linked to the central unit reveal the macro-structure of scientific abstracts (IMRaD structure) and the ambiguity of signals for rhetorical relations, among others.

2

Methodology used for annotating and evaluating relational discourse structure

In this chapter, we describe the methodology we propose for annotating Basque language texts with relational discourse structure, based on the steps proposed by Hovy (2010). Since we divided the annotation process into different phases, we will explore some of the methodological aspects linked to theoretical concepts in more detail in the corresponding chapter. We opted for this structure in order to enable each phase to be explained in its entirety.

The majority of works analyzing rhetorical structure phenomena in the Basque language are based on a formal approach. Thus, when describing coherence, these works seek to explain explicit coherence relations by describing their form-based components (mainly discourse markers).[1]

The most significant monographic works focusing on the Basque language from a formal perspective have analyzed grammar, since they are limited to semantic relations. It is within this approach that we can place the works carried out by Euskaltzaindia – the Basque Language Academy (1990, 1994, 1999, 2005) on connectives, coordinating conjunctions and subordinating conjunctions.

Although some other works do indeed adopt a discourse perspective, they do not offer a comprehensive description of coherence relation categorization, nor do they analyze implicit relations. For example, Esnal (2008) analyses discourse markers within the context of writing strategies for educational texts. Ibarra (2013) and García (2010) study discourse markers in spoken texts, with Ibarra (2013) focusing on those used in the spoken language of young Basque speakers and García (2010) on those appearing in reformulations of students' spoken texts. Aierbe (2008) also analyses reformulation, although in this case in administrative texts, and in addition to reformula-

[1] We use the term discourse marker here in its broadest sense, without taking into account the diverse designations or limitations proposed in the literature.

tion, Urrutia (2008) analyses other discourse markers in administrative texts. From a contrasting perspective, Barandiaran eta Casadok (2011) study reformulation, analyzing and comparing reformulations in Basque and Spanish. Also, Zabalak (1996), from an educational perspective, analyses the discourse markers used in the appositions not studied in the work carried out by Euskaltzaindia 1990, and Larringan 1995 analyses discourse markers in different types of texts (informative and argumentative). The EUDIMA project (Alberdi eta Garcia 2012) aims to create a kind of dictionary of discourse markers, adapting the work carried out in Spanish (*Diccionario de partículas discursivas del español*) to the Basque language. In this project, the authors analyze the reformulation discourse markers established in previously published monographic works (Alberdi eta Landa 2013; Alberdi eta Garcia 2012; Alberdi 2011a; Alberdi 2011b; Azkarate 2013; Garcia eta Iruskieta 2013, among others).

In order to perform a number of tasks linked to coherence within the field of computational linguistics, it is necessary to move beyond the formal perspective. We cannot limit ourselves to analyzing only explicit relations; we must describe coherence relations or the coherence relational discourse structure of the whole text. Since most of the relations in a text are implicit, any analysis of coherence relations must necessarily be carried out from a semantic-pragmatic perspective, within the field of computational linguistics. However, formal considerations (analysis of discourse markers) should also be taken into account, even if it is with a clearly utilitarian purpose (i.e. to define relation patterns). Since we are dealing mainly with implicit relations, research into rhetorical structure within the field of computational linguistics is no simple task, since the complexity of the issue being studied cannot be described using general terms and a small number of rules.

One solution is to manually annotate large-scale corpora so that, subsequently it becomes possible for a machine to use the annotated corpus to learn patterns based on rhetorical structure and to automatically describe the rhetorical structure of non-annotated texts.

Providing the size and quality of the annotated corpus are adequate, we can analyze rhetorical structure using machines, and this in turn enables us to develop advanced language-based applications, such as automatic discourse segmentation, automatic summarization and the machine translation of certain phenomena that signal rhetorical structure, among others.

As the complexity of the topic being studied increases, so must the size of the corpus, although of course, quality is also of vital importance in the complexity/size ratio. According to Hovy (2010), topics studied at the discourse level are more complex than those studied at other linguistic levels. If this is true, then the evaluation of annotated corpora becomes of primary importance as a means of determining the quality of annotated texts.

According to Hovy (2010), the annotation process must be reliable in order to enable the object to be automatically analyzed in an appropriate

manner. Moreover, the information added to the corpus must be in-depth in order for the conclusions drawn from it (both theoretical and practical) to be of interest.

This section is structured as follows: First of all, section 2.1 outlines the main characteristics of the preparation phase for corpus annotation. Next, section 2.2 describes, phase by phase, how the corpus was annotated and evaluated. Section 2.3 describes the delivery phase for the results obtained and finally, section 2.4 offers a summary of the chapter contents.

2.1 Preparation phase: corpus and annotators

This section focuses on the criteria used to build the corpus. It also describes the annotators who carried out the task.

2.1.1 Description of the corpus

The following criteria were taken into account when establishing the corpus:

Corpusa eratzeko, hautatu ditugun irizpideak honako hauek dira:

i) In order to compensate the domain effect in the analysis of rhetorical relations, tests were selected from a number of different domains, with the same number of texts being chosen from each. The corpus used in this thesis was drawn from three different domains, as described in Table 2.1.

Domain	Sub-corpus	Texts	Sentences	Words	Annotators
Medicine	GMB	20	198	3010	E_1, E_2
Terminology	TERM	20	253	5664	E_1, E_2, E_4
Science	ZTF	20	352	6892	E_1, E_2, E_3
Total		**60**	**803**	**15566**	

2.1 taula – Description of the Basque language corpus being studied

ii) Texts were required to be well structured and brief. Texts were required to be well-structured for two reasons. Firstly, in order to identify the rhetorical structure of different text types,[2] and secondly, in order to ensure as high a level of inter-annotator agreement as possible.[3] They were also required to be brief, in order to enable relational discourse structures to be manually compared and precisely evaluated. The texts which best meet these criteria are abstracts of scientific papers. As regards the communication aims of the abstracts, they can be classified

[2]To determine the influence between macro-structure and rhetorical structure.

[3]In order to ensure that the impact of inter-annotator agreement regarding macro-structure had a positive impact on inter-annotator agreement regarding rhetorical relations (i.e. rendering it more reliable).

as specialist texts (Cabré 1998), since their objective is to present and convey specialist knowledge and both the authors and the target readers are experts in their field. The abstracts consist of the title of the paper and a brief summary of its contents.

iii) Texts were required to be written in more than one language; this was to allow contrasting analyses to be carried out and to enable the corpus to be used in machine translation tasks.
The majority of corpora are in three languages. We outline here the methodology used to annotate the Basque language corpus for the purposes of this thesis, even though in other works, a significant part of the corpus created for this thesis is analyzed in English and Spanish (Iruskieta *et al.* Forthcominga; da Cunha eta Iruskieta 2010; Iruskieta eta da Cunha 2010a; Iruskieta eta da Cunha 2010b; da Cunha eta Iruskieta 2009).

2.1.2 Description of the annotators and the super-annotator

All annotators involved in this thesis were linguists. The majority had experience annotating texts at other language levels (morphologic, syntactic and semantic). None had any prior experience annotating with phenomena at the discourse level. They therefore relied on RST when annotating rhetorical structure. After presenting RST, we outlined a series of annotation criteria and introduced the annotators to the RSTTool. Although on certain occasions we were obliged to clarify specific doubts regarding the conceptualization of certain structures, there was no training phase as such.

We decided not to establish a training phase for annotators because one of the criticisms levelled at RST is that it is subjective, and we wanted to identify and analyze inter-annotator disagreements, since our aim is to establish specific criteria for annotating rhetorical structure.

The annotator who had most experience annotating and evaluating using RST was selected as the judge or super-annotator (Hovy 2010). The super-annotator annotated each phase before looking at the annotations made by the other annotators, and once all the annotations had been collected, checked for inter-annotator agreement/disagreement. In the case of inter-annotator agreement, he/she established the corresponding criteria, which were then used to harmonize results in the event of disagreements or when the criteria were not followed.

2.2 Annotation, evaluation and harmonization phases

According to RST, when an annotator wants to represent the rhetorical structure of a text, first of all he or she has to segment the text, and then specify

the relations which exist between the different discourse units. In order to avoid circularity, we designed a non-retroactive phase-based annotation process as follows:

1) Discourse segmentation was carried out on the basis of syntactical function (segmenting adjoining clauses) and form (dividing segments containing verbs).
2) We identified the macro-structure or central unit.
3) The annotation of rhetorical relations was carried out in accordance with meaning, with no prior training period and with no signal criteria being given.
4) The annotation of signals was based on form-related criteria.

2.2.1 Segmentation

An obligatory first step in the annotation of any reference corpus (at any segmentation level) is to identify the discourse units. This is known as the segmentation phase. The aim of segmentation is to mark the elementary units of the text, or in other words, to establish the basic elements of each language analysis level in order to enable the subsequent identification of the relation that exist between them. Different definitions of what an elementary discourse unit (EDU) actually is have been proposed within RST. Although it is never explicitly stated, segmentation proposals are based on the following three basic concepts:

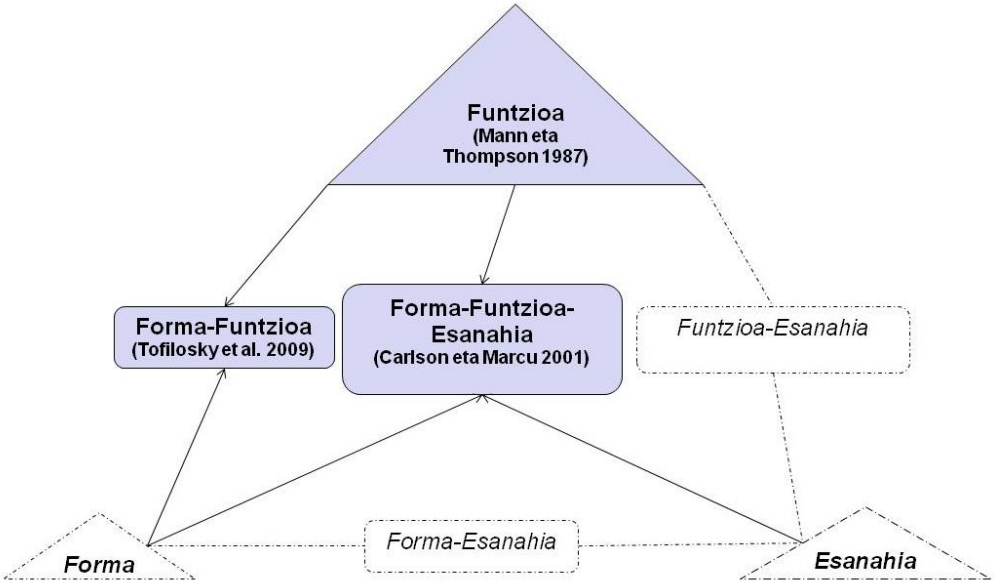
i) Linguistic "form" (or category).
ii) "Function" (the function of the syntactical components).
iii) "Meaning" (the coherence relation between propositions).

The possible combinations which exist between these basic concepts used in discourse segmentation and those proposed by RST are highlighted in gray in Figure 2.1. The basic concepts or combinations that have not been proposed are marked with a dotted line. Concepts are shown in triangles and combinations in rectangles.

The best-known segmentation proposals within RST are outlined in a paper by van der Vliet van der Vliet (2010b). The basic concepts used in the most important of the segmentation proposals listed in that work are:

- The original RST proposal in English (Mann eta Thompson 1987): all clauses are EDUs, except for restrictive relative clauses and clausal subject or object components (syntactical function). This proposal is based solely on syntactical function.
- The first RST-based annotated corpus in English (Carlson eta Marcu 2001): in addition to that outlined in the original proposal, here both the components of attribution clauses (criterion based on function and meaning) and those phrases that begin with a discourse marker (e.g. because of, in spite of, according to, etc.) are also segmented (criterion based on form and semantics). This proposal uses all three basic

2.1 irudia – The basic concepts of discourse segmentation: form, function and meaning



concepts: form, function and meaning.

– A segmentation proposal in English that adheres more closely to the original RST proposal (Tofiloski *et al.* 2009): it segments verb clauses, coordinated clauses, adjunct clauses and non-restrictive relative clauses marked by a comma (it is a proposal based on form restriction and syntactical function). Unlike in the proposal tabled by Carlson eta Marcu (2001), in this method phrases beginning with discourse markers are not segmented, since they contain no verbs. In the Spanish corpus, da Cunha *et al.* (2010) follow this segmentation method.

When attempting to define what a "discourse unit" actually is, these three basic concepts (form, function and meaning) pose a number of problems. These problems are as follows: *a*) if we based our analysis on form alone, many of the segmented elements would not be discourse units. For example, if we asked annotators to segment verb clauses with the "-*tzeko*" form, then they would also segment clauses that are not discourse units. *b*) if we based our analysis on function alone, then we would only be able to give annotators overly generalized definitions and imprecise segmentation criteria, such as adjunct clauses with verbs, etc. *c*) And finally, if we based our analysis solely on meaning, we would encounter the problem of circularity between the segmentation annotation phase and the rhetorical relation annotation phase. The clearest example of this is that in order to annotate ATTRIBUTION relations, we would first have to segment the attribution clauses in the segmentation phase, resulting in a mixing of the two phases.

The sub-phases carried out within this annotation phase are as follows:

a) Annotation. Without moving too far away from the original RST proposal, and based on the syntactic criteria proposed by Tofiloski *et al.* (2009), the annotators were asked to divide the texts into elementary discourse units. This was carried out using the RSTTool (O'Donnell 2000).
b) Evaluation. To measure inter-annotator agreement for segmentation, we used the Kappa score.
c) Harmonization. Following the evaluation, the super-annotator compared and harmonized the segments identified by the other annotators. Next, the F-score was used to measure the agreement level obtained by each annotator in relation to the harmonized text. To calculate the Kappa and F-scores, the super-annotator used the XIRABA application (Zapirain 2004).
d) The work carried out in this annotation phase was later used to fulfill another of the thesis's objectives, namely to establish an annotated reference corpus with fine grained segmentation. This segmented reference corpus will serve in the future for developing automatic segmentation tools and measuring their reliability.

2.2.2 Identifying the macro-structure

In order to determine coherence, in addition to local level phenomena (related to the meaning linking words and sentences), global or macro-structure level phenomena (related to the connections between the text's main theme and other themes) also need to be identified (van Dijk 1980a). In other words, if a discourse is to be truly coherent, it must be so at all levels: local, global and as regards the linkage between the two.

In RST, the rhetorical structure at a local level is hierarchical; in other words, some discourse units (nuclear units) are more important than others (satellite units). Evidence of this hierarchy is provided by the fact that, if we take away the nuclear units, the text becomes incoherent (Mann eta Thompson 1987). As result of this hierarchical structure of discourse, texts can be automatically summarized, as shown in a number of different studies: Ono *et al.* (1994), Rino eta Scott (1996) and da Cunha (2008).[4]

To our mind, the failure to take the global level into account has an impact on a number of inter-annotator agreement factors:

However, nuclearity at a global level is not contemplated in the guidelines for annotating with RST. These guidelines only take local nuclearity into consideration (Carlson *et al.* 2001).

Bearing in mind the importance of annotation in this thesis, we wanted to analyze the consequences of beginning the annotation process from the same or different macro-structures.

[4]For a thorough and critical analysis of nuclearity in RST, see the work by Stede (2008b).

- Agreement regarding the global perspective of the text. In other words, if two annotators base their annotation on different views of the text's global perspective or macro-structure (i.e. if they choose different EDUs as the main theme), then they will build different rhetorical structures.
- Agreement regarding the rhetorical relations linked to the most important unit of the tree structure. Since rhetorical relations have to be identified in order to determine the impact of global nuclearity on local rhetorical structures, this factor (which influences inter-annotator agreement) is discussed in the appropriate sub-section.

If we demonstrate that the macro-structure influences rhetorical relations, then after the segmentation phase but prior to the building of the rhetorical structure, we must decide which unit is the central unit and must design a new phase for identifying the global level within the RST annotation method (something which is not proposed in the literature). This may increase inter-annotator agreement regarding relations. We analyzed the indicators[5] for identifying the central unit or macro-structure (Iruskieta *et al.* Forthcomingb). Since this annotation phase is one not contemplated in other works within the RST approach, we will provide here a detailed description of it and highlight its advantages.[6]

The sub-phases carried out within this annotation phase are as follows:

a) Annotation. The annotators annotated the text's central unit (Iruskieta *et al.* Forthcomingb).
b) Evaluation. The super-annotator measured the inter-annotator agreement level attained in relation to the central unit.
c) Harmonization. Following a set of structural criteria, the super-annotator then resolved any cases of disagreement in order to establish a corpus with harmonized central unit annotation.
d) Indicator annotation. We analyzed the indicators for the central unit and studied their strength.

2.2.3 Relational discourse structure

As regards rhetorical relations and their annotation, in this thesis we analyze the five main questions that have prompted most discussion in the literature and which impact the theoretical-methodological framework: *i*) the nature of rhetorical relations (Taboada eta Das 2013), *ii*) the classification of rhetorical relations (Mann eta Taboada 2010), *iii*) circularity between rhetorical rela-

[5]In accordance with Paice (1980), in this thesis we use the term indicator to refer to any word or structure from any category that can be used to indicate the central theme.

[6]Selecting the central unit has advantages at two different levels: *i*) selecting the most important idea in the tree structure results in greater inter-annotator agreement and *ii*) it also results in greater inter-annotator agreement regarding the relations linked to the key idea at the first level.

tions and their signals (Spenader eta Lobanova 2009), *iv*) the signaling power of signals (ambiguity) (Mann eta Thompson 1987; van Dijk 1998; Taboada 2006) and *v*) inter-annotator subjectivity and its assessment (Marcu 2000a; da Cunha eta Iruskieta 2010; Mitocariu *et al.* 2013).

This annotation phase was divided into the following sub-phases:

1) Annotation. In accordance with the proposals of van Dijk (1980b), Thompson *et al.*ek (1985) and Pardok (2005), we established a specific annotation method taking macro-structure into account, from left to right, in an incremental and modular fashion: *i*) Since macro-structure has an impact on the low level discourse relations of the tree structure, we take the macro-structure or central unit into account when establishing the tree structure representation (van Dijk 1980b).
   ii) The discourse units are linked from left to right within the same sentence (Thompson *et al.* 1985).
   iii) The discourse units are annotated incrementally (from bottom up, i.e. by first joining EDUs and then establishing relations between all tree units) (Pardo 2005).
   iv) Annotation is modular (first units are related within the same sentence, then sentences are related within the same paragraph, and then finally, relations are established between the paragraphs themselves) (Pardo 2005).
   For text annotation we used the extended classification of rhetorical relations provided by RST, and the RSTTool graphic environment. As regards double discourse unit relations, although Mann eta Thompson (1987) defend the view that a text can have more than one correct interpretation, we decided to use a single rhetorical structure for each text.
2) Evaluation. We measured a number of different phenomena using a quantitative-qualitative evaluation method (Iruskieta *et al.* Forthcominga): *i*) whether inter-annotator agreement was greater at a low or high level of the tree, in the GMB sub-corpus (Iruskieta *et al.* 2011b); *ii*) whether there was any correlation between the central unit and rhetorical relations (Iruskieta *et al.* Forthcomingb); *iii*) whether there was greater inter-annotator agreement regarding rhetorical relations linked to the central unit than regarding those not linked to the central unit (Iruskieta *et al.* Forthcomingb); and *iv*) we measured mean inter-annotator agreement in pairs and groups of three, using two statistical measurements: F-score and Fleiss's Kappa score 1971. Furthermore, we also identified the principal confusion matrixes for the relations.
3) Harmonization. Since annotating a corpus with rhetorical structure is a complex process in which annotators may come up with different interpretations, any method which aims to increase the reliability of the task cannot rely on signal-based training or annotation criteria, as this would give rise to circularity. Therefore, in order to increase the reliabi-

lity of the annotation process, we decided to appoint a super-annotator to harmonize rhetorical structures and resolve any inter-annotator disagreements. To this end, the super-annotator laid down a general set of guidelines and then used these guidelines to resolve disagreements or cases which did not follow the established criteria. The harmonization process described above is a proposal made in this thesis, and as such we would like to highlight the fact that, just as with segmentation criteria, the suitability of the harmonization criteria must be evaluated in order to determine the reliability of the proposed method.

In order to determine whether the method chosen was adequate, we based our analysis on two principles: *i*) inter-annotator subjectivity must have as little influence as possible, and *ii*) we must be able to describe inter-annotator disagreements as precisely as possible. In accordance with these two principles, we used two different evaluation methods: the quantitative evaluation method described by Marcuren (2000a) and the quantitative-qualitative evaluation method described by da Cunha eta Iruskietaren (2010) (from hereon, the qualitative evaluation method). This second method is an improvement on the quantitative evaluation method, since it measures nuclearity, linkage (relation), attachment point and composition factors independently. Another advantage of qualitative evaluation is that it can also be used to compare different languages; in other words, it helps identify the disagreement problems between different languages (da Cunha eta Iruskieta 2010; Iruskieta *et al.* Forthcominga).

1) In relation to the question of composition, our aim was to determine the rhetorical structure level at which inter-annotator subjectivity is lowest. To do so, we evaluated intra-sentential rhetorical structure and inter-sentential rhetorical structure (Iruskieta *et al.* 2011b). If there is greater agreement at the lower levels of the tree structure (segments with a simple composition)[7] than at the higher levels, then this would justify a bottom up (incremental and modular) annotation method,[8] and would enable us to measure correlation between syntax and discourse (Soricut eta Marcu 2003).
2) As regards the question of attachment points, our aim is to analyze whether selecting the same attachment for the text's most important idea (macro-structure or central unit) or a different one affected agreement regarding rhetorical relations. Moreover, in order to analyze whether or not the text's most important idea or its macro-structure influenced rhetorical relations, we will compared the agreement found regarding rhetorical relations linked to the central unit with agree-

[7]The attachment point unit factor may also have an impact. At low tree structure levels, the attachment point tends to be located within a sentence; therefore, since it is simpler, agreement regarding it should be greater.

[8]A bottom up annotation method is vital to avoiding the problem of circularity between segmentation and rhetorical structure.

ment found regarding other rhetorical relations (Iruskieta *et al.* Forthcomingb). If the central unit influences agreement regarding rhetorical relations, then this would indicate that the central unit should be annotated before rhetorical relations.

3) As for relation, our aim was to determine which rhetorical relations have the lowest degree of subjectivity and why they tend to be ambiguous. This will enable us to determine whether or not these rhetorical relations can be detected automatically, or at what level they can be detected, and will serve as the basis for the design of the first automatic discourse analyzer prototype (Iruskieta *et al.* 2011a).

Thus, more than finding a means of increasing inter-annotator agreement, the key aim of this thesis was to describe the problems that may arise during rhetorical structure annotation, and to propose possible solutions.

2.2.3.1 Signals for rhetorical relations

After annotating the texts with rhetorical relations, the signals for these relations must be annotated.

a) Annotation. Following Taboada and Das's proposal 2013, a single annotator annotated the signals for all rhetorical relations using the Rhetorical DataBase tool (Pardo 2005).

b) Evaluation. In order to evaluate the work carried out by the single annotator, another two annotators annotated the three relations in the cause subgroup (CAUSE, RESULT and PURPOSE). We measured the mean agreement between all annotators in order to specify the reliability level of the signals for these three relations.

c) Harmonization. The super-annotator resolved any disagreements arising within the cause subgroup.

2.3 Delivery phase

Since no previous Basque language corpus annotated with rhetorical structure existed, we decided to publish the work carried out by the super-annotator, in the hope that this would enable any gaps to be filled in. The main aim of the delivery phase was to describe the possible uses of the corpus. When doing so, we took into account the key criteria used to describe the corpus annotated by Ide eta Pustejovskyk (2010): *i*) description of the theoretical framework, *ii*) annotation guidelines, *iii*) project documents, *iv*) characteristics of the annotated corpus and *v*) the uses to which it can be put.

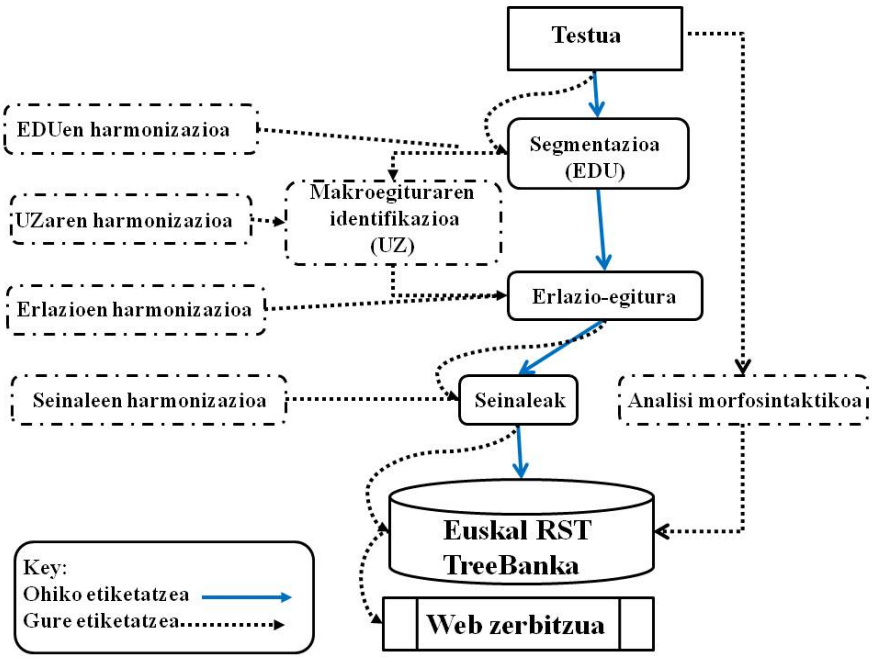
In this phase, we will present the Basque RST TreeBank tool, following the criteria established by Ide and Pustejovsky 2010 for disseminating their annotated corpus. This corpus is the first corpus in the Basque language that has been annotated with rhetorical structure at the discourse level. Although the main innovation offered by the Basque RST TreeBank is its language

(the Basque language), it also offers a series of other innovations that are not linked to (the Basque) language. For example, a number of operations can be carried out with this annotated corpus that cannot be carried out with other language corpora and which may be interesting and useful from both a theoretical and practical perspective. These operations are: *i*) all occurrences of each relation in the corpus can be viewed; *ii*) the relations or elementary discourse units of a text can be consulted; *iii*) the central unit or most important discourse unit of each text can be seen, along with the relation that links it to the central unit; *iv*) the rhetorical relation signals can be viewed, along with their degree of ambiguity; and *v*) searches can be conducted based on morphosyntactic information.

2.4 Summary

In this chapter we describe the methodology used to annotate the corpus with rhetorical structure. The proposed methodology is shown in Figure 2.2 (both the annotation method used in RST and the annotation method followed here).

2.2 irudia – RST annotation method and the Basque RST TreeBank annotation method



Moreover, we outline the characteristics of both the corpus itself and the annotators, we describe the specific phases of corpus annotation, we propose a new annotation evaluation method and we define the path to be followed in order to obtain a reference corpus annotated with rhetorical structure.

We would like to highlight that, as shown in Figure 2.2, this thesis proposes an innovative method within the field of RST. The method proposed here differs from the standard one in the following ways:

a) Annotation phase: we propose that the central unit be annotated prior to the rhetorical relations.

b) Evaluation phase: we propose a qualitative-quantitative method for evaluating rhetorical relations.

c) Harmonization phase: we propose a means of harmonizing each phase.

As regards number of texts, the corpus created for the purposes of this thesis is similar in size to other corpora in the literature (Taboada eta Renkema 2011; Pardo eta Seno 2005; van der Vliet *et al.* 2011). The information contained in it is comprehensive and thorough, and the method proposed for ensuring its reliability is also innovative, since: *i*) a new evaluation system for measuring inter-annotator agreement regarding relations is proposed, and *ii*) the set of criteria followed by the super-annotator when resolving disagreements are established first of all, in accordance with RST.

3 Conclusions and future work

In this last chapter, we will first summarize the main contributions made by this thesis to the field, then we will outline the main conclusions, before identifying some of the future areas of research opened up by this work.

3.1 Contributions

With the aim of complementing the language levels at which the processing of the Basque language has been analyzed in the past, we established a methodology for manually annotating texts with rhetorical structure at the discourse level, and then annotated a corpus in the Basque language. We then verified the innovative nature and reliability of this method with the findings of this thesis. Moreover, we believe that this methodology may prove useful to others working in different languages within the field of RST.

3.1.1 Contributions linked to the Basque language

Within the field of the rhetorical structure of discourse, the main contribution made by this thesis is the Basque RST TreeBank. First of all, Basque language texts were annotated with rhetorical structure using RST. The corpus is made up of 60 texts from three different domains within the same genre (Medicine, Terminology and Science and Technology, all within the scientific paper abstract genre). In total, it contains 15,566 words, 1,355 elementary discourse units (EDUs), 1,315 rhetorical relations and 783 signals. In order to enable this harmonized corpus to be used for language processing tasks, the website of the IXA group is available to all members of the scientific community at the following address: `http://ixa2.si.ehu.es/diskurtsoa/en/`. The website contains the following resources:

- The texts which make up the corpus, in *txt* format.
- The files that have been automatically annotated with morphosyntactic information, in *kaf* format.

- The corpus segmented at the intra-sentential level and the texts annotated with rhetorical structure, in the original *rs3* or *xml* format.
- The files annotated with signals for rhetorical relations, in *RhetDB* format.

A more in-depth description of what consultations can be carried out on the website and what information is available is given in Chapter 7. We will now outline the contributions of our work at each annotation level.

- **Segmentation**: We adapted adjunct verb clause-based segmentation Tofiloski *et al.* (2009) to the Basque language. In order to avoid the problem of circularity between segmentation and rhetorical relations (Taboada eta Mann 2006), we established a set of criteria that are unrelated to either RST or rhetorical relations. The mean inter-annotator agreement level (F_1) for Basque language intra-sentential discourse segmentation, measured using the XIRABA application, was 81.14%.

  We also developed a set of criteria to enable the super-annotator to resolve cases of inter-annotator disagreement, and in accordance with these criteria, we obtained a reference segmented text made up of 60 separate texts.[1]

  We developed and manually evaluated a prototype automatic discourse level segmentation program, taking advantage of the automatic clause identifier developed by the IXA group (Arrieta 2010). The reliability of this prototype is 57.81% (F_1) for EDUs, and 66.94% for end boundaries. The result obtained for end boundaries using Constraint Grammar-based rules was 69.69% (F_1), and that obtained using syntactic dependency based heuristics was 80.68% (F_1).

- **Nuclearity**:

  When annotating the central unit identified as the macro-structure, the mean inter-annotator agreement rate was 61.42%. The super-annotator then harmonized any disagreements based on a set of structural criteria.

  In order to identify the central unit, we described the indicators for noun and verb category central units and used the Basque RST TreeBank to propose a method for calculating the frequency with which these indicators appear in the central unit.

- **Rhetorical relations**:

  Our results were similar to those obtained by other RST annotation projects with similar characteristics (Carlson *et al.* 2001; van der Vliet *et al.* 2011). The mean inter-annotator agreement for rhetorical relations, measured in pairs, was 61.81% (F_1) (3,189 to 1,971).

[1] `http://ixa2.si.ehu.es/diskurtsoa/segmentuak.php`.

We annotated the signals for all rhetorical relations. We also established a process for harmonizing the rhetorical relations of the cause subgroup (CAUSE, RESULT and PURPOSE). The mean inter-annotator agreement obtained for these relations was 60.52% (measured in threes) and 76.82% (measured in pairs).

3.1.2 General contributions

Even though the annotation of a language of a different typology is, in general, always interesting for those analyzing relational discourse structure, we believe that the innovations presented within the RST annotation project may also arouse interest in other languages. The contributions that we believe may be of general interest are as follows:

- **Regarding the corpus**:

  In addition to being used in this thesis project, the corpus created for the purposes of this study has also been used in other RST research initiatives and projects in order to fill in the gaps identified within the theoretical framework of rhetorical structure theory.

  DiSeg,[2] the first automatic segmentation tool in Spanish, was assessed in accordance with the gold standard corpus made up by the segmented sub-corpora GMB (abstracts of scientific papers in the medical field) and TERM (abstracts of scientific papers in the field of terminology) (da Cunha *et al.* 2010).

  In the *RST Spanish TreeBank*[3] Spanish RST corpus and website, the texts of the GMB (abstracts of scientific papers in the medical field) and TERM (abstracts of scientific papers in the field of terminology) sub-corpora have been annotated with Spanish rhetorical relations (da Cunha *et al.* 2011).

  The web applications developed in the Basque RST TreeBank have been used to carry out a number of consultations within the Multilingual RST TreeBank[4] (consisting of the Basque, English and Spanish versions of 15 texts from the TERM corpus: sub-corpus of scientific abstracts in the terminology field) (Iruskieta *et al.* 2013b). The consultations carried out include (among others): *i*) search for the relations and key idea (central unit) of a specific tree structure; and *ii*) search for a specific rhetorical relation in the corpus, in three different languages.

- **Regarding the annotation phase**:

[2] `http://daniel.iut.univ-metz.fr/DiSeg/WebDiSeg/`.
[3] `http://corpus.iingen.unam.mx/rst/`.
[4] `http://ixa2.si.ehu.es/rst/`.

We propose a new phase within the RST annotation method to annotate the text's central unit. This new phase comes after discourse segmentation but before the annotation of rhetorical relations (Iruskieta *et al.* Forthcomingb). Annotators who base their work on the central unit take the macro-structure into account when building the tree structure. The justification for this phase is based on the following three findings: *i*) if the central unit is identified prior to the building of the rhetorical structure, inter-annotator agreement is higher; *ii*) inter-annotator agreement regarding relations is higher in tree structures that have the same central unit than in those with different ones; *iii*) mean inter-annotator agreement is higher in relations linked to the central unit than in those not linked to the central unit. According to these three findings, adding this phase would result in an increase in inter-annotator agreement regarding relations and a more coherent tree structure.

We propose, for the first time in the field of RST, a method for harmonizing rhetorical structure. Thanks to the methodology employed by the super-annotator to resolve disagreements, we avoided the use of either a training phase or an annotation guide (Carlson *et al.* 2001) based on signals for rhetorical relations, thus avoiding circularity between the annotation of relations and their signals (Spenader eta Lobanova 2009).

– **Regarding the evaluation method**:

We propose a qualitative-quantitative methodology for measuring agreement regarding rhetorical relations (Iruskieta *et al.* Forthcominga). Firstly, we conducted an in-depth analysis of the method used to date to evaluate RST structures, and identified a series of problems or drawbacks (van der Vliet 2010a; da Cunha eta Iruskieta 2010; Iruskieta *et al.* 2013b). Next, we proposed an evaluation method that avoided these drawbacks. The advantages of the evaluation method proposed here are as follows: *i*) it assigns the correct weight to agreement regarding rhetorical relations and enables confusion matrixes to be described in an appropriate manner. We used the confusion matrixes obtained from the qualitative evaluation to guide the super-annotator's harmonization work. *ii*) the factors evaluated (rhetorical relations, nuclearity, discourse units) are independent, thus providing a qualitative description of agreement regarding rhetorical relations. Moreover, this also provides a qualitative description of disagreement. *iii*) since translation strategies between the different disagreement types are described, the method can also be used for rhetorical structure level translation tasks.[5]

[5]This is clearly illustrated da Cunha eta Iruskietaren (2010) in the fact that it was men-

– **Regarding reliability**:

We proposed, for the first time in RST, a set of criteria for resolving inter-annotator disagreements by appointing a super-annotator. The super-annotator's work takes all annotation phases into account and means that annotators start each new phase from the same annotation base. We therefore describe each phase's inter-annotator agreement and disagreement rates from both a quantitative and qualitative perspective. The super-annotator's work helped us to increase the reliability that is so vital at the discourse level, and as a result, we had no need to propose here any training phase or rhetorical relation signal-based annotation guidelines aimed at enhancing reliability. This also has the added advantage of avoiding circularity between the relation and signal annotation processes (Spenader eta Lobanova 2009).

– **Regarding the annotation phases**:

The delivery phase is more developed than any other similar phase in either RST or any other rhetorical structure analysis theory. We have overcome the drawbacks present in other delivery phases carried out to date and offer the chance to conduct more advanced consultations. While some of the programs we used to achieve this are linked to the Basque language, such as those which automatically add linguistic information to non-annotated texts, others have no such link and may be of general interest to all those working within the field of RST. These programs include: *i*) a program which identifies rhetorical relations on the basis of a tree structure in *rs3* format, and *ii*) a program which retrieves signals for rhetorical relations from files in *RhetDB* format.

3.2 Conclusions

We shall now present the main conclusions drawn as a result of this thesis.

– **Regarding segmentation and rhetorical relations**:

According to Iruskieta *et al.*en (2011b), intra-sentential discourse segmentation is harder when carried out with no specific segmentation criteria, since inter-annotator disagreement is higher. In this study, the mean inter-annotator agreement rate in the GMB sub-corpus (abstracts of scientific works in the medical field) was 13.74% lower at the intra-sentential level than at the inter-sentential level.

That said, since no rhetorical relations can be established between units that were not segmented at an intra-sentential level with no specific

tioned in the text presented at the international conference entitled Genre- and Register-related Text and Discourse Features in Multilingual Corpora, held from 11 to 12 January 2013 in Brussels: `http://www.mariehaps.be/`.

segmentation criteria (the annotation process is not retroactive), no information was gathered regarding the rhetorical structure between no-segmented units. Consequently, the harmonization of the segments identified by annotators proved extremely useful, since it meant that less information was lost regarding rhetorical structure at this level.

Intra-sentential rhetorical structure, on the other hand, was easier to establish, since inter-annotator agreement was higher. The mean inter-annotator agreement for intra-sentential rhetorical relations was 14.19% higher than for inter-sentential rhetorical relations.

Thus, once disagreements regarding segmentation have been resolved, intra-sentential rhetorical structure is more reliable than inter-sentential rhetorical structure.

– **Regarding the central unit and rhetorical relations**:

Within the rhetorical relations linked to the central unit, those with an IMRaD structure appear most frequently: PREPARATION (26.77%), BACKGROUND (15.44%), MEANS (9.12%), PURPOSE (6.32%) and RESULT (4.21%) are the relations linked to the central unit which appear most frequently. Added to this, the general relations ELABORATION (17.19%) and LIST (6.32%) complete the list of relations linked to the central unit. In no case does the frequency of any of the other relations linked to the central unit exceed 3%.

– **Regarding rhetorical relations**:

Based on two phenomena found in our corpus, we can assert that if the rhetorical structure is built taking the central unit into account, then the resulting annotation is more reliable. In other words, inter-annotator agreement was greater in relation to the following two phenomena: *i*) inter-annotator agreement was between 10% and 30% higher when the central unit was identified, even though the probability of selecting the same central unit was smaller; *ii*) having the same central unit increases inter-annotator agreement by 6.17%, when measuring rhetorical relation agreement using the t-test ($p < 0.013$). *iii*) When measured using the t-test ($p < 0.000000001$), the F score for rhetorical relations linked to the central unit was 11.52% higher than for rhetorical relations not linked to the central unit, and was statistically significant.

– **Regarding signals for rhetorical relations**:

Thanks to signals, the problems posed by those approaches based on discourse markers (implicit relations (Taboada 2006) and ambiguity (van Dijk 1998)) are, to a certain extent avoided when identifying rhetorical relations. The result is a greater number of signaled relations,

which in turn gives us the opportunity of identifying more relations. The annotation of signals, on the other hand, is more subjective than approaches based on discourse markers. This became evident when we evaluated the annotation of certain signals for rhetorical relations; for example, in the RESULT signals of the cause subgroup, the agreement rate between three annotators was lower (37.31%), yet this same rate was higher in the GOAL signals (75.45%). Thus, the more phenomena we take to be signals, the more important it becomes to measure subjectivity.

3.3 Future work

This annotation project had a set of clear aims right from the start: to obtain a corpus annotated with rhetorical structure at the discourse level, to analyze inter-annotator disagreement problems and to ensure that the annotated information was as reliable as possible, in order to enable complex language processing tasks to be carried out using the annotated corpus. The theoretical framework chosen and the methodology we designed demonstrated to us that it is indeed possible to establish tools which can be used within the field of language processing to carry out tasks such as: discourse segmentation, automatic summarization and automatic discourse analysis, among others. Following on from the work carried out in this thesis, it would be interesting to pursue research aimed at extending or exploring in more detail the linguistic description of the different phenomena of the annotation phases. At the same time, it would also be interesting to explore ways of automatically identifying the phenomena described in detail in this thesis report. In this way, we may be able to make applications used in other languages available to the Basque language community.

In this section we will list some of the work that could be carried out in the mid-term, based on the results of this thesis. These possible future works will be described in the same order as the annotation phases to which they correspond.

- **Regarding the corpus**. In relation to size, the corpus section annotated by two or three annotators is comparable to other annotated corpus sections described in the literature. Nevertheless, the corpus has some constraints. Regarding the genre and domains of its texts, we annotated texts from a single genre (abstracts of scientific papers) and three domains (medicine, terminology and science). Consequently, other genres also need to be analyzed if the tool obtained is not to be linked only to a specific genre and set of domains.

  We should therefore annotate texts from different genres and domains. For example, it would be interesting to annotate a sample of journalism

texts from the EPEC corpus (Aldezabal *et al.* 2007b), since this corpus contains manually-annotated information at different language levels, and for certain tasks, this kind of information is more reliable than the automatically obtained kind.

Also, and following da Cunha *et al.* (2007), in addition to a sample from the EPEC corpus, we could also annotate a series of whole papers corresponding to the abstracts contained in the TERM sub-corpus (sub-corpus of abstracts within the field of terminology). In this way, by comparing the rhetorical structure of the whole text with that of its abstract, we would be able to determine which rhetorical relations to eliminate and which to maintain in order to carry out automatic summarization.

– **Regarding segmentation**:

  If a group of different linguists were to re-segment the corpus and we were to re-evaluate it, then we could measure the adequacy of the segmentation criteria proposed in this thesis.

  We could build an automatic discourse segmenter based on automatic learning.

– **Regarding nuclearity**:

  We believe that taking the most important unit in the paragraph into account, as well as the central unit, may have an influence on the annotation of rhetorical relations. To test this hypothesis, we could analyze the effect of adding another phase to the annotation process, i.e. annotating the key unit in the paragraph after annotating the central unit, but before annotating the rhetorical relations.

  Since we now have a corpus annotated with central units and have established a process for selecting the central unit, we can now detect a text's central unit on the basis of a set of rules or automatic learning. This in turn could be used to design an automatic summarization system based on the central unit.

  We could analyze the levels at which an abstract complies with the IMRaD structure. To this end, we could design a system to calculate the extent to which the key relations of the IMRaD structure are repeated.

– **Regarding rhetorical relations**:

  We could analyze the confusion matrix for rhetorical relations linked to the central unit, and compare it to the confusion matrix for relations in general, in order to determine whether or not they are the same.

  We could measure the subjectivity of the criteria used by the super-annotator to harmonize rhetorical relations. If we explored whether or

not two super-annotators obtained the same result when harmonizing a rhetorical tree with the same criteria, then we would be able to measure the adequacy of the criteria established for the super-annotator in our study.

Just as Maziero eta Pardo (2009) and Marcu (2000a) have automated their methodology, we could automate the qualitative-quantitative method used in this study to describe inter-annotator agreement, and adapt it to different languages.

Since we carried out an in-depth study of the signals for the rhetorical relations of the cause subgroup, we could analyze whether or not it would be possible to detect these rhetorical relations on the basis of rules or automatic learning.

– **Regarding relation signals**:

We could give the rhetorical relations not analyzed here to other annotators, so that they could annotate their corresponding signals. In this way, once the super-annotator has resolved any disagreements, the reliability of the relations would be greater.

Bibliografia

Aduriz, Itziar, Izaskun Aldezabal, Iñaki Alegria, JoseMari Arriola, Arantza Diaz de Ilarraza, Nerea Ezeiza, eta Koldo Gojenola. 2003. Finite state applications for basque. In *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*, Budapest, Hungary.

——, María Jesus Aranzabe, JoseMari Arriola, Arantza Diaz de Ilarraza, Koldo Gojenola, Maite Oronoz, eta Larraitz Uria. 2004. *A cascaded syntactic analyser for Basque*, 124–134. Computational Linguistics and Intelligent Text Processing. Springer.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria, E. Izagirre, K. Mendizabal, E. Pociello, eta M. Quintian. 2006a. A methodology for the joint development of the Basque WordNet and Semcor. In *5th International Conference on Language Resources and Evaluations (LREC)*, Genoa, Italy.

——, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, eta Karmele Mendizabal. 2005a. EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa. *Euskalingua* 237–266.

——, Izaskun Aldezabal, Jone Etxeberria, I. Izagirre, K. Mendizabal, E. Pociello, eta M. Quintian. 2005b. EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editatze-, etiketatze-eta epaitze-lanak. Technical report, EHU.

——, ——, ——, ——, ——, ——, eta ——. 2006b. Improving the Basque WordNet by corpus annotation. In *3rd International WordNet Conference*, 287–290, Jeju Island, Korea.

Aierbe, Axun. 2008. Birformulazio-estrategiak eta komunikagarritasuna administrazioko testuetan. Technical report.

Alberdi, Xabier. 2011a. Diskurtso-markatzaile berri bat: hurrenez hurren birformulatzailea. *ASJU* XLV.301–325.

——. 2011b. Erran nahi baita birformulatzaile esplikatiboa gaurko euskaran. *ASJU* XLV.

——, eta Julio Garcia. 2012. Diccionario de marcadores discursivos del euskera (i). In *V Congreso Internacional de Lexicografía Hispánica*, Madrid.

——, eta Josu Landa. 2013. EUDIMA corpusetik adibideak erauzteko lan-tresna: diskurtso unitate fraseologikoak aztertzeko lanabesa. *ASJU* 45.

Aldezabal, Izaskun, María Jesus Aranzabe, JoseMari Arriola, Arantza Diaz de Ilarraza, Ainara Estarrona, K. Fernandez, L. Uria, eta M. Quintian. 2007a. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua. Technical Report LSI/TR 12, UPV/EHU.

——, Klara Ceberio, I. Esparza, Ainara Estarrona, Jone Etxeberria, Mikel Iruskieta, E. Izagirre, eta L. Uria. 2007b. EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua. Technical Report LSI/TR 11, UPV/EHU.

Alegria, Iñaki, Xabier Artola, Arantza Diaz de Ilarraza, Kepa Sarasola, eta Itziar Aduriz. 2011. Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta ixa taldearen adibidea. *Linguamática* 3.13–31.

——, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, eta Ruben Urizar. 2003. Named entity recognition and classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información*, 1–8, Madrid.

Arrieta, Bertol, 2010. Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean. Doktore-tesia, Euskal Herriko Unibertsitatea, Donostia.

Asher, Nicholas, eta Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge Univ Pr.

Azkarate, Miren. 2013. *Hain zuzen (ere) diskurtso-markatzailea: lokailu ala operadore?*. Eridenen du zerzaz kontenta. Sailkideen omenaldia Henrike Knörr irakasleari (1947-2008). Bilbo: EHU.

Barandiaran, Asier, eta Manuel Casado. 2011. *Marcadores discursivos: calas contrastivas en los reformuladores del español y el euskera*, 375–396. Marcadores del discurso: de la descripción a la definición. Madrid-Frankfurt: Iberoamericana-Vervuert.

Barrutieta, Guillermo, Joseba Abaitua, eta Josuka Díaz. 2001. Grossgrained RST through XML metadata for multilingual document generation. In *MT Summit VIII*, 39–42, Santiago de Compostela, Spain.

Bouayad-Agha, Nadjet. 2000. Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts. In *38th Annual Meeting ACL*, volume 38, 16–22, Hong Kong.

Bunt, Harry, eta William Black. 2000. The abc of computational pragmatics. *Abduction, Belief and Context: Studies in Computational Pragmatics* 1–35.

Burstein, Jill C., Daniel Marcu, eta Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *Ieee Intelligent Systems* 18.32–39.

Cabré, MaríaTeresa. 1998. El discurs especialitzat o la variació funcional determinada per la temàtica: Noves perspectives. *Revista Internacional de Filologia* 173–194.

Carlson, Lynn, eta Daniel Marcu. 2001. Discourse tagging reference manual. Technical report.

——, ——, eta Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, p. 10, Aalborg, Denmark. Association for Computational Linguistics.

Carreras, Xavier, 2005. Learning and inference in phrase recognition: a filtering-ranking architecture using perceptron. Doktore-tesia, Universitat Politecnica de Catalunya.

Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational Linguistics* 13.11–24.

da Cunha, Iria, 2008. Hacia un modelo lingüístico de resumen automático de artículos médicos en español. Doktore-tesia, IULA, Universitat Pompeu Fabra.

——, eta Mikel Iruskieta. 2009. La influencia del anotador y las técnicas de traducción en el desarrollo de árboles retóricos. un estudio en español y euskera. In *7th Brazilian STIL*, Sao Carlos, Brazil.

——, eta ——. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies* 12.563–598.

——, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, eta Irene Castellón. 2010. Diseg: Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Natural* 45.

——, Juan-Manuel Torres-Moreno, eta Gerardo Sierra. 2011. On the Development of the RST Spanish Treebank. In *5th Linguistic Annotation Workshop (LAW V '11)*, 1–10, Portland, USA. Association for Computational Linguistics.

——, L. Wanner, eta MaríaTeresa Cabré. 2007. Summarization of specialized discourse. *Terminology* 13.249–286.

Diaz de Ilarraza, Arantza, Koldo Gojenola, eta Maite Oronoz. 2005. Design and Development of a System for the Detection of Agreement Errors in Basque. In *Computational Linguistics and Intelligent Text Processing*, 793–802. Springer.

Esnal, Pello. 2008. *Testu-antolatzaileen erabilera estrategikoa*, volume 51. Bilbo: Euskaltzaindia.

Euskaltzaindia. 1990. *Euskal gramatika. Lehen urratsak III (Lokailuak)*. Bilbo: Euskaltzaindia.

——. 1994. *Euskal gramatika: lehen urratsak (EGLU) VI (juntagailuak)*. Bilbo: Euskaltzaindia.

——. 1999. *Euskal gramatika. Lehen urratsak-V (mendeko perpausak-I)*. Bilbo: Euskaltzaindia.

——. 2005. *Euskal gramatika. Lehen urratsak-VI (Mendeko perpausak-II)*. Bilbo: Euskaltzaindia.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76.378–382.

Forbes, Katherine, Eleni Miltsakaki, Rashmi Prasad, A. Sarkar, Aravind Joshi, eta Bonnie L. Webber. 2003. D-ltag system: Discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Logic, Language and Information* 12.261–279.

Garcia, Julio, eta Mikel Iruskieta. 2013. *Birformulatzaile zuzentzaileak testu idatzietan*. Eridenen du zerzaz kontenta. Sailkideen omenaldia Henrike Knörr irakasleari (1947-2008). Bilbo: EHU.

García, Inés M. 2010. *Estrategias textuales y discursivas en el aprendizaje de la exposición oral de dos materias distintas*, 155–162. Modos y formas de la comunicación humana. Vigo: AESLA.

Ghorbel, Hatem, Azfal Ballim, eta Giovanni Coray. 2001. Rosetta: Rhetorical and semantic environment for text alignment. In *Corpus Linguistics*, 224–233, Lancaster University (UK).

Grosz, Barbara J., eta Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12.175–204.

Haouam, Kamel, eta Fahri Marir. 2003. SEMIR: Semantic indexing and retrieving web document using Rhetorical Structure Theory. In *4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, 596–604, Hong Kong.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive science* 3.67–90.

Hovy, Eduard. 1993. In defense of syntax: Informational, intentional, and rhetorical structures in discourse. In *Intentionality and Structure in Discourse Relations Workshop*, 35–39, Ohio, USA.

——. 2010. Annotation: A tutorial. In *48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

——, 2011. The Three (and a Half) Futures of NLP. http://www.iis.sinica.edu.tw/page/events/FILE/12031310107Slides.pdf.

Ibarra, Orreaga. 2013. Sobre estrategias discursivas del lenguaje de los jóvenes vascoparlantes: aspectos pragmáticos y discursivos (conectores, marcadores). *ASJU* 395–411.

Ide, Nancy, eta James Pustejovsky. 2010. What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. In *2nd Int. Conf. Global Interoperability Lang. Res*, Hong Kong.

Iruskieta, Mikel, 2012. Pragmatika. http://www.ehu.es/seg/hizk/1/6.

——, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, eta 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.

——, eta Iria da Cunha. 2010a. El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español. *Calidoscópio* 8.181–202.

——, eta ——. 2010b. Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera. In *XXVIII Congreso Internacional AESLA: Analizar datos > Describir variación*, 13–159, Vigo. Servicio de Publicaciones.

——, ——, eta Maite Taboada. Forthcominga. A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation* .

——, Arantza Diaz de Ilarraza, eta Mikel Lersundi. 2008. Análisis de los marcadores del discurso para el euskera: denominación, clases, relaciones semánticas y tipos de ambigüedad. In *Proceedings of 26th AESLA International Conference*, 1271–1282.

——, Arantza Diaz de Ilarraza, eta Mikel Lersundi. 2009. Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso. In *Proceedings of 27th AESLA International Conference*, 963–971.

——, Arantza Diaz de Ilarraza, eta Mikel Lersundi. 2011a. Bases para la implementación de un segmentador discursivo para el euskera. In *8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*.

——, Arantza Diaz de Ilarraza, eta 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.

——, Arantza Diaz de Ilarraza, eta Mikel Lersundi. 2013b. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory* 0.1–32.

——, Arantza Diaz de Ilarraza, eta Mikel Lersundi. Forthcomingb. Detecting the central unit in rhetorical structure trees: A key step in annotating rhetorical relations.

Jurafsky, Dan, eta James H. Martin. 2000. *Speech and Language Processing*. Pearson Education India.

Karlsson, F., A. Voutilainen, J. Heikkila, eta A. Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Larringan, Luis, 1995. Testu-antolatzaileak bi testu motatan: testu informatiboa eta argudiapenezkoa. Doktore-tesia, Euskal Herriko Unibertsitatea, Gasteiz.

Litman, Diane J., eta James F. Allen. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science* 11.163–200.

Mann, Willian C., eta Maite Taboada, 2010. RST web-site. http://www.sfu.ca/rst/.

——, eta Sandra A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. *Text* 8.243–281.

Marcu, Daniel. 2000a. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics* 26.395–448.

——. 2000b. *The theory and practice of discourse parsing and summarization*. Cambridge: The MIT press.

Matthiessen, C., eta Sandra A. Thompson. 1987. *The Structure of Discourse and 'Subordination'*, p. 328. Clause Combining in Discourse and Grammar. John Benjamins.

Maziero, Erick G., eta Thiago A. S. Pardo. 2009. Metodologia de avaliação automática de estruturas retóricas. In *7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, eta Bonnie L. Webber. 2004. Annotating discourse connectives and their arguments. In *HLT/NAACL Workshop on Frontiers in Corpus Annotation*, 9–16, Boston, USA.

Mitocariu, Elena, Daniel Alexandru Anechitei, eta Dan Cristea. 2013. *Comparing Discourse Tree Structures*, 513–522. Computational Linguistics and Intelligent Text Processing. Springer.

Moser, Megan, Johanna D. Moore, eta E. Glendening. 1996. Instructions for coding explanations: Identifying segments, relations and minimal units. Technical Report Technical Report 96-17, Department of Computer Science.

O'Donnell, Michael. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *First International Conference on Natural Language Generation INLG '00*, volume 14, 253–256, Mitzpe Ramon. ACL.

Ono, Kenjl, Kazuo Sumita, eta Seijl Miike. 1994. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, 344–348. Association for Computational Linguistics.

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*, 172–191, Cambridge. Butterworth and Co.

Pardo, Thiago A. S., 2005. Métodos para análise discursiva automática. Master's thesis.

——, eta Eloize R. M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. *Anais do V Encontro de Corpora* 24–25.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12.601–638.

Rino, Lucia H. M., eta Donia R. Scott. 1996. A discourse model for gist preservation. *Advances in Artificial Intelligence* 131–140.

Salaburu, Peio, 2012. Menderakuntza eta menderagailuak (Sareko Euskal Gramatika: SEG). http://www.ehu.es/seg/morf/5/2/2/2.

Soricut, R., eta 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, 149–156. Association for Computational Linguistics.

Spenader, Jennifer, eta Anna Lobanova. 2009. Reliable discourse markers for contrast relations. In *8th International Conference on Computational Semantics*, Tilburg, The Netherlands.

Stede, Manfred. 2008a. Disambiguating rhetorical structure. *Research on Language and Computation* 6.311–332.

——. 2008b. *RST revisited: Disentangling nuclearity*, 33–57. 'Subordination' versus 'coordination' in sentence and text. Amsterdam and Philadelphia: John Benjamins.

Taboada, Maite. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics* 38.567–592.

——, eta Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse* 4.249–281.

——, eta Willian C. Mann. 2006. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies* 8.423–459.

——, eta Jan Renkema, 2011. Discourse relations reference corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Thompson, Sandra A., Robert Longacre, eta Shin Ja J. Hwang. 1985. *Adverbial clauses*, volume 2 of *Language Typology and Syntactic Description: Complex Constructions*, 171–234. New York: Cambridge University Press.

Tofiloski, Milan, Julian Brooke, eta Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, 77–80, Suntec, Singapore. ACL.

Uria, L., Ainara Estarrona, Izaskun Aldezabal, María Jesus Aranzabe, Arantza Diaz de Ilarraza, eta Mikel Iruskieta. 2009. Evaluation of the Syntactic Annotation in EPEC, the Reference Corpus for the Processing of Basque. *Computational Linguistics and Intelligent Text Processing* 72–85.

Urizar, Ruben, 2012. Euskal lokuzioen tratamendu konputazionala. Master's thesis.

Urrutia, Andrés. 2008. Legeen eta administrazioaren hizkera, testu-antolatzaileen ikuspegitik. *Euskera: Euskaltzaindiaren lan eta agiriak* 53.525–546.

van der Vliet, Nynke, 2010a. Inter annotator agreement in discourse analysis. http://www.let.rug.nl/ñerbonne/teach/rema-stats-meth-seminar/.

——. 2010b. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*, 203–210, Ljubljana, Slovenia.

——, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, eta Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. *Bochumer Linguistische Arbeitsberichte* 3.157–171.

van Dijk, Teun A. 1980a. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. L. Erlbaum Associates Hillsdale, NJ.

—— 1980b. The semantics and pragmatics of functional coherence in discourse. *Speech act theory: Ten years later, Versus* 26.49–65.

—— 1997. *The study of discourse: An Introduction*, volume 1 of *The study of discourse*, 1–34. London: Sage.

—— 1998. *Texto y contexto: semántica y pragmática del discurso*. Cátedra.

Wolf, Florian, eta Edward Gibson. 2004. Representing discourse coherence: A corpus-based analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, 134–140. Association for Computational Linguistics.

——, eta ——. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics* 31.249–287. PT: J; UT: ISI:000230227800005.

Zabala, Igone. 1996. *Testuan iruzkinak sartzeko funtzioa izan dezaketen antolatzaileak: hau da, hain zuzen ere, adibidez, batez ere*, 113–130. Testuloturarako baliabideak: euskara teknikoa. Bilbo: EHU.

Zapirain, Beñat. 2004. Xiraba: tresna linguistikoen sendotasuna ebaluatzeko tresna. Technical report.

4

Bases para la implementación de un segmentador discursivo para el euskera

Bases para la implementación de un segmentador discursivo para el euskera

Mikel Iruskieta[1], Arantza Diaz de Ilarraza[2], Mikel Lersundi[3]

IXA group for NLP. Faculty of informatics. University of the Basque Country
Post code 10018 – Donostia – Basque Country - Spain

[1]Department of Didactics of Language and Literature
[2]Department of Computer Science
[3]Department of Basque Philology

{mikel.iruskieta, a.diazdeilarraza, mikel.lersundi}@ehu.es

Abstract. *In this paper we study how to adapt an automatic clause parser to discourse segmentation task. Considering a manually tagged corpus according to Rhetorical Structure Theory (RST), we have processed it with an automatic clause parser and the results were studied by comparing the agreement between both annotation systems: automatic and manual. As a result of this comparison we indicate where the intersection among the automatic clause segmentation and discursive segmentation is.*
Keywords. *Discourse segmentation; Rhetorical Structure Theory, parser.*

Resumen. *Presentamos un estudio para adaptar el segmentador automático de cláusulas y oraciones de carácter general para el euskera a la tarea de segmentación discursiva. Partiendo de un corpus anotado manualmente según la Rethorical Structure Theory (RST), hemos procesado el texto de manera automática por medio del segmentador automático y hemos estudiado los resultados comparando las coincidencias y desacuerdos entre la anotación automática y la manual. Los resultados de esta comparación señalan los criterios comunes para adaptar el segmentador a tareas discursivas.*
Palabras clave. *Segmentación discursiva, Teoría de la Estructura Retórica, segmentador.*

1. Introducción

En este artículo presentamos un estudio para adaptar el segmentador de cláusulas y oraciones de carácter general que disponemos para el euskera a la tarea de segmentación discursiva. El segmentador que analizamos ha sido utilizado, en concreto, para tareas de corrección de puntuación en textos (Arrieta 2010) y está implementado mediante la combinación de gramáticas basadas en reglas y técnicas de aprendizaje automático.

En este trabajo trataremos de responder a las siguientes cuestiones: ¿es adecuado abordar la tarea de la segmentación discursiva partiendo de un segmentador de cláusulas y oraciones de carácter general?, ¿cuáles son los criterios comunes entre la segmentación sintáctica y la segmentación discursiva?, y ¿cuándo podemos concluir que es aceptable la segmentación automática discursiva?

Aplicaciones avanzadas, tales como la búsqueda de información basada en conocimiento semántico, la elaboración automática de resúmenes o la traducción automática, precisan herramientas sofisticadas de procesamiento del lenguaje que, a su vez, necesitan basarse en el conocimiento presente en el corpus. Por ello, y para poder llevar a cabo este tipo de aplicaciones, es necesario contar con corpus de referencia

Anais do III Workshop "A RST e os Estudos do Texto", páginas 18–29, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

etiquetados a diferentes niveles lingüísticos: fonético, morfológico, sintáctico o discursivo.

El etiquetado de corpus de referencia en cualquiera de los niveles de análisis lingüístico tiene como primer paso la segmentación. Ésta consiste en identificar y marcar las unidades básicas a considerar en cada nivel lingüístico de análisis, para después determinar las relaciones entre dichas unidades. La identificación de fonemas y su anotación en los corpora es una tarea necesaria para el tratamiento del habla, como es la identificación por un lado de lexemas y morfemas, y por otro de sintagmas y dependencias son necesarias en el etiquetado de corpora a nivel morfológico y sintáctico. También es ineludible la segmentación a nivel discursivo para identificar la estructura relacional de un texto. Este trabajo trata precisamente de la segmentación de este último nivel: el nivel discursivo.

Atendiendo a la granularidad con la que se establece la unidad de discurso, encontramos en la literatura diferentes propuestas para la segmentación discursiva. Las propuestas varían según la aproximación teórica usada y según la finalidad para la que se realiza el trabajo de etiquetado. En general podemos distinguir entre dos niveles en la segmentación discursiva: segmentación de nivel alto y segmentación de nivel bajo. En esta última, a su vez se distinguen dos subniveles: intra-oracional (mayor granularidad) e inter-oracional (menor granularidad). Por ejemplo la segmentación intra-oracional donde se establecen unidades de discurso a nivel de cláusula es utilizada en Marcu (2000) para tareas de resumen automático. La segmentación de alto nivel donde se establecen pasajes o párrafos es utilizada en tareas de recuperación de la información (Girill 1991) o detección de cambios de tópico (Hearst 1997). En este trabajo abordaremos la segmentación intra-oracional, ya que nuestro objetivo es la anotación de corpus válidos para una amplia variedad de aplicaciones.

En la literatura se referencian segmentadores de discurso "independientes de lenguaje" (Kiss y Strunk 2006) que detectan segmentos únicamente a nivel inter-oracional. En el corpus sobre el que hemos trabajado los segmentos a nivel intra-oracional suponen alrededor de un 9%. En la actualidad conocemos herramientas de segmentación discursivas de nivel bajo para inglés, portugués y español (Tofiloski, Brooke y Taboada 2009, Pardo 2006, da Cunha*, et al* 2010). Hasta el momento no existe una herramienta de dichas características en euskera y este es el objetivo que nos proponemos a corto plazo. Este trabajo supone un paso importante en la consecución en ese objetivo.

2. Estado del arte: teorías y corpora

Existen diferentes teorías discursivas que formalizan la estructura referencial; cada una de estas teorías proporciona corpora anotados según sus criterios: i) Segmented Discourse Representation Theory (SDRT) (Asher y Lascarides 2003); ii) Discourse-Lexicalized Tree Adjoining Grammar (D-LTAG) (Webber*, et al* 2003); y, iii) Rhetorical Structure Theory[i] (RST) (Mann y Thompson 1987). Esta última teoría describe la coherencia y relación entre fragmentos textuales haciendo corresponder la idea de nuclearidad, o importancia de un fragmento del discurso, con el efecto que produce en el lector la presentación de dicha relación. Cuenta con varios corpus para diferentes lenguas: i) para el inglés, un corpus de 385 textos periodísticos (Carlson, Okurowski y

Marcu 2002) y otro de 65 textos de géneros diferentes (Taboada y Renkema 2011); ii) para el español un corpus de 267 textos (da Cunha, Torres-Moreno y Sierra 2011); iii) para el portugués el corpus TCC de 100 textos científicos (Pardo y Nunes 2006), y iv) para el alemán el corpus PCC de 170 textos etiquetados (Stede 2004). Existen segmentadores discursivos para el inglés (Marcu 2000, Tofiloski, Brooke y Taboada 2009), para el portugués (Pardo y Nunes 2008) y el español (da Cunha*, et al* 2010)[ii]. La RST ha sido implementada para diversas aplicaciones de PLN según Taboada y Mann (2006a).

El marco teórico sobre el que desarrollamos este estudio empírico es la RST. Según esta teoría, las relaciones que se establecen entre los segmentos del texto pueden ser paratácticas (N-N)[iii], cuando se establece la relación entre fragmentos con el mismo grado de importancia en la intención del autor (LISTA, CONTRASTE, DISYUNCIÓN…), o hipotácticas (N-S), cuando se establece una relación entre una unidad menos importante con otra más importante en cuanto a la intención del autor (ELABORACIÓN, MÉTODO, PREPARACIÓN, CONCESIÓN, CAUSA, RESULTADO…). Las relaciones se definen en base a las restricciones presentes entre el núcleo y satélite, y describiendo el efecto que crea en el lector.

El corpus sobre el que hemos realizado el estudio es un corpus de resúmenes de artículos médicos extraídos de la Gaceta Médica de Bilbao[iv], que contiene todos los resúmenes de artículos en euskera desde sus inicios en el año 2000 hasta el 2008. El corpus está compuesto por 20 documentos y contiene 273 unidades elementales de discurso (EDU); a nivel intra-oracional cada EDU tiene como media unas 11 palabras, y el corpus tiene 3.024 palabras. Este corpus ha sido utilizado en trabajos anteriores (da Cunha y Iruskieta 2010) donde se sugiere que se pueden detectar estrategias de traducción mediante la comparación de árboles retóricos en idiomas diferentes. La anotación de este corpus está disponible tanto en español (da Cunha, Torres-Moreno y Sierra 2011) como en euskera[v].

Aunque en la RST existen diferentes propuestas para la segmentación de textos, el corpus en el que nos basamos se ha segmentado siguiendo la definición original de unidad básica de Mann y Thompson (1987) que dice fundamentarse en una clasificación teórica neutral en la que las unidades debieran caracterizarse por una integridad funcional independiente.

3. Segmentación manual y automática. Comparación

Para determinar si el segmentador automático es un buen punto de partida en la construcción de un segmentador discursivo, vamos a comparar el resultado del segmentador de cláusulas y oraciones con nuestra anotación discursiva manual que sigue la segmentación original de la RST y establecer criterios comunes para definir reglas básicas de implementación válidas en el marco de la RST.

En lo referente a la segmentación manual, y, tras un proceso escalonado para establecer los criterios de segmentación (Iruskieta, Díaz de Ilarraza y Lersundi En prensa), se han fijado las siguientes reglas de segmentación a nivel inter-oracional e intra-oracional: i) en el nivel inter-oracional se van a considerar unidades de discurso aquellas oraciones con verbo conjugado no subordinadas[vi], y ii) en el nivel intra-oracional se consideran unidades de discurso oraciones con verbo (tanto conjugado

como no conjugado). Los complementos verbales no se consideran unidad del discurso aunque posean formas verbales (por ejemplo, complementos de verbos declarativos).

En referencia a la segmentación automática, nuestro sistema para la segmentación del corpus médico utiliza el sistema descrito en (Alegria*, et al* 2008) que identifica cláusulas mediante la combinación de gramáticas basadas en reglas y técnicas de aprendizaje automático. Las reglas establecen los puntos donde finalizan las oraciones y mediante las técnicas de aprendizaje automático se reconocen el comienzo y final de las estructuras sintácticas parciales basándose en la información lingüística asociada a cada palabra de la oración (Carreras 2005). La información se ha obtenido tras la aplicación de la siguiente secuencia de tratamientos lingüísticos:

1. Análisis morfo-sintáctico (MORPHEUS[vii] (Aduriz*, et al* 1998)). Proceso por el cual se establece la segmentación de cada palabra, su categoría, subcategoría y otras características lingüísticas tales como caso, número, etc. El principal problema de este paso de análisis es la gran cantidad de análisis asociados a cada palabra, ya que el análisis de la palabra se realiza sin tomar en cuenta el contexto en el que se encuentra.

2. Lematización e identificación de funciones sintácticas. Estos dos procesos se realizan en secuencia mediante la aplicación EUSTAGGER[viii] (Aduriz*, et al* 2003). La principal tarea del lematizador es resolver la ambigüedad que resulta del proceso de análisis morfo-sintáctico tratando de dar un único análisis para cada palabra de la frase basándose en la información contextual. La identificación de funciones sintácticas se realiza mediante reglas basadas en conocimiento lingüístico que siguen el formalismo establecido en las gramáticas de restricciones (Karlsson*, et al* 1995).

3. Identificación de unidades multi-palabra cuyo objetivo es determinar las unidades que se componen de dos o más palabras, considerando sólo los casos en que estas asociaciones de palabras sean siempre fijas.

4. Identificación de entidades nombradas (EIHERA[ix] (Alegria*, et al* 2003)).

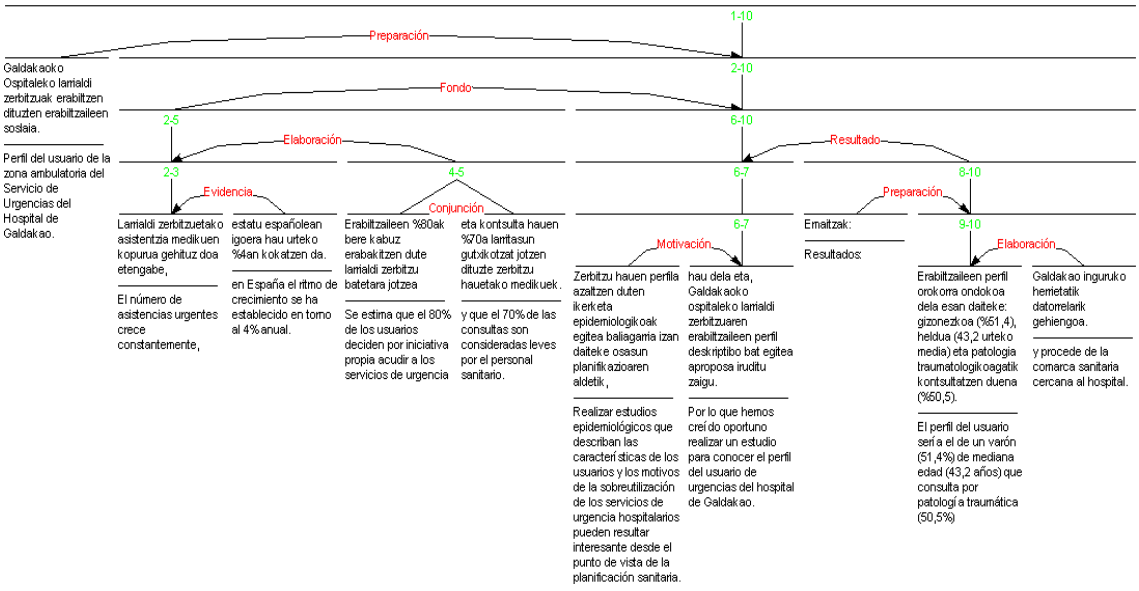


Figura 1. Árbol retórico (GMB_04_01)

En la Tabla 1 se presenta de forma gráfica la segmentación manual y automática[x] del ejemplo (1) tomado del texto del corpus representado en la Figura 1.

(1) a. <[<*Erabiltzaileen %80ak bere kabuz erabakitzen dute*> <*larrialdi zerbitzu batetara jotzea*>] [<*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*>]> GMB_04_01

b. <[<Se estima que el 80% de los usuarios deciden por iniciativa propia> <acudir a los servicios de urgencia>] [<y que el 70% de las consultas son consideradas leves por el personal sanitario.>]>

EDUs en segmentación manual		**EDUs en segmentación automática**		
M1	**M2**	**A1**	**A2**	**A3**
Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea	*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*	*Erabiltzaileen %80ak bere kabuz erabakitzen dute*	*larrialdi zerbitzu batetara jotzea*	*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*
Se estima que el 80% de los usuarios deciden por iniciativa propia acudir a los servicios de urgencia	y que el 70% de las consultas son consideradas leves por el personal sanitario.	Se estima que el 80% de los usuarios deciden por iniciativa propia	acudir a los servicios de urgencia	y que el 70% de las consultas son consideradas leves por el personal sanitario.

Tabla 1. Comparación segmentaciones (fragmento de GMB_04_01)

Como hemos comentado la segmentación automática tiene un componente basado en reglas mediante las que se establece la identificación de limites clausales y oracionales. Presentamos en la Tabla 2, a modo de ejemplo, dos de las reglas que se aplicarían para identificar los limites clausales en el texto del ejemplo (1).

N°	**Explicación de la regla**
11	`MAP (}MUGA) TARGET (ADL) IF (1 (LOT)+(JNT))` `(NOT 1 ("baita")OR("ezta"));`

68	`MAP (}MUGA) TARGET (ADIZE) IF (0 (DEK)) (NOT 1 PUNTUAZIOA)`
	`(NOT 1 ("aritu")+(ADOIN))`
	`(NOT -2 ("aritu")+(ADOIN));`

Tabla 2. Reglas de segmentación utilizadas en el ejemplo 1

La regla 11 asigna la marca de fin de segmento de A1 tras el verbo auxiliar (ADL) (*erabakitzen dute)* 'deciden', si y solamente si: i) viene seguido un conector (LOT) y que es a su vez conjunción coordinante (JNT) y ii) inmediatamente a la derecha del auxiliar no están las palabras *baita* 'también' y *ezta* 'tampoco'.[xi]

La regla 68 asigna la frontera de A2 a la nominalización (ADIZE) *jotzea* 'pegar', si y solamente si la nominalización posee alguna marca de declinación (DEK) y i) no tiene signos de puntuación a su derecha, ii) no tiene el verbo *aritu* 'ocuparse'[xii] más una forma verbal sin terminación aspectual (ADOIN) a su derecha o iii) a una distancia de dos palabras a la izquierda. Esta regla y el final del segmento anterior A1 son suficientes para determinar el segmento A2.

Entre la segmentación manual y automática hay diferencias de granularidad que indican que la segmentación automática es más fina que la manual, ya que se consideran criterios más relacionados con la función sintáctica que cumplen las oraciones (esto no se ajusta a lo establecido por nuestras guías de anotación previamente definidas). Por ejemplo, la segmentación automática considera la nominalización *jotzea* 'acudir' como EDU y la segmentación manual lo descarta por considerarse un complemento verbal y, por tanto, no considerar que exista una relación RST.

Las demás reglas de la gramática, al igual que las reglas explicadas que son utilizadas por el segmentador, determinan únicamente el final de cada segmento. Esto no es problema cuando al finalizar un segmento empieza otro, tal como sucede en el ejemplo (1); pero para detectar, además de estos segmentos en secuencia, segmentos subsumidos en otros (como en el ejemplo (2) donde la unidad 3 de la Figura 2, una cláusula adverbial de modo que se enlaza, en este caso, con la relación de MÉTODO, está subsumida dentro de otra unidad formalizada por la construcción SAME-UNIT) hemos utilizado técnicas de aprendizaje automático. Este problema es más crítico en un corrector de signos de puntuación. En la segmentación intra-oracional, sin embargo, las unidades que rompen una EDU no son tan abundantes; en este corpus dichas construcciones constituyen únicamente el 0,03% de todas las unidades.

(2) a. <[<*Ikerketa berriek,*} [*"microarrays" teknika erabiliz,*>] {<*pronostiko txarra duen*> *bularreko minbiziaren azpitalde bat hauteman dute*.]> GMB_07_02

b. <[<Estudios recientes} [utilizando la técnica de "microarrays">] {<han identificado un subgrupo de cánceres de mama <con pésimo pronóstico.>]>

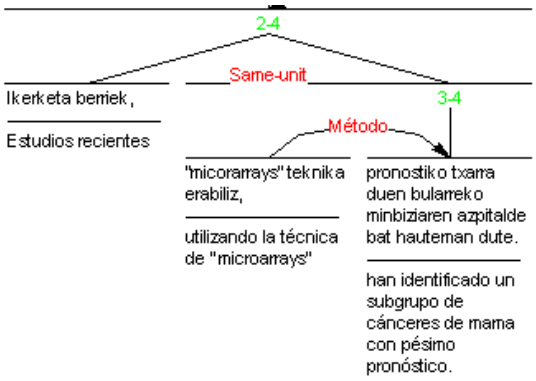


Figura 2. Árbol retórico del ejemplo 2

4. Resultados y evaluación

Actualmente podemos encontrar herramientas de segmentación automática con un *F-measure* en torno a 80%: *SLSeg* en inglés obtiene un 79% de *F-measure* (Tofiloski, Brooke y Taboada 2009) y *DiSeg* en español un 80% (da Cunha*, et al* 2010). Aunque los datos del segmentador automático general que utilizamos son bajos, un *F-measure* de 57,81%, se detectan la mayoría de EDUs; parte de los segmentos (S) que no se detectan se debe a que algunos segmentos intra-oracionales se formalizan de modo diferente, como cláusulas adverbiales y coordinaciones de EDUs. En la Tabla 3 se presentan los acuerdos obtenidos sobre las marcas de inicio de segmento (2ª fila: <S), final de segmento (3ª fila S>), donde el acuerdo es mayor, y la comparación entre las marcas de principio y final de segmento automático y marcas de EDUs (3ª fila: EDU), donde el acuerdo baja considerablemente. La cobertura que mide el grado de marcas automáticas que coinciden con las manuales señala que se han detectado la mayoría de las EDUs de un modo más que aceptable el final de cada segmento. Sin embargo la precisión, que mide el grado de marcas correctas de todas las marcas puestas por el segmentador, disminuye de modo considerable lo que indica una granularidad mayor del segmentador.

	Automático	Manual	Acuerdo	Cobertura	Precisión	F-measure
<S	450	273	223	81,68%	49,56%	61,69%
S>	450	273	242	88,64%	53,78%	66,94%
EDU	450	273	209	76,56%	46,44%	57,81%

Tabla 3. Evaluación del segmentador

Hemos hecho un estudio en detalle para ver por qué no coinciden los segmentos marcados automáticamente con los anotados manualmente y hemos detectado dos fenómenos:

i) Sobre-segmentación: el segmentador automático identifica más segmentos de los anotados manualmente. El segmento A2 *larrialdi zerbitzu batetara jotzea* 'acudir a los servicios de urgencia' del segmentador no se ha considerado en el modo manual, ya que el verbo nominalizado es parte en un sintagma nominal y su relación es puramente sintáctica con referencia al segmento A1.

ii) Falta de segmentación: el segmentador automático no detecta algunos segmentos o no formaliza adecuadamente una EDU con ambas marcas de inicio y final adecuadamente. El segmentador al establecer el segmento M1 *erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea* 'se estima que el 80% de los usuarios acuden a los servicios de urgencia por iniciativa propia' en varios segmentos A1 y A2, no formaliza de manera adecuada dicho segmento, es decir que las marcas de inicio y final no coinciden con las de una EDU. En otros casos la falta de segmentación es debida a diferentes modos de formalización en la segmentación automática y manual.

La Tabla 4 y la Tabla 5 muestran numéricamente la frecuencia de aparición de estos dos fenómenos: sobre-segmentación y falta de segmentación, respectivamente. Explicaremos brevemente los casos en que se dan estos fenómenos e intercalaremos algún ejemplo para dar una mayor claridad a la explicación.

En cuanto al fenómeno de la sobre-segmentación hemos identificado los siguientes casos: i) la segmentación automática ha detectado una oración principal, pero no incluye todas las palabras incluidas en la segmentación manual (Oración incompleta): ejemplo (3). El primer segmento automático es adecuado, porque coincide con el segmento manual, pero el segundo segmento automático, no considerado en la segmentación manual, recoge sólo en parte la oración principal; ii) la segmentación automática no formaliza adecuadamente y agrupa varias EDUs, por ejemplo, cuando establece como una EDU construcciones que componen más de una unidad, (Composición de EDUs); el ejemplo (4) se debe a una diferencia de formalización porque la marca de inicio de la segunda EDU no se ha colocado en la posición del segmento manual donde finaliza la oración subordinada adverbial sino que se ha colocado a su inicio, segmentando de este modo toda la oración compuesta[xiii]; iii) los segmentos corresponden a complementos de verbo (complementos, oraciones interrogativas indirectas, nominalizaciones…) y/o modificadores de sintagmas nominales (oraciones de relativo) que no consideramos EDUs en la segmentación manual (Complemento); en el ejemplo (5) observamos una oración relativa; iv) el segmentador identifica como unidad la coordinación de varios complementos o elementos coordinados sintácticamente que no constituyen EDUs (Coordinación de complementos), y finalmente v) cláusulas que no se han considerado

EDUs en la segmentación manual y se han segmentado automáticamente debido a la puntuación (Puntuación).

Oración incompleta	Composición de EDUs	Complemento	Coordinación	Puntuación	Total
13	26	87	74	38	238
5,46%	10,92%	36,55%	31,09%	15,97%	100,00%

Tabla 4. EDUs sobre-segmentados

(3) a. <[1996ko urtarriletik 1996ko ekainera arte, kolapsoterapia hartzen duten 30 gaixo, <batez beste 70.8 ±17 urtekoak (60-83 urte), aztertu ditugu guztira.>]> GMB_00_01

b. <[Desde Enero de 1996 hasta Junio de 1996 <hemos revisado a un total de 30 pacientes con colapsoterapia, con 70.8±17 años (60-83 años) de edad media.>]>[xiv]

(4) a. <[<Prebentzio metodoen eta artroplastiako teknika modernoen laguntzaz horrelako kasuak murriztu diren arren,>] [infekzio hori sendatzea erronka bat da oraindik ere.]> GMB_08_02

b. <[<Aunque su incidencia ha disminuido a lo largo de los años gracias a la evolución de los métodos de prevención y a las técnicas de artroplastia modernas,>] [su tratamiento sigue siendo un reto.]>

(5) a. <[<eta gaur egunera arte deskribatu diren adibideetan daukaten> maiztasuna alderatu da.]>GMB_05_03

b. <[y se compara su frecuencia <entre las series más numerosas de la literatura descritas hasta la actualidad.>]>

En cuanto al fenómeno de la falta de segmentación hemos identificado estos otros casos: i) una oración principal no detectada (Oración principal); en el ejemplo (6) la segmentación automática formaliza de forma diferente el primer segmento, ya que su cierre se introduce al final de la oración y no antes del segundo segmento, por lo que no coinciden las marcas de inicio y final de ambas segmentaciones; ii) no se detectan cláusulas adverbiales (Cláusula adverbial); iii) no se formalizan adecuadamente las unidades por separado que están coordinadas (Coordinación), y finalmente iv) EDUs que no se segmentan de modo adecuado debido a la puntuación (Puntuación).

Oración principal	Cláusula adverbial	Coordinación	Puntuación	Total
20	20	7	17	64
31,25%	31,25%	10,94%	26,56%	100,00%

Tabla 5. EDUs falta de segmentar

(6) a. <[Ultzera mingarri batzuk bezala agertzen da,] [<tamainu, kokapena eta iraunkortasuna aldakorra izanik.>]> GMB_03_01

b. <[Se caracteriza por la aparición de úlceras dolorosas] [<siendo de tamaño, localización y duración variable.>]>

Por último presentamos las unidades discursivas detectadas correctamente por el segmentador automático (Tabla 6).

Únicamente principal	Principal con subordinación	Cláusula adverbial	Yuxtap. o coordinación	Puntuación	Título	Total
64	57	18	51	6	13	**209**
30,62%	27,27%	8,61%	24,40%	2,87%	6,22%	**100,00%**

Tabla 6. EDU detectados correctamente

La comparación realizada entre la anotación de segmentos realizada automáticamente y la manual nos señala cómo adaptar la herramienta automática a la segmentación de discurso. En la Tabla 7 presentamos las conclusiones de la comparación.

	Forma lingüística	Segmentador general	Segmentador discursivo
Principios generales	Oración o cláusula verbal	sí	sí
	Same-unit construcción	sí	sí
Subordinación	Cláusulas adverbiales	sí	sí
	Complementos con clausulas verbales	sí	no
	Oración interrogativa indirecta	sí	no
	Cláusulas comparativas	ssi cláusulas verbales	no
	Nominalización	sí	no
	Clausulas de relativo	sí	no
Coordinación y/o yuxtaposición	de cláusulas verbales que difieren en un argumento	sí	sí
	de cláusulas verbales sin argumentos propios	sí	no
	de clausulas adverbiales	sí	sí
	de clausulas no-adverbiales	sí	no
	de cláusulas no verbal con marcador	sí	no
	Locuciones con función relacional	sí	no
Puntuación	Cláusulas verbales parentéticas	sí	sí
	Cláusulas no-verbales parentéticas	no	no
	Cláusulas de aposición	no	no
	Punto oracional con o sin verbo	sí	sí
	Dos puntos	sí	ssi EDU después
	Punto y coma	sí	ssi EDU después

Tabla 7. Criterios generales de adaptabilidad

5. Conclusiones y trabajo futuro

El estudio demuestra que aunque el porcentaje de EDUs segmentados correctamente (precisión en Tabla 3) por el segmentador automático es bajo y, por ello dicha segmentación no es la adecuada para la posterior anotación retórica en el marco de la RST; el método seguido para lograr un segmentador discursivo automático es un buen punto de partida, ya que el segmentador ha segmentado adecuadamente la mayoría de EDUs (cobertura Tabla 3). Para lograr ese objetivo, hemos detectado de manera precisa en qué situaciones no coinciden las marcas identificadas por el segmentador automático y qué nuevos criterios de segmentación debemos incorporar en el segmentador

automático, lo que supone un primer paso en la consecución de segmentador de discurso automático válido para la RST.

Teniendo en cuenta que el segmentador automático del que partimos está basado en reglas lingüísticas y algoritmos de aprendizaje automático, en el futuro nos proponemos realizar la tarea de adaptación de algunas de las reglas y adición de nuevas reglas del segmentador automático. Además tendremos que llevar a cabo el reentrenamiento del componente basado en aprendizaje automático tomando como base el corpus etiquetado que se obtendría al aplicar la gramática "adaptada" a un conjunto de textos (corpus de entrenamiento).

Agradecimientos

Este trabajo ha sido realizado en el marco de los siguientes proyectos: Grupo IXA, Grupo consolidado 2010-2015 (IT344-10) (Gobierno Vasco); KNOW2: Tecnologías de comprensión del lenguaje para el acceso multilingüe a la información orientada a dominios (TIN2009-14715-C04-01) (MICINN); Hibrido Sint: analizadores sintácticos basados en reglas y estadísticos. Integración en una plataforma para gestión de corpus basada en estándares XML (TIN2010-20218) (MICINN); Desarrollo de un entorno para extraer terminología y neología a partir de corpus etiquetados lingüísticamente GARATERM2 (US10/01).

Referencias

Aduriz, I., E. Agirre, I. Aldezabal, I. Alegria, O. Ansa, X. Arregi, J. Arriola, X. Artola, A. Díaz de Ilarraza y N. Ezeiza, 1998. A framework for the automatic processing of Basque. En *Proceedings of the First International Conference on Language Resources and Evaluation*.

Aduriz, I., I. Aldezabal, I. Alegria, J. Arriola, A. Díaz de Ilarraza, N. Ezeiza y K. Gojenola, 2003. Finite state applications for basque. En *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*.

Alegria, I., B. Arrieta, X. Carreras, A. Díaz de Ilarraza y L. Uria, 2008. Chunk and clause identification for basque by filtering and ranking with perceptrons. *Procesamiento del lenguaje natural,* (41): 5-12.

Alegria, I., I. Balza, N. Ezeiza, I. Fernandez y R. Urizar, 2003. Named entity recognition and classification for texts in basque. En *II Jornadas de Tratamiento y Recuperación de Información*, 1-8.

Arrieta, B., 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* Tesis doctoral. EHU: Euskal Herriko Unibertsitatea.

Asher, N. y A. Lascarides, 2003. *Logics of conversation.* Cambridge Univ Pr, Cambridge.

Carlson, L., M.E. Okurowski, D. Marcu, 2002. RST Discourse Treebank[Corpus]. *Linguistic Data Consortium.*

Carreras, X., 2005. *Learning and inference in phrase recognition: a filtering-ranking architecture using perceptron.* Tesis doctoral. Polytechnic University of Catalunya.

da Cunha, I. y M. Iruskieta, 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies,* 12 (5): 563-598.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes y I. Castellón, 2010. Discourse segmentation for Spanish based on shallow parsing. En *Proceedings of the 9th Mexican international conference on Advances in artificial intelligence: Part I*, 13-23.

da Cunha, I., J. Torres-Moreno y G. Sierra, 2011. On the Development of the RST Spanish Treebank. En *Proceedings of the 5th Linguistic Annotation Workshop*, 1-10.

Girill, T., 1991. Information chunking as an interface design issue for full-text databases. *Interfaces for Information Retrieval and Online Systems: The State of the Art,* 149-158.

Hearst, M.A., 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics,* 23 (1): 33-64.

Iruskieta, M., A. Díaz de Ilarraza y M. Lersundi, En prensa. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. En *XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

Karlsson, F., A. Voutilainen, J. Heikkila y A. Anttila, 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text.* Walter de Gruyter, .

Kiss, T. y J. Strunk, 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics,* 32 (4): 485-525.

Lehmann, C., 1985. Towards a typology of clause linkage. En *Conference on Clause Combining*, 181-248.

Mann, W.C. y S.A. Thompson, 1987. Rhetorical Structure Theory: A Theory of Text Organization. *Text,* 8 (3): 243-281.

Marcu, D., 2000. *The theory and practice of discourse parsing and summarization.* The MIT press, Cambridge.

Pardo, T.A.S., 2006. SENTER: um segmentador sentencial automático para o português do Brasil. En: Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional:1-6.

Pardo, T.A.S. y M. Nunes, 2006. Review and Evaluation of DiZer–An Automatic Discourse Analyzer for Brazilian Portuguese. En *International Workshop on Computational Procesing of Written and Spoken Portuguese*, 180-189.

Pardo, T.A.S. y M.G.V. Nunes, 2008. On the development and evaluation of a Brazilian Portuguese discourse parser. *Journal of Theoretical and Applied Computing,* 15 (2): 43-64.

Pardo, T.A.S.y M.G.V. Nunes, 2002. Segmentação Textual Automática: Uma Revisão Bibliográfica. En: Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional.

Stede, M., 2004. The Potsdam commentary corpus. En *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 96-102.

Taboada, M. J. Renkema, 2011. Discourse Relations Reference Corpus.

Tofiloski, M., J. Brooke y M. Taboada, 2009. A syntactic and lexical-based discourse segmenter. En *Proceedings of the ACL-IJCNLP 2009*, 77-80.

Webber, B., M. Stone, A. Joshi y A. Knott, 2003. Anaphora and discourse structure. *Computational Linguistics,* 29 (4): 545-587.

---

[i] Página Web de la RST: http://www.sfu.ca/rst/

[ii] Más información sobre segmentadores en Pardo y Nunes (2002).

[iii] Utilizamos N-N (Núcleo-Núcleo) para señalar las relaciones paratácticas o relaciones multinucleares con más de un núcleo, mientras que utilizamos N-S (Núcleo-Satélite) para señalar las relaciones hipotácticas o relaciones nucleares con solo un núcleo, pudiendo ser su orden Núcleo-Satélite o Satélite-Núcleo.

[iv] La fuente de los ejemplos se indica primero por el acrónimo, seguido del año de publicación y un número que distingue los números publicados en un mismo año. Los artículos se han extraído de la página Web de la revista Gaceta Médica de Bilbao: http://www.elsevier.es/en/revistas/gaceta-medica-bilbao-316.

[v] El corpus anotado en diferentes niveles puede ser consultado en la página del grupo IXA dentro de la sección de recursos: https://ixa.si.ehu.es/Ixa/resources/Euskal_RSTTreebank.

[vi] En algunos casos algunos signos de puntuación (punto y dos puntos) pueden crear un segmento de discurso a pesar de no poseer un verbo conjugado.

[vii] MORPHEUS puede ser probado en: http://ixa2.si.ehu.es/demo/analisianali.jsp.

[viii] EUSTAGER puede probarse en: http://ixa2.si.ehu.es/demo/analisimorf.jsp.

[ix] EIHERA puede probarse en: http://ixa2.si.ehu.es/demo/entitateak.jsp.

[x] Utilizamos el carácter '[' para señalar el inicio de la segmentación manual y el ']' para el final. Y para la segmentación automática los caracteres '<' de inicio y '>' de final de segmento. Los caracteres de '{' y de '}' se utilizan para representar el inicio y final de la construcción SAME-UNIT en la segmentación manual.

[xi] Aunque esta regla es suficiente para la segmentación de A1, para la segmentación de A3 es necesario detectar la función sintáctica de *zerbitzu hauetako medikuek* 'el personal sanitario de estos servicios' que es el sujeto del verbo *jotzen dituzte* 'son consideradas' y, por tanto, parte del segmento.

[xii] Ofrecemos la primera acepción del diccionario OEH (http://www.euskaltzaindia.net/oeh): Ocuparse, estar en actividad; actuar, comportarse; hablar, tratar (sobre).

[xiii] En este caso la oración principal y la subordinada han sido detectadas y formalizadas correctamente tal y como se diseñaron para la segmentación automática, la cláusula adverbial subordinada dentro de la oración principal. Esa formalización es adecuada para las construcciones SAME-UNIT, cuando una EDU divide la otra EDU. Pero en este ejemplo no estamos ante tal construcción y pensamos que la formalización no coincide con la segmentación discursiva, ya que ambos segmentos se consideran EDUs y no hay un segmento que divida otro. Por lo tanto, de dicha formalización surgen dos diferencias: i) el primer segmento automático se considera sobre-segmentado, una composición de EDUs y ii) la marca de inicio del segundo segmento automático no coincide con el manual, se considera que hay una unidad que falta segmentar, en este caso una oración principal.

[xiv] La traducción de los ejemplos que se ofrece han sido extraídos del mismo artículo original, en los casos en los que la traducción se alejaba de la versión en euskera se ha modificado mínimamente acercándonos lo máximo posible a la explicación del fenómeno.

5

Detecting the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations

Detecting 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 identify superficial markers which help to determine the central unit of rhetorical structure trees. 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). This study analyzes how agreement regarding the central unit influences agreement when establishing rhetorical relations. These results help to establish criteria to be used in RST-based annotation 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 an automatic system for identifying the central unit in rhetorical structures.

1 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). Furthermore, 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 clearly establish rhetorical relations (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). If texts are taken from parallel corpora, it is possible that some discrepancies which arise when assigning coherence relations may be the result of different translation strategies (da Cunha and Iruskieta, 2010).

In searching for linguistic evidence which can be used 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 study focuses on how to identify the unit associated with the main node in the rhetorical structure tree or, in other words, the "central unit" (CU) (Stede, 2008), e.g. 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 the-

oretical and methodological implications.

This paper aims to answer the following research questions:

(i) Does agreement about the CU 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 CU of a rhetorical structure? If there are multiple CUs, have they similar marks, that is, do they contain verbs from the same semantic class?

In order to answer these questions, Section 2 of this paper describes the theoretical framework, corpus and methodology utilized in this study. Section 3 lays out the results obtained and the final section presents conclusions and suggests directions for future research.

2 Theory, corpus and methodology

2.1 Theory

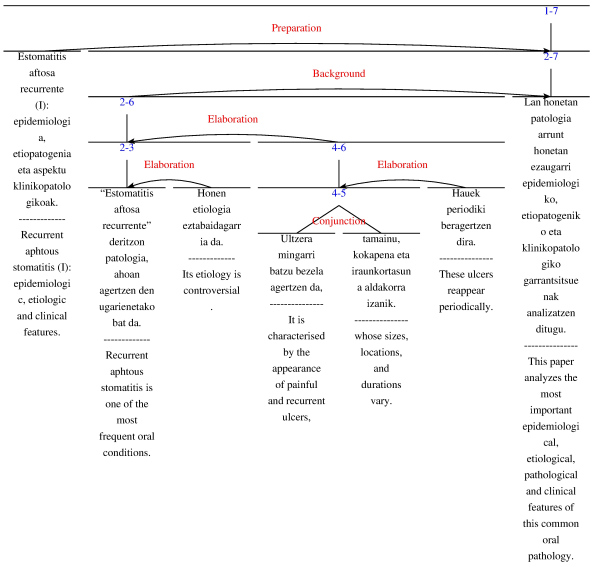
Various theories describe the relational structure of a text (Asher and Lascarides, 2003; Grosz and Sidner, 1986; Mann and Thompson, 1987). This study is based on Mann and Thompson's (1987) Rhetorical Structure Theory (RST), an applied, language-independent theory that describes coherence between text 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 example, in paratactic relations –that is, relations between two fragments which are equally important– neither unit can be eliminated without affecting the coherence or contents of the text. 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 (2011).

According to RST, hypotactic and paratactic relations connect elementary discourse units (EDUs) either a nucleus or groups of nuclear 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 subject 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 (however many nuclei are in the relation).

For example,[1] 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 is composed by the relation PREPARATION. 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). Consequently, it is the most important unit in the structure, which is indicated by the verb *analizatu* 'analyze'.

Figure 1: A rhetorical structure tree for text GMB0301 (A1)



Determining nuclearity in a relation –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

[1] Examples are extracted from the Basque corpus used in this study.

rhetorical structure is represented in RST based on several reasons:

i) It is not clear what grounds are used to make the decision (e.g. if it is made because of nuclearity or if it is because the effect of a rhetorical relation).[2]

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

In this study, the authors also identified the same problems. Examples (1) and (2) demonstrate how different choices of nuclearity affect agreement in rhetorical relations.

(1) [*Emaitza:*]$_1$ [*Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (% 51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (% 50,5).*]$_2$ GMB0401
[Results:]$_1$ [The average user is as follows: male (51.4%), middle-aged (43.2 years old), and treated for trauma (50.5%).]$_2$

A1 decides that the second unit in Example (1) is more important than the first unit. The other annotator (A2), however, makes the exact opposite decision. Both annotators arrive at their conclusions based on structural reasons. Disagreements about the importance of each text fragment impact 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*]$_1$ [*eta kontsulta hauen % 70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*]$_2$ GMB0401
[It is calculated that about 80% of users come to emergency services on their own initiative]$_1$ [and that 70% of visits are considered minor by health care personnel.]$_2$

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. A2, on the other hand, annotates the same discourse segment as a paratactic relation, 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 (note that A1 annotates the relation with a hypotactic label, ELABORATION, while A2 uses the paratactic label CONJUNCTION).

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 less agreement. This hypothesis underpins the methodology used to answer the second research question of this study.

Next subsection describes the corpus utilized for this study.

2.2 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 is composed of three sub-corpora from the same genre (abstracts) but from different domains. The communicative goal of these texts is to present specialized knowledge, since both the writer and readers are experts (Cabré, 1998), The corpus is trilingual, containing texts written in Basque, English and Spanish. It includes relevant texts in three specialized domains: medicine, terminology and science. Medical texts include the abstracts of all medical articles written in Basque in the *Gaceta Médica de Bilbao* (GMB) 'Medical Journal of Bilbao' be-

[2]Recall that in RST a rhetorical relation consists of constraints on the nucleus, constraints on the satellite, constraints on the combination of nucleus and satellite, and the effect. Thus, at times it is unclear whether annotators make their decisions based on nuclearity or the effect of the relation.

tween 2000 and 2008. Texts related to terminology are abstracts from the proceedings of the *Congreso Internacional de Terminología* (TERM) 'International Conference on Terminology' put on 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 Investigacin de la Facultad de Ciencia y Tecnología* (ZTF) 'Research Conference of the Faculty of Science and Technology', which took place in 2008.

Table 1 provides an overview of the size of the corpus.

	GMB	**TERM**	**ZTF**	**Total**
Texts	20	20	20	60
Sentences	198	253	352	803
Discourse Units	273	527	555	1355
Word count	3024	4416	6693	14133

Table 1: Breakdown of the corpus

After the annotation process (central unit and rhetorical relations among other issues), the annotated corpus was evaluated and harmonized by a judge. The harmonized corpus can be consulted in the RST Basque TreeBank[3] (Iruskieta et al., 2013a).

2.3 Methodology

The methodology presented below was designed in order to define the linguistic forms mainly verbs (which would allow identifying the CU).

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 CU and the annotation of the rhetorical structure lead us to redefine the annotation strategy for TERM and ZTF in the sense that we asked the annotators that before tagging the rhetorical structure they have to identify first the CU (one or more).

So 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[4] (O'Donnell, 1997) program.

$(B.)$ CU identification (TERM and ZTF). Both annotators determined the CU[5] and which verbs were present in the CU of a scientific abstract in TERM and ZTF domains.

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.

$(D.)$ CU identification (GMB). Both annotators extracted the CU from the rhetorical tree structures and verbs present in the CU of a scientific abstract in GMB domain.[6]

E. Evaluation. Agreement in rhetorical tree structures were manually evaluated following the qualitative methodology proposed in da Cunha and Iruskieta (2010).

F. Interpretation. We compared the results when there was agreement in CU and where there was not to check if there was any correlation using a t-test formula at 99.5% confidence.

G. SUMO class annotation. Determine the semantic classes of the Basque verbs extracted in the previous phase. Given the lack of sufficiently robust materials on the semantic classes of Basque verbs, the verbs were translated from Basque into English and their semantic classes were determined using the SUMO (Niles, 2003) ontology category. This process of searching for equivalences was conducted using free tools readily available for the English language. It consisted of the following steps:

i. Determining the exact translation of

[3]The RST Basque TreeBank is available at `http://ixa2.si.ehu.es/diskurtsoa/en/fitxategiak.php`.

[4]The website for the rhetorical structure tree graphic editing tool is `http://www.wagsoft.com/RSTTool/`.

[5]We want to illustrate the complexity of the CU selection reporting the average number of EDUs: the average number (which was calculated based on the number of EDUs, over the number of texts) of 22.58 EDUs for CU candidates per text illustrates the complexity of the CU selection task. In order to do so, the authors compared the main verbs and indicators indicating the CU and the relations affiliated with it in the corresponding rhetorical structure trees. These results indicate the correlation between the degree of inter-annotator agreement in determining the CU and the degree of agreement in assigning the RR attached to CU or for the full text structure.

[6]The central units (CU) can be consulted also in RST Basque TreeBank.

each Basque verb[7] as per the Unified Verb Index.

ii. Searching for the Basque equivalent which matches the appropriate meaning in the Multilingual Central Repository 3.0 (MCR) (Atserias et al., 2004), which includes information about various ontologies and knowledge bases.

iii. Annotating the semantic classes of each verb in the SUMO ontology (which is included in the aforementioned MCR).

H. Verbs ambiguity computation. Study ambiguity[8] in the verbs detected in the central unit. In order to determine the degree of ambiguity of the verbs used in the central unit, the three sub-corpora were automatically analyzed[9] and the corresponding morphological analysis of every word (at EDU level) was stored in the "search section" of the RST Basque TreeBank,[10] enabling researchers to search this information by lemma and category. This database provided an easy way to determine the general behavior of the verbs used in the central units, allowing the researchers to consider whether these verbs are highly frequent and how often they appear in the central unit.

3 Results and discussion

Our main hypothesis is that an agreement on CU leads us to a more 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 bigger.

3.1 Benefits from indicating the CU before rhetorical structure annotation

The result of all sub-corpora considered are presented in Table 2.

[7]Verb translations were extracted from parallel versions of the same corpus when both forms corresponded to the same synset in the MCR. If not, the Elhuyar Basque-English dictionary was utilized.

[8]In this study, ambiguity refers to the fact that a verb can have two purposes: it can indicate a nucleus unit of a rhetorical relation or the central unit of tree structure.

[9]Morphosyntactic analysis: MORPHEUS (Aduriz et al., 1998); lemmatization and identification of syntactic functions: EUSTAGGER (Aduriz et al., 2003); identification of multiword units and name entities: EIHERA (Alegria et al., 2003).

[10]The "search section" of the RST Basque TreeBank is available at `http://ixa2.si.ehu.es/diskurtsoa/en/bilaketak.php`.

GMB		TERM		ZTF	
Match	**F_1**	**Match**	**F_1**	**Match**	**F_1**
13 of 20	0.65	16 of 20	0.8	16 of 20	0.8

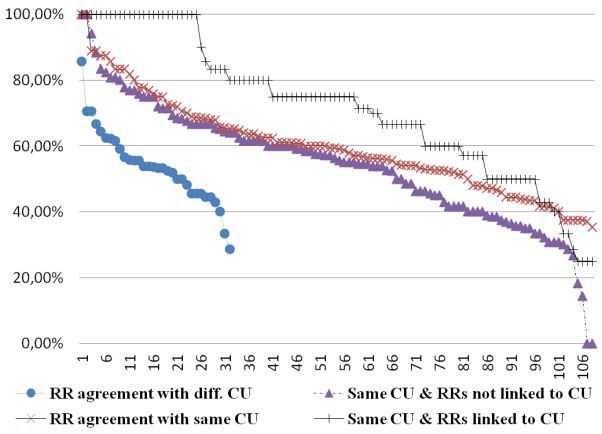
Table 2: Global results for agreement regarding the central unit

As Table 2 indicates, the change in methodology improved CU agreement between annotators. This underscores the benefits of a first step which entails detecting the CU. But this does not answer our first question: does agreement about the CU affect inter-annotator reliability when annotating rhetorical relations?

3.2 Correlation between agreement on RRs and agreement on CU

The observation we did in the GMB, first annotated GMB subcorpus, that there was more agreement in rhetorical relations when there CU was the same, was maintained after considering results of a more extended corpus with two new subcorpus (TERM and ZTF) and two more annotators.

Figure 2: Representing agreement about the CU and RRs



We can see in Figure 2 that there are differences in both populations (average agreement about rhetorical relations which is represented in Figure 2 with red crosses when the CU was the same and with blue circles when the CU was different).

The results confirm this fact although the figures have been substantially reduced when more data (all the corpus) were considered, from a difference of 0.1497 to a difference of 0.0426. Table 3 presents the global results of the comparison be-

tween the CU and rhetorical relations for the corpus as a whole.

GMB			Corpus		
= CU	≠ CU	Diff.	= CU	≠ CU	Diff.
0.7456	0.5959	**0.1497**	0.5915	0.5489	**0.0426**

Table 3: Table 3: Agreement about the CU and RRs

Even so, we wanted to see whether this small difference was significant or not. To check it, we look 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: there is observed a significant agreement when we compare agreement in RR when there was match in CU than where there was not (in RR linked to the CU). But it is very blurry to say which RR are linked to CU when annotators did not consider the same CU. Although, this answers positively and partially to the first research question: Does agreement about the central unit affect inter-annotator reliability when annotating rhetorical relations?

3.3 Correlation between agreement on RRs linked or not to CU

After our main hypothesis was confirmed, we go ahead in the tree structure and we check another hypothesis if there is more agreement in rhetorical relations linked to the CU (considering the structures when there was agreement in CU), than in the other relations of tree structure. For example, in the rhetorical structure tree presented in Figure 1, we consider two relations linked to CU PREPARATION (1>2-7)[11] and BACKGROUND (2-6>7), while the other four relations are not linked to CU (ELABORATION (3<2), ELABORATION (4-6<2-3), ELABORATION (6<4-5) and CONJUNCTION (4=5). Table 4 presents the results of relations linked to CU with relation not linked to CU:

[11] This hypotactic relation can be stated as 1 > 2-7. The unit represented by span 1 is the satellite of the hypotactic relation whose nucleus is represented by span 2-7. The symbol '>' represents the direction of the relation from the satellite toward the nucleus and the symbol '=' represents the connection in paratactic or multi-nuclear relations.

GMB			Corpus		
Linked	Not	Diff.	Linked	Not	Diff.
0.7454	0.5881	**0.1573**	0.7179	0.5449	**0.1730**

Table 4: Comparison between RRs linked and no-linked to CU in structures with the same CU

In structures with the same CU we made a comparison between the agreement in rhetorical relations linked to the CU and all the other relations. Percent agreement is substantially higher when we observe the relations linked to the CU: 17.3% higher than agreement in the relations there were not linked to the CU. This populations representing average agreement about rhetorical relations are represented in Figure 2 with black crosses when the RRs are linked to CU and with violet triangles when RRs are not linked to CU.

Populations being compared (average agreement about rhetorical relations linked to CU in texts when the CU was the same and average agreement about rhetorical relations not linked to CU in texts when CU was the same) follows a normal distribution (p-value of K-S test was 0.93) but have not the same variance (p-value of F-test is 1.296). The result of the null hypothesis (the average RR agree on a text according to the central unit is no different in the average percentage of agreement in the rhetorical relations linked to the UC to those not linked) could not be confirmed (p-value of t-test was smaller than 0.001), so we can establish a correlation.

These results help to answer the second research question of this study and seem to indicate that there is a correlation between these two kinds of agreement: greater agreement about detecting the central unit correlates with greater agreement in the annotation of rhetorical relations, also with those which are linked to the CU.

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.

ii) In Computational Linguistics, a process which helps to automatically identify the CU is important for determining rhetorical structure.

In order to discuss this results, first of all we have to consider that the CU is a nuclear unit

that relations are linked at various levels (intra-sentential level and inter-sentential level); but especially there are more relations linked from inter-sentential level. For example, in Figure 1 two relations linked to CU are only from inter-sentential level. Because of that reason, we think that these results (RRs linked to CU) are not so trivial, since the degree of agreement expected at higher level tree structures are lower. In other words, the agreement at lower levels is higher than in the high level. For example, Marcu and Echihabi (2002) argue that automatic annotation of certain rhetorical relations should address it initially in intra-sentential level as the less ambiguous. In line Soricut and Marcu (2003) mention that some of the rhetorical relations are derived from syntactic structures.

In the framework of the evaluation of a discourse parser based automatic linguistic patterns for Brazilian Portuguese, Pardo and Nunes (2008) have obtained more agreement about the intra-sentential level when annotating rhetorical relations. At that level, but for English, Soricut and Marcu (2003) also have achieved more agreement in rhetorical relations, obtaining for a statistical model a similar degree of agreement achieved by human annotators. However, according to Pardo and Nunes (2008) a statistical model of annotation that cannot be extended to inter-sentence level, with the same results. In this regard, the results of Iruskieta et al. (2011b) confirm the aforementioned works (Pardo and Nunes, 2008; Soricut and Marcu, 2003; Marcu and Echihabi, 2002) with higher agreement at this level, intra-sentential, 11.50% than from inter-sentential level in the GMB corpus.

3.4 Identifying the semantic class of verbs in the CU

After comparing the CU from both annotators, results help to answer an aspect of the second research question: Is it plausible to have multiple CUs in a text? There are multiple EDU functioning as the CU of the text was selected by A1 or by A2 in the three subcorpora: 9 multiple EDU functioning as CU in GMB, 2 multiple EDUs in TERM and 3 multiple EDUs in ZTF.

In order to describe how the CU is indicated in each domain in greater detail, the meanings of verbs were analyzed and their semantic class determined as per the SUMO ontology (cf. section 3.4). After consulting the Unified Verb Index, the relation between meaning and semantic class was obtained by means of the MCR semantic database, which includes various lexico-semantic and ontological databases. Data from the GMB, TERM, and ZTF sub-corpora are grouped in Table 5 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$_1$, show$_2$, base$_1$	0.4615	0.2273	0.0870
	Comparing	value$_2$, compare$_1$	0.2692		
	Classifying	classify$_1$			0.0870
	Learning	review$_1$	0.0385		
	Guiding	take$_3$		0.0455	
	Process	gain$_4$			0.1739
IP-IPP		recognize$_2$, determine$_8$, hold$_6$, focus$_1$	0.0385	0.0909	0.0435
IP-SI	Communication	present$_2$, addres$_9$, recount$_1$, propose$_1$	0.0385	0.4545	0.0435
IP		perform$_1$, target$_1$, set-up$_{15}$, work$_1$, make$_3$, use$_1$	0.1154	0.0909	0.0870
IP	Searching-Investigating	investigate$_1$			0.0435
IP	Organizational Process	serve$_2$			0.0435
IC		palliate$_2$		0.0455	
Predicate		be$_1$, develop$_5$, constitute$_1$, hold$_4$	0.0385	0.0455	0.3913

Table 5: 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 sub-corpus, the central unit was usually marked with verbs from the IPP category. On the other hand, in the TERM sub-corpus, verbs from the IPP and SI category. Verbs in the central unit of the ZTF sub-corpus are marked with IPP and Predicate class.

Therefore, the results demonstrate that:

i) A study is needed to know which is the SUMO class of the verbs to mark a specific domain, for example in our corpus the central unit is indicated with verbs from IPP class for three domains, but other class also has to be considered SI for TERM and Predicate for ZTF.

ii) In the case of weak verbs, other indicators help to indicate the central unit, as is the case in Example (3), where the central unit is indicated by the phrase *komunikazio honen gaia* "the topic of this paper".[12]

[12]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 sub-corpus derives from texts published in a periodical while the TERM and ZTF sub-corpora come from texts published in the proceedings of conferences. In other words, it could be argued that the medium influences the writing style and consequently impacts the classes of verbs used in the texts. This is in line with the main argument of this study,

(3) Komunikazio honen gaia izango da zelan jarri martxan terminoen asmatzaile automatizatu bat eta termino-banku baten zati osagai bilakatu.TERM_32
The topic of this paper is facilitating automated coinage of terminology and making this an integral part of an online term bank.

Indicator analysis throws more light on why the agreement in CU was much higher in TERM corpus and ZTF corpus. Although the TERM and ZTF corpora are bigger and more difficult to agree in CU, because there are more EDUs (see Table 1), the texts in those corpora are more marked by indicators than in GMB corpus (see Table 6). In GMB are 9 EDUs marked and in TERM are 16 EDUs marked with these phrases and in ZTF are 20 EDUs marked with indicators. Another reason is that the direct observation of the CU makes more consistent the CU selection, an evidence of that is that all the verbs in CU are from the same SUMO class in TERM and ZTF corpora by both annotators.

	GMB	**TERM**	**ZTF**
EDU	273	527	555
Agreement in CU	13	17	17
CUs indicators	9	16	20

Table 6: Indicators influence in CU selection

3.5 Degree to which verbs indicate the central unit

So far, this paper has provided a partial answer to the first research question. However to automatically detect the central unit by means of verbs it is necessary to consider this 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 synset of analyzed verbs are necessary in order to know what SUMO class these verbs belong to.[13]

iii) The central unit is not always indicated with a verb.

since different verbs are used to indicate the central unit in the TERM and ZTF sub-corpora, which share the same medium but come from different fields.

[13] 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.

The next phase of this research considered whether MCR synset or better say verb forms which appear in the central unit unequivocally mark 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. Table 7 presents the results of the frequency with which these MCR synsets indicate the central unit grouped in SUMO classes.

SUMO	**MCR Synset**	**GMB**		**TERM**		**ZTF**	
IP-IPP-Reasoning	examine$_1$	11 of 20	0.55	2 of 8	0.25	1 of 36	0.0278
	base$_1$	1 of 3	0.3333	1 of 7	0.1429		
	show$_2$			1 of 3	0.3333		
IP-IPP-Comparing	value$_2$	6 of 6	1				
	compare$_1$	1 of 1	1				
IP-IPP-Learning	review$_1$	1 of 1	1				
IP-IPP-Guiding	take$_3$			1 of 6	0.1667		
IP-IPP-Classifying	classify$_1$					2 of 5	0.4
IPP	recognize$_2$	1 of 2	0.5				
	determine$_8$			1 of 4	0.25		
IP-SI-Communication	recount$_1$			1 of 9	0.1111		
	present$_2$	1 of 2	0.5	5 of 17	0.2941		
	address$_9$			1 of 4	0.25		
	propose$_2$			1 of 4	0.25	1 of 1	1.00
IP	target$_1$	1 of 1	1				
	use$_1$			1 of 44	0.0227	1 of 56	0.0179
	perform$_1$	2 of 43	0.0465				
	make$_3$					1 of 49	0.0204
IP-Investigating	investigate$_1$	1 of 2	0.5			1 of 22	0.0454
OOTP	palliate$_2$			1 of 1	1.00		
Predicate	be	1 of 214	0.0047	14 of 326	0.0429	11 of 373	0.0295

Table 7: Frequency with which MCR synsets indicate the CU

With much more data we will confirm if the results show any tendency. But with this data we can only confirm that results from indicating the CU with different SUMO categories also repeats when considering ambiguity: in the GMB sub-corpus, the least ambiguous verbs are those in the IPP category (10 of 13, 76.92%) while in the TERM subcorpus such verbs are in the SI category (8 of 34, 23.53%) and in IPP (4 of 20, 20.00%).

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

i) Verb that indicate the CU with a high enough frequency in GMB is from IPP category *baloratu* 'value2'; other verbs can be considered but they have not enough frequency, e.g. *alderatu* 'compare$_1$', *gainbegiratu* 'review$_1$', *aztertu* and *analizatu* 'analyze$_1$', and *ezagutu* 'recognize$_2$'.

ii) While in TERM, there is a MCR synset 'present$_2$' (composed with the verbs *plazaratu*, *aurkeztu*, *aipatu*, *berri eman* and *jardun*) that has enough frequency but it does not indicate the CU with a high enough frequency.

iii) Some verbs rarely indicate the central unit: this is especially common in the ZTF sub-

corpus, which has no unambiguous verbs with enough frequency which indicate the CU.

4 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 relations linked to the central unit. And since 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 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 indicate the central unit. Verbs utilized to indicate the central units vary in different domains: in the GMB sub-corpus which was analyzed, the central unit was more frequently and less ambiguously indicated with verbs from the IPP category, while in the TERM subcategory, SI verbs were most frequent and least ambiguous.

Testing these results in a larger corpus 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 difficulties in automatically detecting the central unit based on the ambiguity of the verb which marks the central unit.

5 Acknowledgemnts

This study was carried out within the framework of the following projects: IXA Group. Natural language processing (GIU09/19) [UBC-EHU]; Berbatek: Tools and Technologies to promote the Language Industry (IE09-262) [Basque Government]; IXA Group. Consolidated research groups grant 2007-2012 (IT-344-10) [Basque Government]; OPENMT-2 Hybrid machine translation and advanced evaluation (TIN2009-14675-C03-01) [Spanish Ministry of Science and Innovation]; KNOW2: Language understanding technologies for multilingual domain-oriented information access (TIN2009-14715-C04-01) [Spanish Ministry of Science and Innovation]; IMLT: A framework for the integration of linguistic resources based on a general XML annotation model (TIN2007-63173) [Spanish Ministry of Education]; GARATERM-2: Developing an environment to extract terminology and neology based on linguistically processed corpora (US10/01).

References

[Aduriz et al.1998] Aduriz, Itziar, Eneko Agirre, Izaskun Aldezabal, Iaki Alegria, Olatz Ansa, Xabier Arregi, JoseMari Arriola, Xabier Artola, Arantza Diaz de Ilarraza, and Nerea Ezeiza. 1998. A framework for the automatic processing of basque. In *First International Conference on Language Resources and Evaluation*, Granada, Spain, May 28-30.

[Aduriz et al.2003] Aduriz, Itziar, Izaskun Aldezabal, Iaki Alegria, JoseMari Arriola, Arantza Diaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite state applications for basque. In *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*, Budapest, Hungary, 13-14 April.

[Alegria et al.2003] Alegria, Iaki, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2003. Named entity recognition and classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperacin de Informacin*, pages 1–8, Madrid.

[Asher and Lascarides2003] Asher, Nicholas and Alex Lascarides. 2003. *Logics of conversation*. Cambridge Univ Pr, Cambridge.

[Atserias et al.2004] Atserias, Jordi, Lus Villarejo, German Rigau, Eneko Agirre, John M. Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *GWC*, page 2330, Brno, Czech Republic, 20-23 January.

[Bateman and Rondhuis1997] Bateman, John A. and Klaas Jan Rondhuis. 1997. Coherence relations: Towards a general specification. *Discourse Processes*, 24(1):3–49.

[Cabré1998] Cabré, MaraTeresa. 1998. El discurs especialitzat o la variaci funcional determinada per la temtica: Noves perspectives. *Revista Internacional de Filologia*, (25):173–194.

[Carlson et al.2001] Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, page 10, Aalborg, Denmark, 1-2 September. Association for Computational Linguistics.

[da Cunha and Iruskieta2010] da Cunha, Iria and Mikel Iruskieta. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

[da Cunha2008] da Cunha, Iria. 2008. Hacia un modelo lingstico de resumen automtico de artculos mdicos en espaol. Doktore-tesia, 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 retricas 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 implementacin 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 retricas: 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.

[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 Echihabi2002] Marcu, Daniel and Abdessamad Echihabi. 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 online 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. Rst-tool: 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. Relaes retricas e seus marcadores superficiais: Anlise de um corpus de textos cientficos em portugułs 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 and Nunes2008] Pardo, Thiago A. S. and Maria G. V. Nunes. 2008. On the development and evaluation of a brazilian portuguese discourse parser. *Revista de Informtica Terica e Aplicada*, 15(2):43–64.

[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.

[Taboada and Renkema2011] Taboada, Maite and Jan Renkema. 2011. Discourse relations reference corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

[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: semntica y pragmtica del discurso*. Ctedra.

6

Discourse unit and rhetorical relations. A study about discourse units in the annotation of a corpus in Basque

Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera

Discourse unit and rhetorical relations. A study about discourse units in the annotation of a corpus in Basque

Mikel Iruskieta
IXA NLP Group
Department of Didactics of Language and Literature
University of the Basque Country
Ramón y Cajal 72
48014 Bilbao
mikel.iruskieta@ehu.es

Arantza Díaz de Ilarraza
IXA NLP Group
Department of Computer Science
University of the Basque Country
Manuel Lardizabal 1
48014 Donostia
a.diazdeilarraza@ehu.es

Mikel Lersundi
IXA NLP Group
Department of Basque Philology
University of the Basque Country
Sarriena auzoa z/g
48940 Leioa
mikel.lersundi@ehu.es

Resumen: En este artículo se describe el estudio realizado sobre las características del etiquetado de la estructura de discurso, según la Teoría de la Estructura Retórica, en los niveles inter-oracional e intra-oracional. El corpus etiquetado está compuesto por textos médicos escritos en euskera y extraídos de la Gaceta Médica de Bilbao siendo nuestro objetivo final establecer una metodología general para la anotación de corpus a nivel discursivo. En este trabajo se analizan los acuerdos y desacuerdos de la anotación realizada por dos anotadores en cada nivel. Los resultados obtenidos sugieren que la segmentación en unidades de discurso es más compleja en el nivel intra-oracional mientras que la asignación de relaciones retóricas lo es en el nivel inter-oracional. Además hemos detectado que hay relaciones que aparecen con mayor frecuencia en cada nivel y otras se dan indistintamente en ambos niveles inter- e intra-oracional. Este estudio sienta las bases para el futuro desarrollo de un anotador automático de relaciones.
Palabras clave: anotación, análisis del discurso, segmentación, relaciones retóricas.

Abstract: This article describes the study on the features used for labelling the discourse structure, according to the Rhetorical Structure Theory, at the inter-sentential and intra-sentential levels. The tagged corpus is composed of medical texts written in Basque and extracted from the medical journal 'Gaceta Médica de Bilbao'. The difficulties encountered both while identifying the discourse units and while establishing the relations are analysed at each level based on the observation of agreement and disagreement identified in the texts annotated by two annotators. The results obtained suggest that the segmentation into units of discourse is more complex at the intra-sentential level while the assignment of rhetorical relations is more difficult at the inter-sentential level. We also note that some relations occur more frequently at the intra-sentential level and others at the inter-sentential level. However, there are relations that can appear indistinctively in both levels intra- and inter-sentential. This study will lay the foundations to carry out the automatic annotation process that the authors intend to perform shortly.
Keywords: Annotation, Discourse Analysis, Segmentation, Rhetorical Relations.

1 Introducción

El desarrollo de aplicaciones avanzadas basadas en el procesamiento del lenguaje, tales como búsqueda y extracción de información basada en conocimiento semántico, elaboración automática de resúmenes o traducción automática, precisan de corpus de referencia etiquetados a diferentes niveles lingüísticos: morfológico, sintáctico, semántico, etc. En este artículo trataremos del etiquetado de corpus a nivel discursivo.

La segmentación discursiva del corpus, al ser el primer estadio de la anotación de la estructura discursiva, tiene una importancia crucial y ha sido analizada desde diferentes puntos de vista y con finalidades diversas.

Existe una gran controversia sobre cuáles son los criterios de segmentación más adecuados para establecer las unidades de discurso. Cuando se trata de realizar la anotación discursiva de un corpus, normalmente se opta por realizar la segmentación considerando un alto nivel de granularidad estableciendo unidades de discurso a nivel intra-oracional (Carlson, Okurowski y Marcu 2002). El nivel inter-oracional (de menor granularidad) comprende unidades entre conjuntos de oraciones (párrafos, enunciados), unidades relacionadas mediante conjunciones coordinativas y unidades relacionadas de modo adverbial[1] (oraciones compuestas); en el nivel intra-oracional (de mayor granularidad) se consideran las unidades relacionadas con conjunciones subordinantes[2] y coordinativas (cláusulas con relaciones adverbiales); finalmente, el nivel de complementos verbales con sólo relaciones sintácticas hace referencia a los complementos de verbos declarativos, verbos que tienen como complementos otros verbos.

Cuando el objetivo es ofrecer un corpus de referencia enriquecido con información discursiva a la comunidad científica se suele optar por un alto nivel de granularidad, sin embargo Tofiloski, Brooke y Taboada (2009) subrayan que una granularidad tan fina, sobre los complementos de los verbos declarativos (*attributive and cognitive verbs*) no recoge información sobre relaciones retóricas, sino que recoge información de otras cuestiones del discurso. Limitándose a las relaciones retóricas Tofiloski, Brooke y Taboada (2009), descartan el último nivel (cláusulas con sólo relaciones sintácticas) y consideran únicamente niveles inter-oracional e intra-oracional. Esta distinción es útil, por ejemplo, para la clasificación de textos de diferentes géneros (Webber 2009); sin embargo, no lo es para tareas de resumen automático (Marcu 1999), donde es más conveniente considerar la granularidad inter-oracional.

En cuanto a la segmentación de alto nivel, Girill (1991) propone unidades discursivas más amplias (el pasaje) para tareas de recuperación de la información; en este sentido Hearst (1997) determina los multiparágrafos como unidades discursivas en la detección de cambios de tema.

Por tanto, del estudio bibliográfico se observa que la granularidad puede ser determinante para el éxito o no en ciertas tareas de etiquetado.

En este sentido, nuestro objetivo general es doble: i) establecer la metodología de anotación de la estructura relacional del discurso (anotación de segmentos y relaciones retóricas) y ii) llevar a cabo el proceso de anotación inter- e intra-oracional en un corpus.

De las diferentes teorías discursivas que formalizan la estructura referencial (Webber, et al 2003, Asher y Lascarides 2003, Polanyi 1988, Wolf y Gibson 2004), el marco teórico sobre el que desarrollamos este estudio empírico es la Teoría de la Estructura Retórica[3] (RST) de Mann y Thomson (1987), que es válida según Taboada y Mann (2006a) para aplicaciones avanzadas.

Con el fin de establecer la metodología de anotación nos preguntamos si existe el mismo grado de ambigüedad, en cuanto a las relaciones de la RST, en los niveles inter-oracional e intra-oracional. El objetivo concreto de este estudio es determinar, con la menor ambigüedad posible, el tipo de relaciones o las relaciones fácilmente identificables en cada nivel de manera que sirva como base en la implementación de un analizador automático de discurso.

Marcu y Echihabi (2002) sostienen que la anotación automática de ciertas relaciones retóricas conviene abordarla inicialmente en el nivel intra-oracional por ser el menos ambiguo. En la misma línea Soricut y Marcu (2003: 234) mencionan que algunas de las relaciones retóricas se derivan de las estructuras sintácticas:

> *Our experiments empirically show that, at the sentence level, there is an extremely strong correlation between syntax and discourse. This is even more remarkable given that the discourse*

[1] Thompson et al. (1985) detallan una tipología y sus funciones de oraciones adverbiales a ambos niveles: intra-oracional e inter-oracional.

[2] En este trabajo utilizamos el concepto de subordinación de modo tradicional. Véase Lehmann (1985) para una clasificación exhaustiva sobre los diferentes grados de dependencia entre sintagmas relacionales y un acercamiento funcional de la combinación de cláusulas.

[3] Página Web de la RST: http://www.sfu.ca/rst/

corpus (RST-DT, 2002) was built with no syntactic theory in mind. The annotators used by Carlson et al. (2003) were not instructed to build discourse trees that were consistent with the syntax of the sentences. Yet, they built discourse structures at sentence level that are not only consistent with the syntactic structures of sentences, but also derivable from them.

Pardo y Nunes (2008) han obtenido en la anotación de relaciones retóricas un grado más alto de acuerdo a nivel intra-oracional, en la evaluación de un analizador discursivo automático basado en patrones lingüísticos para el portugués de Brasil y en ese mismo nivel Soricut y Marcu (2003) han logrado para el inglés con un modelo estadístico un grado de robustez parecido al conseguido por anotadores humanos. Sin embargo, según Pardo y Nunes (2008), ese modelo estadístico de anotación no puede extenderse al nivel inter-oracional.

La estructura de este artículo es la siguiente: en la sección 2 explicamos el marco teórico y la metodología empleada para la anotación del corpus y su evaluación. Los resultados de las anotaciones de los niveles inter- e intra-oracional y su interpretación se presentan en las secciones 3 y 4 respectivamente. Finalmente, en la sección 5, establecemos las conclusiones y el trabajo futuro.

2 Teoría y metodología

2.1 Teoría

La RST es una teoría de carácter aplicado e independiente del idioma que nos permite describir la coherencia entre fragmentos textuales combinando la idea de nuclearidad, o importancia de un fragmento del discurso, con la identificación de las relaciones retóricas que unen los fragmentos del texto. Se entiende que el autor va guiando al lector, mediante el texto, comunicándole explícitamente o implícitamente qué fragmento es más importante y su relación con los demás fragmentos. Las relaciones se definen en base a las restricciones que se establecen entre el núcleo (N) y satélite (S), y el efecto que crea en el lector. Estas relaciones, según la teoría, pueden ser paratácticas (N-N), cuando se establece la relación entre fragmentos con el mismo grado de importancia en la intención del autor (LISTA, CONTRASTE, DISYUNCIÓN, etc.), o hipotácticas (N-S), cuando se establece una relación entre una unidad menos importante: satélite (S) con otra más importante: núcleo (N) siempre según la intención del autor. (ELABORACIÓN, MÉTODO, CONCESIÓN, CAUSA, RESULTADO, etc.). Las relaciones hipotácticas se clasifican en relaciones de presentación (P) y de contenido (C)[4].

Dado que éste es el primer estudio de estas características que se realiza para el euskara, nuestro objetivo es establecer las relaciones retóricas entre los fragmentos del discurso siguiendo las definiciones RST pero sin consensos previos ante las diferentes formas lingüísticas que señalan una u otra relación. Después estudiaremos las discrepancias y estableceremos los criterios lingüísticos que nos lleven a una anotación robusta. Por este motivo hemos elegido la clasificación extendida con 29 relaciones (Mann y Taboada 2010), dejando aparte las clasificaciones más complejas como por ejemplo la propuesta por Carlson, Marcu y Okurowski (2001) de 78 relaciones. Para la visualización y etiquetado de los fragmentos y relaciones hemos utilizado la herramienta RSTTOOL (O'Donnell 2000).

2.2 Metodología

La metodología de este estudio incluye tres fases.

1. Constitución del corpus. Se constituye el corpus que contiene todos los resúmenes en euskera extraídos de la Gaceta Médica de Bilbao[5] desde sus inicios hasta el año 2008. El corpus está compuesto por 20 documentos y tiene un tamaño de 3.024 palabras.

2. Niveles de anotación retórica. En primer lugar, tras un proceso en el que se establecieron unos criterios de anotación generales, dos anotadores segmentan los textos del corpus a nivel inter-oracional que comprende oraciones con verbo conjugado y después relacionan las unidades discursivas identificadas utilizando la clasificación RST extendida. En segundo lugar, se pide a los anotadores que vuelvan a segmentar de nuevo los textos del corpus, pero con una mayor granularidad, anotación intra-oracional, que comprende oraciones adverbiales en la misma oración. Para no repetir tareas, sólo se

[4] Véase su distribución en la Tabla 4.

[5] Los artículos se han extraído de la página Web de la revista Gaceta Médica de Bilbao: http://www.gacetamedicabilbao.org/web/es/.

relacionan los segmentos (*spans*[6]) intra-oracionales entre punto y punto.

3. Evaluación del etiquetado. Se evalúan y se comparan las anotaciones y se extraen conclusiones de las tareas de segmentación y del análisis retórico realizadas en ambos niveles: inter-oracional e intra-oracional. El método que hemos utilizado para evaluar las anotaciones retóricas es el propuesto por da Cunha e Iruskieta (2010). La Figura 1 y la Figura 2 ilustran los dos niveles de segmentación. La Figura 1 muestra un ejemplo de segmentación y su anotación retórica a nivel inter-oracional, donde se considera como unidad discursiva (EDU) aquélla que contiene un verbo conjugado, a excepción del título que constituye una EDU aunque no contenga verbo. En la Figura 2 se muestra un ejemplo de la segmentación y su anotación retórica sólo a nivel intra-oracional, donde se considera una unidad como EDU siempre que presente un verbo, conjugado o sin conjugar, sea o no subordinado.

En cuanto a la fase referente al establecimiento de las relaciones retóricas se ha pedido a cada anotador que primero relacione las unidades que van de punto a punto y después los párrafos de manera incremental y modular como propone Pardo (2005).

Observamos en los árboles que representan la anotación retórica a nivel inter-oracional, que hay un desacuerdo en la relación entre los *span*s 2 y 3; el anotador A1 detecta la presencia de una relación hipotáctica de ELABORACIÓN mientras que A2 anota una relación paratáctica de UNIÓN. Este desacuerdo en la interpretación está ligado al concepto de nuclearidad y es debido a la ausencia de elementos discursivos que faciliten la identificación de la relación retórica.

En la Figura 2 se representa la segmentación a nivel intra-oracional[7] del último segmento del ejemplo de la Figura 1. Se ha considerado la conjunción de cláusulas verbales con complementos, en la que hay un verbo no conjugado *aztertu* 'examinar' y otro conjugado *alderatu da* 'se ha comparado'. En este caso se establece la relación de SECUENCIA entre ambas cláusulas que se explicita mediante los verbos *aztertu* 'examinar' y *alderatu* 'comparar' y la conjunción *eta* 'y'.

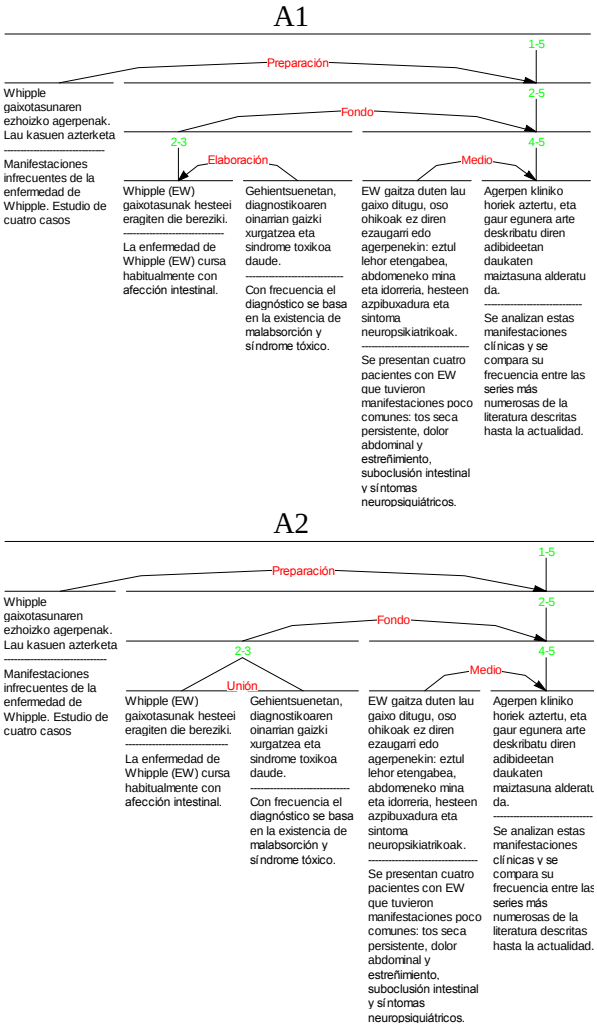


Figura 1: Anotación retórica inter-oracional

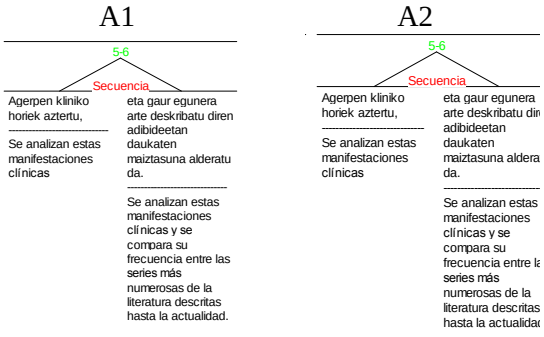


Figura 2: Anotación retórica intra-oracional

[6] El término *span* comprende tanto unidades de discurso como conjuntos de unidades.

[7] El fragmento de la Figura 2 es el correspondiente a la EDU 5 de la Figura 1. No hemos puesto las demás EDUs en la figura porque no reciben relación retórica alguna.

3 Resultados y discusión

En esta sección comparamos los resultados de las diferentes fases de cada tarea: segmentación y análisis retórico.

3.1 Segmentación

En cuanto a la evaluación de la segmentación se utilizan diferentes tipos de medidas: a) acuerdo promedio (*percent agreement*) (Marcu 1999, Hearst 1997, Passonneau y Litman 1993), que se utiliza para medir los posibles acuerdos entre anotadores; b) precisión y cobertura. Passoneau et al. (1993) lo utilizan para evaluar la fiabilidad del algoritmo de segmentación. Afantenos et al. (2010) utilizan *F-score*, medida utilizada en las tres tareas en CoNLL 2001, que combina ambas medidas (precisión y cobertura) para la anotación por pares de humanos y c) coeficiente Kappa. Esta otra medida que substrae el valor de la casualidad (Carletta 1996) es usada para medir el acuerdo entre anotadores en Hearst (1997) y Miltsakaki et al. (2004) y Tofiloski, Brooke y Taboada (2009), estos últimos comparan esta medida con *F-score*.

	EDU
Inter	100,00%
Intra	86,26%

Tabla 1: Acuerdo en la segmentación

En la Tabla 1 presentamos los resultados de la segmentación, donde podemos observar que el acuerdo nivel inter-oracional ha sido mayor que a nivel intra-oracional; ya que es mayor la complejidad de identificar las unidades a nivel intra-oracional por la variedad de casos que se presentan, sobre todo en una lengua aglutinante como el euskera.

3.2 Evaluación cualitativa del acuerdo en la anotación retórica

Con el método cuantitativo propuesto por Carlson et al. (2001) se mide el acuerdo en las anotaciones, dando especial importancia a la nuclearidad, donde se evalúa el acuerdo en los siguientes factores: i) segmentos simples del discurso (EDU), ii) segmentos compuestos, iii) nuclearidad y iv) relación. Aunque la nuclearidad es un factor de interés para muchas aplicaciones –por ejemplo, resumen automático (Marcu 1999) y detección del antecedente anafórico (Danlos 2008, Cristea, Ide y Romary 1998)–, no lo es para el objetivo concreto de este estudio. Nuestra propuesta es realizar una evaluación más cualitativa basada en los siguientes factores que intervienen en la asignación de la relación retórica: i) identificación de unidades núcleo a las que se asocia las relaciones o Asociación (A), ii) identificación de EDU o conjunto de *span*s de las unidades satélite (Composición: C) y iii) relaciones[8] (R). En la Tabla 2 se presenta el acuerdo en ambos niveles[9] atendiendo a los factores mencionados.

	A	**C**	**R**
Inter	71,23%	67,45%	57,00%
Intra	88,37%	92,06%	71,19%

Tabla 2. Cobertura en Asociación, Composición y Relación

Los resultados sugieren que la dificultad de la tarea a nivel inter-oracional es mayor, ya que a ese nivel hay menor acuerdo en todos los factores. Si tenemos en cuenta la cobertura es un 17,14% menor en la Asociación (A), un 24,61% en la Composición (C) y un 14,35% en la Relación (R). De estos datos deducimos que aunque la tarea de la segmentación es más compleja, el resto de las tareas a nivel intra-oracional son más simples. Las razones podrían ser las siguientes:

- La manera en que se combinan los segmentos es más simple a niveles más bajos (Composición).
- La identificación de las unidades núcleo a las que se asocian las relaciones es más sencilla a nivel intra-oracional que a nivel inter-oracional (Asociación).
- Tal y como apuntaban Marcu y Echihabi (2002) y Soricut y Marcu (2003), existe una fuerte relación entre sintaxis y discurso por lo tanto es más sencillo establecer la relación entre las unidades a nivel intra-relacional (Relación).

Nos fijamos ahora en los casos de acuerdo a nivel de relación y observamos más en detalle (Tabla 3) en qué casos se ha producido acuerdo total: i) acuerdo en Composición, Asociación y Relación (CAR); acuerdos parciales: ii) en Asociación y Relación (AR), iii) acuerdo en

[8] En da Cunha e Iruskieta (2010) se propone el modo de evaluar también la nuclearidad con este método cualitativo.

[9] La precisión es la misma para los tres factores. A nivel inter-oracional es de 100,00% y a nivel intra-oracional de 96,39%.

Composición y Relación (CR) y iv) acuerdo únicamente en Relación (R).

	CAR	AR	CR	R
Inter	83,60%	5,74%	4,10%	6,56%
Intra	93,10%	5,17%	1,72%	0,00%

Tabla 3. Tipos de acuerdo en base a la relación

Los resultados de la Tabla 3 sugieren que el acuerdo a nivel intra-oracional además de ser mayor es más consistente, ya que el acuerdo se basa en menor medida en acuerdos parciales (AR, CR y R).

3.3 Descripción de las relaciones retóricas

En este apartado se describe la frecuencia y ambigüedad de las relaciones hipotácticas en los niveles intra-oracional e inter-oracional.

R	Inter	Intra
Lista (N-N)	26,02%	15,52%
Elaboración (C)	21,95%	8,62%
Preparación (P)	17,07%	0,00%
Método (C)	10,57%	10,34%
Resultado (C)	8,94%	8,62%
Fondo (P)	8,13%	3,45%
Circunstancia (C)	0,00%	15,52%
Conjunción (N-N)	0,00%	8,62%
Condición (C)	0,00%	5,17%
Propósito (C)	0,00%	5,17%
Interpretación (C)	3,25%	3,45%
Concesión (P)	0,81%	3,45%
Evidencia (P)	0,81%	3,45%
Causa (C)	0,00%	3,45%
Contraste (N-N)	1,63%	1,72%
Justificación (P)	0,81%	0,00%
Motivación (P)	0,00%	1,72%
Secuencia (N-N)	0,00%	1,72%
Total	**100,00%**	**100,00%**

Tabla 4. Acuerdo en relaciones

En la Tabla 4 se presenta relación por relación[10] el acuerdo habido entre ambos anotadores en los diferentes niveles.

Considerando sólo los casos con un acuerdo superior al 5,00%, observamos que: i) a nivel intra-oracional, entre las relaciones hipotácticas, las más utilizadas y con mayor acuerdo, es decir, las menos ambigüas, son las relaciones de contenido (C) con formas subordinadas: CIRCUNSTANCIA, CONDICIÓN y PROPÓSITO; ii) a nivel inter-oracional, las relaciones de presentación (P): PREPARACIÓN y FONDO y iii) algunas relaciones de contenido (ELABORACIÓN, MÉTODO y RESULTADO) se utilizan en ambos niveles con frecuencia similar.

3.4 Descripción de la discrepancia

En relación con el desacuerdo de anotación las causas más discutidas son: a) la indeterminación de las relaciones retóricas por definición (Stede 2008), b) las diferentes y posibles interpretaciones (Taboada y Mann 2006b)[11] y c) falta de consenso previo (ver Tabla 5). Tras el estudio de las discrepancias se han detectado lo que podemos llamar patrones de confusión, que en nuestro caso se deben a los siguientes factores: a) segmentación, evidencia la dificultad de la segmentación a nivel intra-oracional; b) determinación de la nuclearidad; c) asignación de relaciones paratácticas, y d) asignación de relaciones hipotácticas.

Patrones de confusión	Inter	Intra
Segmentación	0,00%	27,00%
Nuclear (N-S)	54,00%	43,00%
Multinuclear (N-N)	13,00%	11,00%
Nuclear vs Multinuclear (N-S/N-N)	33,00%	19,00%

Tabla 5. Patrones de confusión de relaciones

Confusión en relaciones nucleares	Inter	Intra
Interpretación (C) / Resultado (C)	10,00%	0,00%
Justificación (P) / Causa (C)	1,00%	12,00%
Otras confusiones	43,00%	31,00%

Tabla 6. Patrones de confusión de relaciones

[10] En cada relación se especifica el tipo de relación. Según la RST hay dos tipos de relaciones hipotácticas: relaciones de presentación (P) y relaciones de contenido (C). Las otras relaciones son paratácticas o multinucleares (N-N).

[11] Aunque un texto puede tener más de una interpretación o árbol (Mann y Thompson 1987), se le ha pedido a los anotadores que den únicamente una interpretación.

Dentro de los patrones de confusión en relaciones hipotácticas es notable señalar que también los patrones de confusión (Tabla 6) son sensibles a nivel: INTERPRETACIÓN/RESULTADO a nivel inter-oracional y JUSTIFICACIÓN/CAUSA a nivel intra-oracional.

4 Conclusiones y trabajo futuro

Hemos presentado el estudio realizado sobre las características del etiquetado de la estructura de discurso, según la Teoría de la Estructura Retórica, en los niveles inter-oracional e intra-oracional. Este estudio nos sirve de base para establecer y refinar la metodología de etiquetado de estructuras de discurso. Basándonos en los resultados de este estudio indicamos que el acuerdo, en niveles más bajos del árbol retórico (nivel intra-oracional), es menor en la segmentación, un 13,74% menor; pero es mayor en la asignación de relaciones retoricas por su alto grado de señalización, un 14,35% mayor. Además, los resultados señalan que la configuración de las relaciones es diferente en un nivel u otro. Las relaciones hipotácticas en el nivel inter-oracional de mayor frecuencia y acuerdo son PREPARACIÓN y FONDO; mientras que en el nivel intra-oracional de las relaciones hipotácticas son: CIRCUNSTANCIA, CONDICIÓN y PROPÓSITO. Las relaciones con mayor acuerdo a nivel inter-oracional podrían ser explotadas en tareas de resumen automático y las del nivel intra-oracional en tareas de extracción de información. A pesar de los desacuerdos encontrados los resultados sugieren que la anotación automática de discurso debería considerar las tres relaciones intra-oracionales mencionadas por las siguientes razones: i) están siempre señalizadas y ii) ofrecen un bajo grado de ambigüedad. La identificación de estas relaciones nos puede servir de ayuda en el diseño de un anotador automático de relaciones retóricas. Además, en los patrones de confusión también identificamos claves importantes para dicho diseño en lo referente a las relaciones retóricas INTERPRETACIÓN/RESULTADO a nivel inter-oracional y JUSTIFICACIÓN/CAUSA a nivel intra-oracional.

En trabajos futuros analizaremos las razones lingüísticas de la correlación entre sintaxis y discurso en la anotación automática de relaciones retóricas y abordaremos las razones de los patrones de confusión para realizar árboles de decisiones o manual detallado de las marcas que evidencian las relaciones.

Agradecimientos

Este trabajo ha sido realizado en el marco de los siguientes proyectos: Grupo IXA, Grupo consolidado 2007-2012 (IT-397-07) [Gobierno Vasco]; KNOW2 (TIN2009-14715-C04-01) [MICCIN], Hibrido Sint (TIN2010-20218) [MICCIN], y GARATERM2 (US10/01) [Gobierno Vasco].

Bibliografía

Afantenos, S., P. Denis, P. Muller y L. Danlos, 2010. Learning Recursive Segments for Discourse Parsing. En *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 3578-3584.

Asher, N. y A. Lascarides, 2003. *Logics of conversation.* Cambridge Univ Pr, Cambridge.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22 (2): 249-254.

Carlson, L., D. Marcu y M.E. Okurowski, 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. En *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 85-112.

Carlson, Lynn, M.E. Okurowski, D. Marcu, 2002. RST Discourse Treebank. *LDC*.

Cristea, D., N. Ide y L. Romary, 1998. Veins theory: A model of global discourse cohesion and coherence. En *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 281-285.

da Cunha, I. y M. Iruskieta, 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12 (5): 563-598.

Danlos, L., 2008. Strong generative capacity of RST, SDRT and discourse dependency DAGSs. *Constraints in discourse*, 69–95.

Girill, T., 1991. Information chunking as an interface design issue for full-text databases. *Interfaces for Information Retrieval and Online Systems: The State of the Art*, 149-158.

Hearst, M.A., 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23 (1): 33-64.

Lehmann, C., 1985. Towards a typology of clause linkage. En *Conference on Clause Combining*, 181-248.

Mann, W.C. y M. Taboada, 2010. RST web-site. http://www.sfu.ca/rst/.

Mann, W.C.y S.A. Thompson, 1987. Rhetorical Structure Theory: A Theory of Text Organization. Marina del Rey. CA: Information Sciences Institute.

Marcu, D., 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 123-136.

Marcu, D. y A. Echihabi, 2002. An unsupervised approach to recognizing discourse relations. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 368-375.

Miltsakaki, E., R. Prasad, A. Joshi y B. Webber, 2004. Annotating discourse connectives and their arguments. En *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, 9–16.

O'Donnell, M., 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. En *Proceedings of the First International Conference on Natural Language Generation INLG '00*, 253-256.

Pardo, T.A.S. y M.G.V. Nunes, 2008. On the development and evaluation of a Brazilian Portuguese discourse parser. *Journal of Theoretical and Applied Computing*, 15 (2): 43-64.

Passonneau, R.J. y D.J. Litman, 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. En *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 148-155.

Polanyi, L., 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12 (5-6): 601-638.

Soricut, R. y D. Marcu, 2003. Sentence level discourse parsing using syntactic and lexical information. En *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 149-156.

Stede, M., 2008. Disambiguating rhetorical structure. *Research on Language & Computation*, 6 (3): 311-332.

Taboada, M. y W.C. Mann, 2006a. Applications of rhetorical structure theory. *Discourse studies*, 8 (4): 567.

Taboada, M. y W.C. Mann, 2006b. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8 (3): 423.

Thompson, S.A., R. Longacre y S.J.J. Hwang, 1985. Adverbial clauses. En: Shopen,T. (Ed.), Language Typology and Syntactic Description: Complex Constructions. Cambridge University Press, New York : 171-234.

Tofiloski, M., J. Brooke y M. Taboada, 2009. A syntactic and lexical-based discourse segmenter. En *Proceedings of the ACL-IJCNLP 2009*, 77-80.

Webber, B., 2009. Genre distinctions for discourse in the Penn TreeBank. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 674-682.

Webber, B., M. Stone, A. Joshi y A. Knott, 2003. Anaphora and discourse structure. *Computational Linguistics*, 29 (4): 545-587.

Wolf, F. y E. Gibson, 2004. Representing discourse coherence: A corpus-based analysis. En *Proceedings of the 20th international conference on Computational Linguistics*, 134-140.

7

Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque

Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque[1]

MIKEL IRUSKIETA, ARANTZA DIAZ DE ILARRAZA and MIKEL LERSUNDI

Abstract

This article presents a discourse annotation methodology based on Rhetorical Structure Theory and an empirical study of annotating a corpus of specialized medical texts in Basque. The annotation process includes two phases: segmentation and annotation of rhetorical relations. Phase one entails an initial study which leads to establishing linguistic criteria for sentence-based segmentation; a second phase focuses on annotation of rhetorical relations. After establishing discourse segments and rhetorical relations, the annotation process is analyzed and evaluated by means of the method commonly used in RST (Marcu 2000)*. Inconsistencies detected in the evaluation method lead the authors to redefine some criteria of the evaluation method. As a result of this work, a small annotated Basque-language corpus is provided to scientific community.*

Keywords: natural language processing, discourse structure, segmentation, rhetorical relations, evaluation method

1. Introduction

In the field of computational linguistics, discourse analysis tends to touch on different structural phenomena, including referential and relational structure. The main task of referential structure is coreference resolution, while the main task of relational structure is coherence relation assignment. Although many works refer to each of these phenomena, a limited number of studies have discussed corpus annotation at the discourse level in Basque. Existing studies have, however, considered the two phenomena of referential structure (Arregi et al. 2010; Ceberio et al. 2009) and relational structure (Iruskieta et al. 2011b; Iruskieta et al. 2009; Iruskieta et al. 2008; Barrutieta et al. 2002); the latter studies are related to the topic of this article.

Sophisticated language processing tools founded on knowledge from an annotated corpus are necessary for advanced applications such as information retrieval based on semantic knowledge, automatic text summarization, and machine translation. Consequently, in order to carry out these types of applications, it is important to have a corpus which is annotated at different linguistic levels, including the discourse level, as a point of reference.

This study focuses on discourse-level annotation, and is based on a corpus of abstracts of medical research articles taken from the *Gaceta Médica de Bilbao* (Medical Journal of Bilbao).[2] The corpus includes all 20 abstracts written in the journal in Basque through 2008, and contains 3,024 words. This corpus has been used in other research (da Cunha and Iruskieta 2010; da Cunha and Iruskieta 2009). For the purpose of this study, it will be utilized to describe problems arising during the processes of segmentation and rhetorical annotation.

The corpus annotation process employed herein utilizes a relatively small annotated corpus, but annotation phases and evaluation methods employed were critically analyzed to achieve an optimal annotation methodology. Indeed, a larger corpus and employing more than two annotators make it more difficult to perform a deep, critical analysis.

The general goal of this research is two-fold: i) to set out a methodology for annotating the relational structure of discourse (e.g., for annotating segments and rhetorical relations); and ii) to annotate a more extended corpus in Basque following this procedure. This will provide data about discourse structures for machine learning algorithms. Furthermore, with respect to future corpus annotations and applications, this study will also contribute significantly to the scientific community by providing a small but robust Basque-language corpus which has been annotated on a rhetorical level. Corpora available in other languages include English corpora (Taboada and Renkema 2011; Carlson et al. 2002), a German corpus (Stede 2004), Portuguese corpora (Pardo and Seno 2005; Pardo and Nunes 2004) and a Spanish corpus (da Cunha et al. 2011).

Relational structure is discussed in various discourse theories (Polanyi et al. 2004; Webber et al. 2003; Asher and Lascarides 2003; Moser and Moore 1996; Litman and Allen 1987; Cohen 1987; Grosz and Sidner 1986; Hobbs 1979). This empirical study is founded upon Mann and Thomson's (1987b) Rhetorical Structure Theory[3] (RST) since, apart from being applied to different languages, RST facilitates the representation of coherence in real texts, establishing relations among all the units in a tree-like structure. Furthermore, it is easy to find tools which facilitate working with RST and corpora, such as the RST annotation tool (O'Donnell 2000) and automatic discourse structure evaluation tool (Mazeiro and Pardo 2009). Finally, RST has been used for applications as diverse as text generation and summarization (Taboada and Mann 2006b) and for many other more advanced applications (Taboada and Mann 2006a). Consequently, this paper views RST to be the strongest framework for describing the relational structure of a text so that it can subsequently be implemented in advanced NLP applications.

RST is an applied, language-independent theory describing coherence between text fragments. It combines the idea of nuclearity—that is, the salience or importance of an individual fragment from within the discourse—with the effect that this relation has on the reader. Using the text, the author guides the reader, explicitly or implicitly letting him or her know which fragments are more important in relation to other fragments. As per the theory, these relations can be paratactic (N-N)[4]—when they establish relations between fragments that are equally important to the author (e.g. `LIST`, `CONTRAST`, `DISJUNCTION`, etc.)—or hypotactic (N-S), when they connect a less-important unit with a unit the author views to be more important (e.g. `ELABORATION`, `MEANS`, `PREPARATION`, `CONCESSION`, `CAUSE`, `RESULT`, etc.). Relations are defined in light of the restrictions established between the nucleus and satellite and by describing the effect they have on the reader. A more detailed explanation of RST can be found in Mann and Thompson (1988) and in Mann and Taboada (2010).

For the purpose of this article, the extended classification (Mann and Taboada 2010) is used. The set of 78 rhetorical relations proposed in Carlson et al. (2003) was ruled out due to the fact that it proposes some rhetorical relations which are dubious in terms of RST. For example, Stede (2008a) and Tofiloski et al. (2009) have criticized the `ATTRIBUTION` relation; the same reasoning underlies da Cunha and Iruskieta's (2010) proposal to discard embedded relations. Furthermore, given the initial phase of this study and its goals, it made sense to avoid a mutually agreed upon methodology for inter-annotator rhetorical relations and therefore steer away from Carlson et al. (2003) classification. Fragments and relations were viewed and annotated using the RSTTool[5] (O'Donnell 2000) program.

This study describes the methodological and linguistic elements of carrying out a rhetorical-level annotation on texts in Basque. During the course of research, various linguistic problems regarding the nature of rhetorical structure arose. These necessitated the establishment of a robust rhetorical structure annotation process. This study aims to answer the following basic questions:

- What is meant when describing an "elementary discourse unit" (EDU)? What linguistic forms must make up an elementary unit?
- In a segmented corpus, what should be measured to adequately describe inter-annotator agreement regarding elementary discourse units? In a rhetorical structure tree, what does Marcu's (2000) inter-annotator agreement measuring method involve?

Section 2 of this article lays out the theoretical framework and the methodology utilized to annotate the corpus and evaluate this annotation. Section 3 presents the results of the segmentation and raises some issues regarding it. Section 4 sets out the results of the annotation of rhetorical relations and suggests some shortcomings of the evaluation method which was employed. Finally, section 5 presents conclusions and establishes directions for future work.

2. Theory and methodology

When a human annotator wishes to annotate a text's relational structure, he or she must segment the text and later classify the relations between fragments. Generally speaking, the annotator can utilize one of the following strategies: a) determine relations during the segmentation process; or b) segment the text first and then determine the relations between all fragments, classifying all of these into a single structure (which is usually represented as a tree). In order to avoid circularity—where the analysis depends on the segmentation and the segmentation depends on the analysis (Taboada and Mann 2006b)—this study adopted the latter annotation strategy (strategy b). Consequently, the annotation was carried out by two annotators in two phases: i) first the corpus was segmented into units and ii) then, the rhetorical relations between units were determined. This approach leads to a more exact segmentation, paving the way to later consider the degree of agreement between rhetorical relations in greater detail.

Following Hovy (2010), this paper provides information on the profile of the annotators, annotation and adjudicating criteria used in this study. Both annotators were linguists who have annotated texts at other linguistic levels (morphosyntax, syntax and semantics), although neither had previously annotated texts in the framework of RST. The segmentation phase did not foresee a training phase. Segmentation was evaluated and it was decided that annotation should take place at the inter-sentential level. In subsequent works, the same corpus is annotated at the intra-sentential level (Iruskieta et al. 2011b). Nevertheless, a training phase was proposed as part of the rhetorical annotation phase because it became clear that the definitions of some relations were not well-understood by annotators. After noting how the relations were to be understood, an annotation process was established which was both incremental (bottom-up) and modular (sentence-by-sentence and paragraph-by-paragraph), as proposed in Pardo (2005). Finally, an adjudicator evaluated both annotations and resolved discrepancies, making a final decision by determining the most plausible relation. As a result of this work, this corpus can be consulted at both the intra-sentential and inter-sentential levels.[6]

Phase one was sub-divided into the following sub-phases: i) each annotator segmented the text using a minimal set of criteria; ii) this first segmentation was assessed in order to establish the final criteria for identifying the elementary segmentation unit; iii) the corpus was re-annotated and re-evaluated at the segmentation level; iv) rhetorical relations were annotated; and, finally, v) inter-annotator agreement was assessed using the evaluation system described in Marcu (2000).

The concepts of segmentation and rhetorical annotation can be contextualized using an example taken from the corpus (Figure 1). The Basque and English are extracted from the aforementioned medical journal; the English text was poorly written and thus was modified by the authors in order to make it easier for readers to fully understand the phenomena represented in the examples—as well as the segmentation and rhetorical annotation produced by one of the annotators. As can be observed in Figure 1, the annotation includes various types of elements:

a) Units and nodes. In Figure 1, the elementary unit is marked with horizontal lines (the segments and translations thereof are found underneath these). After segmenting the text, the annotator must relate these units. The text contains 10 units numbered from 1 to 10. The spans or nodes (groups of units) are represented by pairs of numbers which indicate the first and last unit of their component elements. Our example includes nine spans: 2-3, 2-5, 4-5, 6-7, 6-10, 2-10, 9-10, 8-10 and 1-10.

b) Nuclearity and relations. All segments or units are considered to be either a nucleus or a satellite. The concept of nuclearity[7] (nucleus and satellite) is important when establishing rhetorical relations, since it determines whether these relations are paratactic or hypotactic in relation to the other units in the text.

Figure 1: A rhetorical structure tree for text GMB0401

In Figure 1, units below straight vertical lines represent the nuclei of hypotactic relations (2-2,[8] 2-3, 7-7, 6-7, 6-10, 2-10 and 9-10) while those units found underneath diagonal lines are the nuclei of paratactic relations (4-4, 5-5, 9-9, and 10-10). Other elements are satellites of hypotactic relations (1-1, 2-5, 3-3, 4-5, 6-6, 8-8, and 8-10). The span which covers the entire text (1-10) cannot be related to any other span, and consequently, has no nuclearity.

Relations between segments are represented using arrows extending from the satellite towards the nucleus; for example, the BACKGROUND relation connects satellite segment 2-5 to its nucleus, 6-10.[9] As such, annotators interpret which units are most important for understanding the text.

The main concept—that is, the idea presenting the most important unit of tree structure (Mann and Thompson 1987a)—is represented with straight vertical lines if it is a hypotactic relation or under diagonal vertical lines if it is a paratactic relation. In our example (Figure 1), unit 7-7 is the main unit of the rhetorical structure. There are eighteen cases of nuclearity in this example: i) seven units function as satellites: 1-1, 2-5, 3-3, 4-5, 6-6, 8-8 and 8-10 and ii) the other eleven units function as nuclei: 2-2, 2-3, 4-4, 5-5, 7-7, 6-7, 6-10, 2-10, 9-9, 10-10 and 9-10.

In this example, the annotator interpreted the rhetorical relations presented in Figure 1 as follows: i) PREPARATION for the article, by means of the title ([1-1 > 2-10]); ii) laying out the BACKGROUND of the issue to be considered: the profile of users using the emergency services ([2-5 > 6-10]); iii) demonstrating why the study is interesting using the MOTIVATION relation ([6-6 > 7-7]), and iv) highlighting the RESULTS ([6-7 < 8-10]).

Within the BACKGROUND relation there are three other relations explaining how the number of urgent medical visits has risen: two ELABORATIONS ([2-2 < 3-3] and [2-3 < 4-5]) and one multi-nuclear CONJUNCTION[10] relation ([4-4 = 5-5]).[11]

Similarly, the RESULT relation subsumes the PREPARATION relation ([8-8 > 9-10]) and the multi-nuclear CONJUNCTION relation ([9-9 = 10-10]).

Though only a single interpretation has been presented for the example text, Mann and Thompson (1987b) state that one annotator may have more than one valid interpretation of a given text. In light of this, each annotator was asked to present only a single interpretation of each text.

3. Text segmentation

The previous section explained the general methodology employed in this study and provided some comments on the annotation schema. This section begins by explaining the basic principles of segmentation in detail. Then, it will lay out some problems related to segmentation—namely agreement and causes for disagreement between annotators—and finally will conclude by describing the consensually arrived upon decisions taken with regard to the segmentation process.

3.1. Basic principles

Rhetorical segmentation of a text entails specifying the rhetorical units. This is a basic stage in the rhetorical annotation process, since inter-annotator disagreements negatively affect the assignment of later relations.

The literature review pointed out the fact that there is not a clear definition regarding what constitutes an elementary discourse unit. For example, a discourse unit could be: i) a clause or sentence (Carlson et al. 2003); ii) a sentence with a finite verb (da Cunha and Iruskieta 2010) or iii) groups of sentences (Hearst 1997).

Mann and Thompson's (1987b: 224) original definition of an elementary unit aimed to be founded on a "theory-neutral classification" in which units could "have independent functional integrity". Carlson et al. (2003) argue that this definition is not sufficiently explicit since the boundary between discourse and syntactic is at times undefined. Given this, and in order to increase inter-annotator reliability, Carlson et al. (2003) define segmentation more broadly, specifying which kinds of clauses

constitute EDUs and which do not. Their goal is to present the most rhetorically enriched and robust corpus, the RST Discourse Treebank (Carlson et al. 2002), to the scientific community. Consequently, segmentation must be as refined as possible regardless of whether some syntactic forms constitute a rhetorical unit.

Another segmentation proposal, which adopted a less refined granularity but was more faithful to the original nature of RST, was carried out by Tofiloski et al. (2009). For the sake of this study, it seemed most adequate to begin with a deliberate definition of segmentation which would bring out the problems with the process; consequently, this study followed Mann and Thompson's (1987b) original definition of segments.

3.2. Analysis of agreement and decisions

Based on the definition of segment proposed in Mann and Thompson (1987b), the two annotators segmented the corpus independently without consulting each other.

Segmentation agreement was assessed using various measures. Percent agreement (Hearst 1997; Marcu 1999; Passonneau and Litman 1993) is used to measure agreement between annotators. Precision and recall can be used to evaluate the reliability of the segmentation algorithm (see Passonneau and Litman 1993); note that Afantenos et al. (2010) used F-score, a measure evaluating pairs of human annotators, which combines both precision and recall, for this purpose. Finally, the Kappa coefficient subtracts the value of expected chance agreement (Carletta 1996) when computing the agreement between annotators; Kappa was used by Miltsakaki et al. (2004), Hearst (1997) and Tofiloski et al. (2009), the last of whom compare Kappa values with F-scores.

In order to assess the degree of agreement, the segmented texts were manually evaluated. Agreement data were compared using Kappa value. It is generally accepted that Kappa statistics are more robust than percentages or F-score. This article applied the Kappa measures as per Landis and Koch (1977) and interpreted the coefficients for strength of agreement as per Cohen (1987).

The Kappa value measures agreement, correcting the expected chance agreement as follows:

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) represents the proportion of times that annotators' segments match and P(E) represents the proportion of times that annotators would be expected to agree by chance.

Table 1: Segmentation cross tabulation of boundaries

		A2		
		Yes	**No**	**Total**
A1	**Yes**	243	0	243
	No	36	202	238
	Total	279	202	481

$$k = \frac{0.92 - 0.5}{1 - 0.5} = 0.85$$

The Kappa value of 0.85, according to Cohen (1987), is almost perfect (Table 1). The Kappa value was calculated by considering the contents of the body of the document—including titles, parentheses, and verbs—as candidates indicating elementary units. What is remarkable, however, is that all of A1's segment boundaries correspond with A2's. This fact illustrates the different levels of granularity applied by the two annotators, indicating that they interpreted the starting definition differently.

This degree of agreement does not guarantee inter-annotator reliability in the next stage of annotating rhetorical relations. However, agreements in the rhetorical annotation phase depend to a great extent on the results of the segmentation. As the degree of agreement in segmentation is key for the next stage—which compares nuclearity and relations—this segmentation results are lower than those obtained in similar studies and cannot be accepted as valid. Therefore, an analysis of the underlying reasons for disagreement was necessary; this would lead to making some decisions to increase inter-annotator agreement in the segmentation stage.

The aforementioned differences owe to differing levels of granularity: A2 adopted a finer granularity than A1. In fact, annotator A2 established segment boundaries in all of the positions marked by A1 and in 36 other positions. The results in Table 1 prove that the initial definition was not sufficiently explicit to allow two annotators to arrive at a substantial degree of agreement without consulting each other. Thus, explicit decisions are needed with regard to segmentation.

Tables 3, 4 and 5 present inter-annotator agreements, disagreements and decisions; these are explained through examples and commented (see Table 2 for an explanation of the glosses employed in examples) on in subsequent sub-sections.

Table 2: Glosses used in examples

Gloss abbreviations	Explanation	Basque form[12]
A	Absolutive in auxiliary glosses	
AUX	Auxiliary	
COMP	Complementizer	-(e)n- -(e)la
D	Dative in auxiliary glosses	
DET	Determiner (article)	-a
E	Ergative in auxiliary glosses	
IMPF	Imperfective	-t(z)en
INSTR	Instrumental	-(e)z
NOM	Nominalizer	-t(z)e-
PL	Plural	
PRF	Perfective	-i; -tu
PTCP	Adverbial participle	-ta; rik

Table 3: Agreement regarding segmentation

	Linguistic forms	EDU
Agreement	Non-embedded clauses with finite verbs	Yes
	Complement clauses	No
	Relative clauses	No
	Verbal nominalization	No

3.2.1. Agreement in segmentation and establishing the elementary unit

Both annotators considered clauses containing a finite verb without syntactic subordination to be elementary units. Below, the linguistic phenomena on which annotators agreed are explained.

i) Non-embedded clauses with finite verbs.

Example (1) is a typical case in which both annotators segmented the text into two elementary units since there are two finite verbs: one is the verb *da* '(it) is' and the other is *adierazten du* [indicate.IMPF AUX.3A/3E] (it) indicates (that)'.

(1) *[Hipertentsiorako tratamendu farmakologikoa konplexua **da**,] [hori **adierazten du** medikuek errezetutako eta laborategi farmazeutikoek eskainitako farmako aukera zabalak]* GMB0801
[Pharmacological treatment of hypertension **is** complicated;] [the vast quantity of drugs offered by pharmaceutical laboratories and prescribed by physicians **indicates** this.] **Translation**

ii) Complement clauses.

Complement clauses are not new segments. In (2), the complement clause is created by adding the suffix *-(e)la* 'that' to both auxiliary verbs: *da* (in *gertatzen dela* [take place.IMPF AUX.3A.COMP] 'that… (it) takes place', and *luzatzen dela* [prolong.IMPF AUX.3A.COMP] 'that… (it) may be prolonged'). This was not considered an elementary unit. In this case neither of the complement clauses was considered to be connected via the coordinating conjunction, the marker *eta* 'and'.

(2) *[Horrela gauzak, aurreratu behar zaie odoljarioa sarritan gertatzen de**la** eta egun batzuetan luzatzen de**la**, nahiz eta kantitate urria izan.]* GMB0202
[Thus, it is important to stress to patients **that** the probability of bleeding taking place is high and **that** it may be prolonged over time, though this may be limited.] **Translation**

iii) Relative clauses.

Relative clauses are not new segments[13]. See example (3) below: the relative clause *eskaintzen digun* [offer.IMPF AUX.3A/1D.PL/3E.COMP] 'that is offered' was not considered a unit.

(3) *[Merkatuak eskaintzen digu**n** espezialitate merkeena aukeratuko bagenu 6.463.400,35€-ko aurrezpena lortuko genuke.]* GMB0801
[If we selected the most inexpensive medicine **that** is offered on the market we could realize savings of 6,463,400.35€.] **Translation**

iv) Verbal nominalization.

Clauses containing a nominalized verb were not considered elementary units. The presence of the nominalized form *egitea* [execute-NOM-DET] 'the execution') in (4) does not define a segment.

(4) *[Hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskriptibo bat **egitea** aproposa iruditu zaigu.]* GMB0401
[Consequently, we believed that **the execution** of a study designed to determine the profile of Galdakao hospital emergency room users would be appropriate.] **Translation**

3.2.2. Disagreement in segmentation

The 36 cases (Table 1) of segmentation disagreement (Table 4) were classified as follows:

- syntactic subordination:[14]
  - 22 cases (61.1%) involving non-finite verbs and markers of subordination
  - 4 cases (11.1%) involving finite verbs and markers of subordination
- conjunction or juxtaposition with markers and verbal ellipsis: 8 cases (22.2%)
- and segmentation errors or *lapsus*: 2 cases (5.5%).

Table 4: Disagreement regarding segmentation

	Linguistic forms
Disagreement	Syntactic subordination with a non-finite verb
	Syntactic subordination with a finite verb
	Conjunction or juxtaposition with markers and verbal ellipsis

As indicated above, all of these discrepancies are based on the differing grades of granularity applied by the annotators when analyzing the text. Basically, whereas A1 viewed units as functionally independent whenever they included an independent clause or a non-subordinate finite verb (except titles, which had no finite verb but which were nevertheless viewed as units), for A2 clauses with a verb—whether subordinate or non-finite—as well as titles were viewed as units.

Examples of the cases which produced inter-annotator disagreement are presented below.

i) Syntactic subordination with a non-finite verb and marker of subordination.

In example (5), the participle *aztertuta* [analyze.PRF.PTCP] 'after having analyzed' contains a non-finite verb (*aztertu* 'analyze') and a marker of subordination (*-ta* '-(e)d') which conveys the perfect tense. This led to disagreement between the two annotators.

(5) *[7 itemak aztertu**ta**,] [estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05).]* GMB0701
[After having analyze**d** the 7 items,] [statistically significant differences were found between the group of cancer patients and the patients suffering from other pathologies (p<0.05).] **Translation**

In example (6) the modal aspect of the gerund *erabiliz* [utilize.PRF.INSTR] 'utilizing' led to the disagreement.

(6) *[Ikerketa berriek,]*[15] *["microarrays" teknika erabili**z**] [pronostiko txarra duen bularreko minbiziaren azpitalde bat hauteman dute.]* GMB0702
[Recent studies,] [utiliz**ing** the "microarrays" technique,] [have identified a sub-group of breast cancers with a very low prognosis.] **Translation**

ii) Syntactic subordination with a finite verb and marker of subordination.

In the cases shown in (7) and (8), the causal subordinate clauses marked by the subordinating suffixes *–(e)nez* (*zehaztu gabe daudenez* [specify.PRF instead are.3A.COMP.INSTR] '(they) are not specified') and *–(e)lako* (*narriatu delako* [deteriorate.PRF AUX.3A.because] 'because (it) has deteriorated') were treated differently by both annotators.

(7) *[Kitokeratina basalak zehaztu gabe daude**nez**,] [txosten anatomopatologikoetan erabili ohi diren parametroen bidez "basal-like" tumoreen azpitaldea hauteman dezakegu, gaitzaren egoera oso goiztiarrean.]* GMB0702
[**Given** that basal cytokeratins are not specified,] [the use of parameters regularly present routinely in anatomic pathology reports allows us to identify a subgroup of "basal-like" tumors at very early stages of the disease.]**Translation**

(8) *[Bere gorputzaren ohiko funtzionamendua narriatu de**lako**] [dago ospitalean.]* GMB0501
[He is in the hospital] [**because** his general health has deteriorated.] **Translation**

iii) Conjunction or juxtaposition with markers and verbal ellipsis.

Annotators also analyzed the verbal ellipsis in *nabaritzen zen* [notice.IMPF AUX.3A] '(it) was noticed' and its accompanying coordinating conjunction differently (example 9).

(9) *[Zazpi kasutan hiperkapnia **nabaritzen zen**] [eta 26 kasutan hipoxemia.]* GMB0001
[Hypercapnia **was noticed** in 7 cases] [and hypoxemia in 26 cases.] **Translation**

There was also disagreement in contexts where coordinating conjunctions presented contrasting contents. In example (10), the subject quality was negated in the first clause and ellipsis used instead of repeating the verb *ez dituzte adierazten* [Not AUX.3A.PL/3E.PL express.IMPF] '(they do not) express (them)', adding affirmation via the particle *bai* [yes] '(they) do (express)' in the second clause.

(10) *[Tumore horiek **ez** dituzte hormona hartzaileak eta c-erb-B2 onkogenea adierazten;] [eta **bai**, ordea, epitelio basaleko geruzaren zelulei dagozkien kitokeratinak.]* GMB0702
[These tumors **do not** express hormone receptors or the c-erb B2 oncogene,] [however **they do (express)**[16] heir own citokeratins from cells from the basal epithelial layer.] **Translation**

The following section presents all of the decisions which were made to create a broader definition of segmentation at the inter-sentential level.

3.2.3. Decisions taken after evaluating the segmentation process

Table 5 summarizes the decisions taken after assessing the first segmentation attempt. Before moving on to the rhetorical annotation phase, the text is re-segmented with the aim of obtaining a much higher degree of agreement in terms of the segmentation of the corpus.

Table 5: Decisions regarding segmentation

	Linguistic forms	EDU
Decisions	Adverbial subordinate clauses with finite verbs	No
	Adverbial subordinate clauses with non-finite verbs	No
	Conjunctions or juxtaposition with verbal ellipsis	No
	Conjunctions of verbs with only one finite verb	No
	Period with or without a finite verb	Yes
	Colon followed by a finite verb	Yes
	Colon not followed by a finite verb	No
	Semicolon without a finite verb	No
	Discourse marker without a finite verb	No
	Parenthetical clauses without a finite verb	No

The following decisions were made:

i) Do not segment adverbial subordinate clauses with finite verbs.[17]

In example (11) it was decided to classify the verbal suffix and marker of subordination *-(e)lako* 'because' as a single segment.

(11) *[Bere gorputzaren ohiko funtzionamendua narriatu de**lako** dago ospitalean.]* GMB0501
[He was in the hospital **because** his general health had deteriorated.] **Translation**

ii) Do not segment adverbial subordinate clauses with non-finite verbs.

In this case, the participle *aztertuta* [analyze-PRF-PTCP] 'having (been) analyzed', which conveys the perfect tense, was not segmented (example 12).

(12) *[7 itemak aztertu**ta**, estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05).]* GMB0701
[After having analyze**d** the 7 items, statistically significant differences were found between the group of cancer patients and the patients suffering from other pathologies (p<0.05).] **Translation**

iii) Do not segment conjunction and juxtaposition clauses with verbal ellipsis.

In cases of coordination (example 13) or juxtaposition (example 14) which included verbal ellipsis, the fragment was considered to be only one elementary unit.

(13) *[Zazpi kasutan hiperkapnia nabaritzen zen **eta** 26 kasutan hipoxemia.]*GMB0001
[Hypercapnia was noticed in 7 cases **and** hypoxemia in 26 cases.] **Translation**

(14) *[24 pazientek bronkiektasiak zituzten (1998an ingresatuko %12k)**;** 15 pazientek, BGBK.]* GMB0201
[24 patients had bronchiectasis (12% of all sick patients admitted with this diagnostic in 1998); 15 patients (had)[18] COPD.] **Translation**

iv) Do not segment conjunctions of verbs with only one finite verb.

A verb which is part of a verb coordination does not constitute an elementary unit. In example (15) only the second verb which is an object of the coordinating conjunction *areagotzen du* is finite ([increase.IMPF AUX] in the translation, this is indicated in the first verb, 'causes...to increase'); thus, this must be considered a verbal coordination and the entire fragment must be considered a unit.

(15) *[horrek heriotza-tasa **handitu** **eta** ospitaleko ingresu berrien kopurua **areagotzen** du.]* GMB0201
[this **causes** the number of new hospital admissions **to rise and** the mortality rate **to increase.**] **Translation**

v) Segment clauses separated by a period, even if they do not contain a verb.

A period can separate clauses even if there is not a finite verb in the phrase (example 16).

(16) [*Hona hemen oin malgua izateagatik kalkaneo-stop teknika erabiliz gure zerbitzuan ebakuntza egin diegun haurrek izandako emaitzak.*] GMB0601
[(We present)[19] results obtained in patients treated by our department for juvenile onset flexible flat foot using the calcaneus-stop technique.] **Translation**

vi) Segment clauses separated by a colon if the following clause or sentence contains a finite verb.

A colon can have a discourse function if it functions as a title or a cataphoric or syntactic function if it refers to the information contained in the object of the verb. Evidence for this is found in (17): the first colon has a discourse function, since there is a finite verb in the following fragment, while the second colon has a different function, presenting the information which is contained in the complement clause.

(17) *[Emaitzak:] [Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5).*] GMB0401
[Results:] [The average user is as follows: male (51.3%), middle-aged (43.2 years old), and treated for trauma pathology (50.5%).] **Translation**

vii) Do not segment a fragment simply because it contains a semicolon.

A semicolon in and of itself is not sufficient for segmenting a unit into two (example 18).

(18) *[24 pazientek bronkiektasiak zituzten (1998an ingresatuko %12k); 15 pazientek, BGBK.]* GMB0201
[24 patients had bronchiectasis (12% of all sick patients admitted with this diagnostic in 1998); 15 patients (had) COPD.] **Translation**

viii) Do not segment clauses with a discourse marker but no finite verb.

Clauses with a discourse marker but no finite verb were not considered units (example 19).

(19) *[Tumore horiek ez dituzte hormona hartzaileak eta c-erb-B2 onkogenea adierazten; **eta** bai, **ordea**, epitelio basaleko geruzaren zelulei dagozkien kitokeratinak.]* GMB0702
[These tumors do not express hormone receptors or the c-erb B2 oncogene, **however** they do (express)[20] their own citokeratins from cells from the basal epithelial la**y**er.] **Translation**

ix) Do not segment parenthetical clauses without a finite verb.

(20) *[Gure ospitalean PTRko infekzio kroniko guztiak bi alditan eginiko ordezko protesien bidez tratatu ziren (LCCK protesiekin: Legacy Constrained Condylar Knee zementu antibiotikoarkin).]* GMB0802
[The treatment of chronic knee prosthesis infections carried out in our hospital consisted of all of the cases of a two-time prosthesis replacement (with LCCK prosthesis: Legacy Constrained Condylar Knee with antibiotic-loaded cement).] **Translation**

In sum, the units which were segmented at inter-sentence level were clauses containing a finite verb without syntactic subordination. This guideline was only disregarded in the cases of punctuation markers such as a period, colon with a discourse function, and colon which served as the title of an abstract without a finite verb. All of these cases constituted elementary segmentation units. Thus, on the one hand, various forms of syntactic subordination—complement clauses, relative clauses, verbal nominalization, and clauses without a finite verb such as those connected with a discourse marker or parenthetical clauses—were not considered discourse units. On the other hand, cases with verbal ellipsis and conjunctions with only one finite verb were also considered to contain only one unit.

However, the authors of this study are currently developing an automatic discourse parser at intra-sentence level (Iruskieta et al. 2011a) that uses a syntactic parser based on machine learning techniques (Arrieta 2010). So far, this parser obtains an F-score of 57%, which is far from the results—F-scores between 73% and 85%—obtained for other discourse parsers based on machine learning techniques for French (Afantenos et al. 2010), and parsers based on rules for English (Tofiloski et al. 2009; Soricut and Marcu 2003) or Spanish (da Cunha et al. 2010).

4. Evaluation of the rhetorical annotation

After having finalized the set of elementary segments, the corpus was rhetorically annotated by both annotators following an incremental and modular strategy.

4.1. Methodology

The annotation was evaluated as per the methodology proposed in Marcu (2000). Although this method was designed to compare manually created trees with automatically-segmented trees, in this study the same technique was used to evaluate annotations carried out by two different annotators.

In order to describe this evaluation method, another text from the corpus is provided as an example (Figure 2). Table 6 presents agreements on the four factors which were analyzed: i) dividing the text into units (EDU), ii) creating a tree structure for these units (that is, the nodes or spans), iii) determining the most important unit in a relation: nuclearity (N/S), and iv) determining the type of rhetorical relation (RR).

Figure 2: Text GMB0701

The first column of Table 6 (Node) contains all units and spans identified by both annotators and their lengths. The other columns present the evaluation of these units by each annotator. Columns two and three (EDU) show whether each annotator segmented the elementary unit in question. Where the annotator segmented the unit, it is marked with a '√'. Columns four and five (Span) also are marked with a '√' when the annotator identified this EDU or group of units; spans which were not identified are marked with a '-'. Columns six and seven (N/S) describe the nuclearity of the unit: satellites are marked as 'S' and nuclei are marked as 'N'. The final two columns (RR) present the rhetorical relation. This method sets out various indications. On the one hand, it establishes all spans in multi-nuclear relations via the name of the rhetorical relation (LIST), and on the other hand, it establishes all spans with nuclearity value (N) as NUCLEUS and those with value (S) with the name of the corresponding rhetorical relation (ELABORATION, RESULT, PREPARATION, and MEANS). Disagreements are shaded gray to make them easier to identify.

Table 6: Quantitative evaluation of text GMB0701

	EDU		Span		N/S		RR	
Node	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
1-1	√	√	√	√	S	S	Preparation	Preparation
2-2	√	√	√	√	N	N	Nucleus	Nucleus
3-3	√	√	√	√	N	N	Nucleus	Nucleus
4-4	√	√	√	√	S	N	Elaboration	Nucleus
5-5	√	√	√	√	N	N	List	List
6-6	√	√	√	√	N	N	List	List
7-7	√	√	√	√	N	N	List	List
8-8	√	√	√	√	N	N	List	List
6-8			√	√	N	N	List	List
5-8			√	√	S	S	Result	Result
2-8			√	√	N	N	Nucleus	Nucleus
3-4			√	-	S	-	Means	-
2-4			√	-	N	-	Nucleus	-
4-8			-	√	-	S	-	Elaboration
3-8			-	√	-	S	-	Means

Table 6 demonstrates that the annotators completely agreed about the segmentation of the text, since both annotators created the same eight elementary units (EDU). As for groups of units (Span), the annotators disagreed about two groups of units (3-4 and 2-4 for A1 and 4-8 and 3-8 for A2). These two disagreements affected both judgments of nuclearity (N/S) and the identification of the rhetorical relation (RR) (MEANS and NUCLEUS for A1 and ELABORATION and MEANS for A2). Furthermore, annotators disagreed about the nuclearity (N/S) of node 4-4 and its relation (ELABORATION for A1 and NUCLEUS[21] for A2). These observations are analyzed in further detail in subsections 4.2.1 and 4.2.2.

Table 7 provides data on average precision (the number of elements selected correctly in relation to the number of total elements selected) and recall (the number of elements found correctly in relation to the number of total elements found), focusing on the factors analyzed in Table 6–that is, EDU, Span, Nuclearity (N/S), and Relation (RR). As we have seen, the degree of agreement for elementary units (EDU) and groups of units (Span) is key when it comes time to analyze the different interpretations of the relations between nodes. If agreement is low for these first two factors, the factors of nuclearity and rhetorical relation will have a low rate of agreement.

Table 7: Results for text GMB0701

	EDU	Span	N/S	RR
A1	8	13	13	13
A2	8	13	13	13
Agreement	8	11	10	10
Precision	8/8	11/13	10/13	10/13
Recall	8/8	11/13	10/13	10/13

Table 8 presents global data for the corpus annotation.

Table 8. Global quantitative results

	EDU	Span	N/S	RR
A1	233	432	432	432
A2	233	432	432	432
Agreement	233	386	328	252
Precision	100.00%	89.35%	75.93%	58.33%
Recall	100.00%	89.35%	75.93%	58.33%

Table 8 demonstrates that the decisions made regarding segmentation were clear: annotators completely agreed on both precision and recall for elementary units (EDU). Note that although the corpus was annotated incrementally and modularly, there was a relatively high degree of disagreement regarding spans—10.65%. This value affected the two following factors. Disagreements regarding nuclearity rose significantly, to 24.07%, while the biggest disagreement regarded the relation factor, at 41.67% disagreement—that is, 58.33% agreement. With regard to the relation factor, these results are lower than those obtained in similar studies. For example, as da Cunha et al. (2011) mention, analysts of a Spanish text had agreement percentages of 76.81% (precision) and 78.48% (recall), whereas for an English text, analysts obtained values of 83.4% for precision and recall, with automatic parser results of 47.0% (recall) and 78.4% (precision) (Marcu 2000).

4.2. Reflections on methodology

This subsection reflects on the inadequacies of the evaluation methodology which was adopted (Marcu 2000). Some of these inadequacies were detected in da Cunha and Iruskieta (2010), where a qualitative evaluation was proposed to avoid them. Here is an explanation of the methodology:

4.2.1. The relation factor interferes with nuclearity.

Since the annotation of relation bears nuclearity in mind, the aspects of nuclearity and relation are muddled. Consequently, the authors believe that this methodology does not adequately encompass the agreement that there was in regard to relations.

This is made clear by comparing the results presented in Table 6 with the actual relations annotated by annotators A1 and A2 in Figure 2. For example, Table 6 contains thirteen relations: PREPARATION, MEANS, ELABORATION, RESULT, five LIST relations and four NUCLEUS relations. As is clear from the example shown in Figure 2, both annotators identified the same number of relations, six: PREPARATION, MEANS, ELABORATION, RESULT and two LIST relations. We believe that agreement must be evaluated in terms of these six relations (see Table 9). The reason for so much disagreement stems from the fact that Marcu's (2000) method includes the NUCLEUS label among its Relation factors. However, this label cannot be considered a RST relation, since it refers to the spans which constitute the NUCLEUS in hypotactic relations. Therefore, the difference in agreement arises because in this method, every nucleus/satellite has a label describing its relation.

Given RST's definition of rhetorical relations, NUCLEUS cannot be viewed as a RST relation. Consequently, it should not be considered when measuring inter-annotator agreement about relations. Table 9 presents the precision and recall of agreement for RST rhetorical relations.

Table 9: Comparing agreement among relations, GMB0701

A1	6
A2	6
Agreement	5
Precision	5/6
Recall	5/6

In Table 7, the degree of agreement for recall in the Relation factor was 10/13, or 76.92%. In Table 9, however, the agreement between results rises to 5/6, or 83.33%.

Table 10 presents the weight of each relation in terms of agreement about the relation. The first column includes the relations from Table 6, while the second includes the weight of each relation, calculated for the two spans that participate in each relation (cf. the methodology employed in this study) and the third includes its percentage. The fourth column presents the weight of each relation calculated only for RST relations and the fifth presents its corresponding percentage.

Table 10: Comparing weight: span based comparison/relation based comparison, GMB0701

Relation	RR agreement (methodology)	%	RR agreement (RST)	%
Preparation	1/13	7.69%	1/6	16.66%
Means	1/13	7.69%	1/6	16.66%
Elaboration	1/13	7.69%	1/6	16.66%
Result	1/13	7.69%	1/6	16.66%
List	5/13	38.46%	2/6	33.33%
Nucleus	4/13	30.76%	-	-

Table 10 demonstrates that the weight of nuclear relations increases while the weight of multi-nuclear relations decreases.

Agreement regarding the NUCLEUS annotation is more frequent than agreement about actual relations, since only span and nuclearity must overlap for this annotation to be considered an agreement. Note that both annotators labeled different relations, as in Figure 3.

Considering the disagreement about example represented in Figure 3, we can see that the annotators indeed disagreed about the relations: while A1 annotated the span with the ELABORATION relation, A2 interpreted the relation as being more specific and labeled it as EVIDENCE.

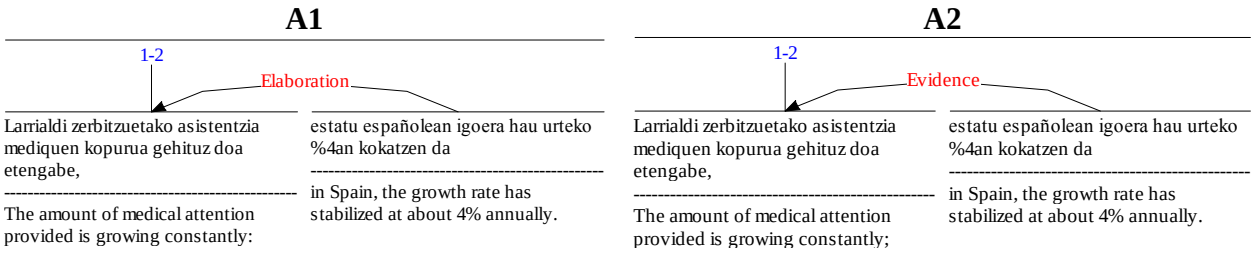


Figure 3: Disagreement regarding relation, GMB0401

A representation of this example using the methodology adopted in this study reveals that there is some degree of agreement with regard to the relation factor (see Table 11).

Table 11: Evaluation of the annotation of Figure 3, GMB0401

	EDU		Span		N/S		RR	
Node	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
1-1	√	√	√	√	N	N	Nucleus	Nucleus
2-2	√	√	√	√	S	S	Elaboration	Evidence

Table 12: Results for Figure 3, GMB0401

	EDU	Span	N/S	RR
A1	2	2	2	2
A2	2	2	2	2
Agreement	2	2	2	1
Precision	2/2	2/2	2/2	1/2
Recall	2/2	2/2	2/2	1/2

It is unjustifiable to argue that there is agreement regarding rhetorical relations in Figure 3 if RST relations are being measured. Agreement for the Relation factor (1/2) established by the methodology adopted in this study, as shown in Table 12, adequately reflects agreements about span and nuclearity but demonstrates a lack of agreement with regard to relation. The evaluation table demonstrates a recall value of 1/2 or 50%, reflecting the fact that the two annotators disagreed about the relation. This degree of agreement does not refer to agreement about the Relation factor (which was 0) but rather refers to the agreement about nuclearity.

4.2.2. Descriptive insufficiency.

The composition of relations is reflected in labels but not in their associations (Marcu 2000: 436):

> *This evaluation assumes that rhetorical labels are associated with the children nodes, and not with the father nodes, as in the formalization. (…) The rationale for this choice is the fact that the analysts did not construct only binary trees; some of the nodes in their manually built representations had multiple children.*

The methodology does not adequately compare the N/S and Relation factors when the annotators disagree about attachment node (da Cunha and Iruskieta 2010).

To illustrate the fact that the methodology does not adequately reflect agreement about relations; consider what happens when two annotators attach the same relation to different levels or nodes of the tree. The agreement reflected in Figure 2 and depicted in Table 6 cannot measure agreement with regard to the ELABORATION relation (4-4 for A1 and 4-8 for A2 are both associated with the same unit 4, and both have the same central unit, 4-4) and MEANS relation (3-4 for A1 and 3-8 for A2 are both associated with unit 2-2, with the same central unit, 3-3), since it cannot compare the spans of these relations. The composition is certainly different in both relations, but this composition is not a consequence of these relations but rather reflects the attachment node of another relation, RESULT. Though both annotators agree that this RESULT relation is a satellite (5-8 for both A1 and A2), agreement about its nucleus is not reflected: even though both are annotated NUCLEUS, they have different nuclearity for A1 (2-4) and A2 (4-4).[22] Moreover, as mentioned previously, according to Marcu's method, agreement for the ELABORATION and MEANS relations cannot be compared; consequently, this is the root of the disagreements about attachment node to another relation, RESULT. The portion of Table 6 which demonstrates this is reproduced in Table 13.

On the other hand, consider an alternative method of comparing the nodes, focusing partially on the nuclearity of unit 4-4. In Figure 2, unit 4-4 is a satellite (S) in the ELABORATION relation for both annotators, but when A2 associates another relation above unit 4-4, it is now the nucleus (N) in this new diagram. In cases with different associations, this method (Marcu 2000) places intense value on the agreement in relations, especially if these occur at the lowest levels of the rhetorical structure tree. In other words, the method is based on comparing the composition of these relations.

Table 13: Descriptive insufficiency, GMB0701

	EDU		Span		N/S		RR	
Node	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
4-4	√	√	√	√	S	N	Elaboration	Nucleus
2-4			√	-	N	-	Nucleus	-
4-8			-	√	-	S	-	Elaboration
3-8			-	√	-	S	-	Means
3-4			√	-	S	-	Means	-
5-8			√	√	S	S	Result	Result

In short, the authors believe that an evaluation method must offer a description of relations without confusing nuclearity and relation, a method which describes the composition and attachment node of the Relation factor.

5. Conclusions and future research

For the first time, this article presents the results of an empirical study which analyzes and discusses a segmentation proposal using RST theory for texts in Basque. This represents a fundamental step forward for rhetorical segmentation tasks in Basque. Two human annotators annotated a specialized corpus comprised of medical texts. The study defined the primary rules for inter-sentential segmentation, and also applied and explained the annotating method. The study clearly established the segmentation criteria and measured discrepancies between annotators. Special emphasis has been placed on identifying segments given their critical place of importance in the rhetoric structure.

Moreover, another interesting contribution of this paper is that the first Basque texts annotated with RST have been made available online.[23]

An annotation performed using the method commonly utilized in RST (Marcu 2000) was analyzed and evaluated, leading to the finding of two main inconsistencies in the method: i) the confusion between the annotation of nuclearity and rhetorical relation and ii) the lack of descriptiveness.

The authors are currently striving to develop an automatic evaluation method which can move beyond the methodological errors mentioned in section 4.1 of this paper, a method which also bears in mind other factors such as the composition and attachment node of relations.

They are also working on how to implement these segmentation decisions automatically (Iruskieta et al. 2011a). Such a method will also consider whether there are linguistic forms which show rhetorical relations on the clause-level and will test the extent to which these relations may derive from syntactic structures (Iruskieta et al. 2011b). By doing so, it will be possible to identify patterns which can later be incorporated into a system to automatically analyze discourse structures in Basque.

Bionotes

Mikel Iruskieta is lecturer of Basque language and literature at the University of the Basque Country. His methodological interests include text parsing and knowledge and discourse representation. He has worked mainly on text analysis applications such as machine translation, text summarization and knowledge extraction. Email: mikel.iruskieta@ehu.es

Arantza Diaz de Ilarraza is professor of computer languages and systems at the University of the Basque Country. She received her PhD in Computer Science from the University of the Basque Country in 1990. She is a researcher in the field of Natural Language Processing. Her research interests include the development of natural language processing resources, machine translation and linguistic annotations. E-mail: a.diazdeilarraza@ehu.es

Mikel Lersundi received his PhD from the University of the Basque Country; his dissertation performed a syntactic and semantic analysis of a Basque dictionary to extract lexical-semantic relations

between words and to build a database containing these relations. He teaches Basque language for scientific purposes at the University of the Basque Country and specializes in lexico-semantic relations, terminology, and machine translation. Email: mikel.lersundi@ehu.es

Notes

[1] This study was carried out within the framework of the following projects: IXA Group: natural language processing (GIU09/19) [UBC-EHU]; IXA Group: consolidated research groups grant 2007-2012 (IT-397-07) [Basque Government]; RICOTERM-3 (HUM2007-65966-CO2-02) [Spanish Ministry of Education]; KNOW2: Language understanding technologies for multilingual domain-oriented information access (TIN2009-14715-C04-01) [Spanish Ministry of Science and Innovation].

[2] The source of examples is indicated as follows: journal acronym, year of publication, issue number (to differentiate the various issues published during a year, sequential numbering is used). Articles were excerpted from the website of the *Gaceta Médica de Bilbao* (Bilbao Medical Journal): http://www.gacetamedicabilbao.org/web/es/

[3] RST website: http://www.sfu.ca/rst/

[4] This article uses N-N (Nucleus-Nucleus) to indicate paratactic or multi-nuclear relations with more than one nucleus and N-S (Nucleus-Satellite) to indicate hypotactic or nuclear relations with a single nucleus, whether their order is Nucleus-Satellite or Satellite-Nucleus.

[5] The website for the rhetorical structure tree graphic editing tool is http://www.wagsoft.com/RSTTool/

[6] https://ixa.si.ehu.es/Ixa/resources/Euskal_RSTTreebank

[7] See detailed discussion of nuclearity in Stede (2008b).

[8] Although this notation (2-2) does not appear in the figure, it is used to refer to a simple segment, in this case segment number 2.

[9] This hypotactic relation can be stated as 2-5 > 6-10. The unit represented by span 2-5 is the satellite of the hypotactic relation whose nucleus is represented by span 6-10. The symbol ">" represents the direction of the relation from the satellite toward the nucleus.

[10] A clarification may be necessary for readers unfamiliar with RST, given that multinuclear relations could almost be confused in some cases. For example, in Figure 2, CONJUNCTION could be confused with JOINT and LIST. The JOINT relation is the declared absence of a relation in RST literature (Taboada and Mann 2006b), because it by definition lacks constraints on both the nucleus and the satellite. Annotators need to determine the most appropriate relation before choosing JOINT instead of CONJUNCTION, LIST or SEQUENCE (Mann and Taboada 2010).

In our example, CONJUNCTION is the most plausible relation, since both nuclei have comparable elements (Mann and Taboada 2010). In the first CONJUNCTION one EDU tells us the percentage of users that come to emergency services while the other EDU reflects the percentage of how these users are considered. In the second, CONJUNCTION the comparison reflects user profiles and where users come from.

An interesting discussion about these relations can be found on the RST web page (Mann and Taboada 2010).

[11] The symbol '=' represents the connection in paratactic or multi-nuclear relations.

[12] Following Hualde and Ortiz (2003) Table 2 shows the list of gloss abbreviations for Basque examples. Note that when a gloss has multiples forms, these are not included.

[13] In contrast to RST we don't distinguish between restrictive and non-restrictive relative clauses.

[14] In this paper, subordination refers exclusively to syntactic subordination, whereas hypotactic refers to rhetorical structure. In this case, the dependent unit or satellite depends on the more important unit, the nucleus.

[15] Although the text in example 16 has been split twice (e.g. into what appears to be three pieces), the annotator has indicated that it contains two elementary units: the clause interpolated by means of the satellite unit using the gerund *erabiliz* ('utilizing') splits the nucleus into two fragments.

[16] Note that this verb is elided in the Basque text.

[17] Note that examples 11 and 12 could be segmented more deeply at the intra-sentential level and annotated with MEANS and CAUSE relations, respectively.

[18] This verb is elided in the Basque text.

[19] This verb is also elided in the Basque text.

[20] Note that the literal translation of *eta* is 'and' and not 'however' and that the verb 'express' is elided in the Basque text.

[21] Marcu uses the label SPAN.

[22] This node (4-4) annotated by A2 can be compared with another node annotated by A1 (4-4) in Table 13 given that the composition of both nodes for A1 and A2 is the same.

[23] https://ixa.si.ehu.es/Ixa/resources/Euskal_RSTTreebank

References

Afantenos, Stergos D., Pascal Denis, Philippe Muller & Laurence Danlos. 2010. Learning Recursive Segments for Discourse Parsing. Paper presented at the Seventh conference on International Language Resources and Evaluation, Paris, France, 19-21 May.

Arregi, Olatz, Klara Ceberio, Arantza Díaz-de-Ilarraza, Iakes Goenaga, Basilio Sierra & Ana Zelaia. 2010. A first machine learning approach to pronominal anaphora resolution in Basque. Paper presented at the 12th edition of the Ibero-American Conference on Artificial Intelligence, Bahía Blanca, Argentine, 1-5 November.

Arrieta, Bertol. 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean [Surface treatment of syntax by machine learning techniques: Basque words and sentences identification and its use in a coma-corrector]*. Donostia: EHU-UPV University of the Basque Country dissertation.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge Univ Pr.

Barrutieta, Guillermo, Joseba Abaitua & Josuka Díaz. 2002. An XML/RST-based approach to multilingual document generation for the web. *Procesamiento del lenguaje natural* 29, 247-253.

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics* 22(2), 249-254.

Carlson, Lynn, Daniel Marcu & Mary E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (eds.), *Current and New Directions in Discourse and Dialogue*, 85-112. Berlin: Springer.

Carlson, Lynn, Mary E. Okurowski & Daniel Marcu. 2002. *RST Discourse Treebank, LDC2002T07 [Corpus]*. Philadelphia: PA: Linguistic Data Consortium.

Ceberio, Klara, Itziar Aduriz, Arantza Díaz-de-Ilarraza & Ines Garcıa. 2009. Empirical study of the relevance of semantic information for anaphora resolution: the case of adverbial anaphora. Paper presented at the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC09), Goa, India, 5-6 November.

Cohen, Robin. 1987. Analyzing the structure of argumentative discourse. *Computational linguistics* 13(1-2), 11-24.

da Cunha, Iria & Mikel Iruskieta. 2009. La influencia del anotador y las técnicas de traducción en el desarrollo de árboles retóricos [*Annotators and translation strategies influences in rhetorical trees structures. A Spanish Basque study*]. Paper presented at the 7th Brazilian Symposium in Information and Human Language Technology (STIL), Sao Carlos, Brazil, 8-11 September.

da Cunha, Iria & Mikel Iruskieta. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies* 12(5), 563-598.

da Cunha, Iria, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes & Irene Castellón. 2010. DiSeg: Un segmentador discursivo automatico para el espanol. *Procesamiento de Lenguaje Natural* 45.

da Cunha, Iria, Juan-Manuel Torres-Moreno & Gerardo Sierra. 2011. On the Development of the RST Spanish Treebank. Paper presented at the 5th Linguistic Annotation Workshop (LAW V '11), Portland, USA, 23 June.

Grosz, Barbara J. & Candance L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12(3), 175-204.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23(1), 33-64.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive science* 3(1), 67-90.

Hovy, Eduard. 2010. Annotation: A Tutorial. *48th Annual Meeting of the Association for Computational Linguistics*.

Hualde, José I. & Jon Ortiz de Urbina. 2003. *A grammar of Basque*. Berlin: Walter de Gruyter.

Iruskieta, Mikel, Arantza Díaz-de-Ilarraza & Mikel Lersundi. 2008. Análisis de los marcadores del discurso para el euskera: denominación, clases, relaciones semánticas y tipos de ambigüedad [*A study of discourse markers in Basque: denomination, classes, semantic relations and ambiguity*]. Paper presented at the 26th AESLA Conference, Almeria, Spain.

Iruskieta, Mikel, Arantza Díaz-de-Ilarraza & Mikel Lersundi. 2009. Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso [*Correlations between rhetorical relations and discourse markers*]. Paper presented at the 27th AESLA Conference, Ciudad Real, Spain.

Iruskieta, Mikel, Arantza Díaz-de-Ilarraza & Mikel Lersundi. 2011a. Bases para la implementación de un segmentador discursivo para el euskera [*Bases for an Implementation of a Discourse Parser for Basque*]. Paper presented at the Workshop A RST e os Estudos do Texto, Mato Grosso, Brazil, 24-26 October.

Iruskieta, Mikel, Arantza Díaz-de-Ilarraza & Mikel Lersundi. 2011b. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera [*Discourse unit and rhetorical relations: a study about discourse units in the annotation of a corpus in Basque*]. *Procesamiento de Lenguaje Natural* 47, 137-144.

Landis, Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159-174.

Litman, Diane J. & James F. Allen. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science* 11(2), 163-200.
Mann, Willian C. & Maite Taboada. 2010. RST web-site. http://www.sfu.ca/rst/ (24/04/2012).
Mann, Willian C. & Sandra A. Thompson. 1987a. Antithesis: A study in clause combining and discourse structure. In R. Steele & T. Threadgold (eds.), *Language Topics: Essays in honor of Michael Halliday*, 359-381. Amsterdam: Benjamins.
Mann, Willian C. & Sandra A. Thompson. 1987b. Rhetorical Structure Theory: A Theory of Text Organization. *Text* 8(3), 243-281.
Mann, Willian C. & 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.
Marcu, Daniel. 1999. Discourse trees are good indicators of importance in text. In Inderjeet Mani & Mark Maybury (eds.), *Advances in Automatic Text Summarization*, 123-136. Cambridge.
Marcu, Daniel. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics* 26(3), 395-448.
Mazeiro, Erick G. & Thiago A. S. Pardo. 2009. Metodologia de avaliação automática de estruturas retóricas [*Methodology for automatic evaluation of rhetorical structures*]. Paper presented at the 7th Brazilian Symposium in Information and Human Language Technology (STIL), São Carlos, Brazil, 8-11 September.
Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi & Bonnie L. Webber. 2004. Annotating discourse connectives and their arguments. Paper presented at the HLT/NAACL Workshop on Frontiers in Corpus Annotation, Boston, USA.
Moser, Megan & Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational linguistics* 22(3), 409-419.
O'Donnell, Michael. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. Paper presented at the First International Conference on Natural Language Generation INLG '00, Mitzpe Ramon (Israel), 12-16 June.
Pardo, Thiago A. S. 2005. *Métodos para análise discursiva automática [Methods for automatic discourse analysis]*. Sao Carlos: Instituto de Ciências Matemáticas e de Computação dissertation.
Pardo, Thiago A. S. & 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*]. *Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional* NILC-TR-04-03.
Pardo, Thiago A. S. & Eloize R. M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente [*Rhetalho: a rhetorically annotated reference corpus*]. Paper presented at the Anais do V Encontro de Corpora, São Carlos, Brazil, 24-25 November.
Passonneau, Rebecca J. & Diane J. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. Paper presented at the 31st annual meeting on Association for Computational Linguistics, Ohio, USA.
Polanyi, Livia, Christopher Culy, Martin van den Berg, G. L. Thione & David Ahn. 2004. A rule based approach to discourse parsing. Paper presented at the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004, 30-1 April-May.
Soricut, R. & Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. Paper presented at the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.
Stede, Manfred. 2004. The Potsdam commentary corpus. Paper presented at the 2004 ACL Workshop on Discourse Annotation, Barcelona, Spain, 25-26 July.
Stede, Manfred. 2008a. Disambiguating rhetorical structure. *Research on Language and Computation* 6(3), 311-332.
Stede, Manfred. 2008b. RST revisited: Disentangling nuclearity. In Cathrine Fabricius-Hansen & Wiebke Ramm (eds.), *Subordination' versus 'Coordination' in Sentence and Text*, 33-57. Amsterdam and Philadelphia: John Benjamins.
Taboada, Maite & Willian C. Mann. 2006a. Applications of rhetorical structure theory. *Discourse studies* 8(4), 567-588.
Taboada, Maite & Willian C. Mann. 2006b. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies* 8(3), 423-459.
Taboada, Maite & Jan Renkema. 2011. Discourse Relations Reference Corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.
Tofiloski, Milan, Julian Brooke & Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. Paper presented at the 47th Annual Meeting of the Association for Computational Linguistics, Suntec, Singapore, 2-7 August.
Webber, Bonnie L., M. Stone, Aravind Joshi & Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics* 29(4), 545-587.

8

Comparing rhetorical structures in different languages: The influence of translation strategies

Article

Discourse Studies
12(5) 563–598

DOI: 10.1177/1461445610371054
http://dis.sagepub.com
SAGE

Comparing rhetorical structures in different languages: The influence of translation strategies

Iria da Cunha
Université d'Avignon et des Pays de Vaucluse, France and Universitat Pompeu Fabra, Spain

Mikel Iruskieta
University of the Basque Country (UPV/EHU), Spain

Abstract
The study we report in this article addresses the results of comparing the rhetorical trees from two different languages carried out by two annotators starting from the Rhetorical Structure Theory (RST). Furthermore, we investigate the methodology for a suitable evaluation, both quantitative and qualitative, of these trees. Our corpus contains abstracts of medical research articles written both in Spanish and Basque, and extracted from *Gaceta Médica de Bilbao* ('Medical Journal of Bilbao'). The results demonstrate that almost half of the annotator disagreement is due to the use of translation strategies that notably affect rhetorical structures.

Keywords
annotation, discourse analysis, evaluation, medical research articles, rhetorical relations, Rhetorical Structure Theory, textual corpus, translation strategies

1. Introduction
Writing abstracts of research articles both in a lingua franca (English, French, etc.) and in local languages (Catalan, Spanish, Basque, etc.) is nowadays usual among the scientific community. In fact, it has become a requisite for the publication in some scientific journals. As a result, it is possible to obtain bilingual corpora to investigate how the

Corresponding author:
Iria da Cunha, Université d'Avignon et des Pays de Vaucluse, Laboratoire Informatique d'Avignon, 339, chemin des Meinajaries, 84911 Avignon, France and Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona, Spain.
Email: iria.dacunha@upf.edu

rhetorical structures of abstracts are shown in each language and how translation strategies affect discourse structure. Some authors have carried out studies about the evaluation of rhetorical structure annotation (Carlson et al., 2001; Marcu, 2000a; Marcu et al., 1999) and about the comparison of rhetorical structures in different languages: Chinese–English (Cui, 1986; Kong, 1998; Ramsay, 2000, 2001), English–Dutch (Abelen et al., 1993), English–French (Delin et al., 1996; Salkie and Oates, 1999), Portuguese–French–English (Scott et al., 1998) and English–Japanese (Marcu et al., 2000), among others. However, to our knowledge, no studies exist on the way that translation strategies affect the process of rhetorical annotation and on the evaluation of annotator agreement.

In this work, we use Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) since it is a language-independent theory. RST is a descriptive theory for textual organization that has been proven to be very useful in describing a document by characterizing its structure with relations maintained among its discursive or rhetorical elements (e.g. Circumstance, Elaboration, Motivation, Evidence, Justification, Cause, Purpose, Antithesis, Condition, List, Contrast, etc.). As Taboada and Mann (2006) state: 'RST addresses text organization by means of relations that hold between parts of a text. It explains coherence by postulating a hierarchical, connected structure of texts, in which every part of a text has a role, a function to play, with respect to other parts in the text.' RST determines a set of relations among the discursive units of texts. As a rule, one of the units is more essential to the speaker's purpose (nucleus), while the other one (satellite) provides some rhetorical information about it. This is the more usual structural model between these two units (almost always adjacent units, although there are some exceptions). These relations are named 'nuclear' relations (e.g. Circumstance, Elaboration, Motivation, Evidence, etc.). In the case of relations with more than one central unit with regard to the author's purposes, the relation is named 'multinuclear' and a coordinated relation is established (e.g. List, Joint, Contrast, etc.). For a more detailed explanation of RST, we recommend reading the article by Mann and Thompson (1988) or the RST web site by Mann (2005).

RST is used to inquire into several theoretical and applied subjects explained in Taboada and Mann (2005) as, for example, automatic generation of texts, automatic summarization, textual analysis, automatic translation, writing teaching, acquisition of discursive knowledge, spoken discourse analysis, information extraction, etc. Some relevant works on these subjects are, among others, Bouayad-Agha (2000), Burstein and Marcu (2003), da Cunha (2008), da Cunha et al. (2007), Ghorbel et al. (2001), Haouam and Marir (2003) and Marcu (2000a). In addition, some rhetorical parsers in different languages are also based on this theory: Sumita et al. (1992) in Japanese, Marcu (1998) in English, and Pardo and Nunes (2008) and Pardo et al. (2004) in Brazilian Portuguese. There is a current project to develop this parser for the Spanish language (da Cunha and Torres-Moreno, 2010). A rhetorical parser is a system that automatically analyzes a text, giving as output the rhetorical tree of this text in terms of RST. This kind of parser has three stages: rhetorical segmentation, determination of RST relations and development of rhetorical trees. They are usually based on lexical-syntactic rules and statistical techniques.

However, though widely used, some objections have been made to RST. Stede (2008), for example, criticizes its ambiguity, since many assumptions that annotators carry out cannot be made explicit in a single tree. The difficulty of obtaining the same rhetorical tree of a text from different annotators would prove this subjectivity:

> An RST-style analysis of a text, on the other hand, cuts 'vertically': It tries to capture the essence of coherence within a single representation structure, making a series of quite different simplifications along the way. We do not doubt that this can be an insightful instrument for studying text – RST has been quite successful for a variety of purposes. But there are inherent limitations on the explanatory power when information from different realms is conflated in a single tree structure: On the one hand, one cannot do full justice to the separate realms; on the other hand, the single tree structure becomes ambiguous, because when crafting it, many underlying assumptions cannot be made explicit. (Stede, 2008: 329)

All the considerations taken into account until now lead us to formulate the following interesting questions:

- Is it possible to compare the rhetorical structures of a parallel corpus of medical texts in two very different languages such as a Romance language (Spanish) and a Non-Indo-European language (Basque) by means of the same theory? Do these texts share a similar superstructure?
- Taking into account the difficulty of two annotators carrying out the same rhetorical analysis with RST relations, how do translation strategies affect the agreement on the rhetorical structure of parallel texts? Which linguistic differences exist in both rhetorical structures?
- Which is the best evaluation method in order to determine the factors affecting the evaluation of rhetorical structure (translation strategies or linguistic differences; theoretical abstraction level or ambiguity of the rhetorical structure)?

In this article we aim to answer these questions. With this intention, an experiment has been designed. First, the corpus was annotated with rhetorical relations (one author annotated the Basque corpus and the other annotated the Spanish one). This corpus contains 20 abstracts in Spanish and Basque, included in medical research articles from the *Gaceta Médica de Bilbao*[1] ('Medical Journal of Bilbao'). Afterwards, both annotations were compared and the differences among them were observed. The methodology used in this experiment is explained in section 2. In section 3, we give the details of the results of the quantitative and qualitative evaluations on spans, nuclearity and rhetorical relations. Conclusions are presented in section 4.

2. Methodology

The methodology of our research included several phases. First, a corpus of analysis was built. Second, departure criteria with regard to the segmentation of the text into units and to the specific relations used were defined. Third, the corpus texts were labeled by the annotators (one in Spanish and one in Basque). Fourth, quantitative analysis was carried out. Fifth, qualitative analysis was performed.

2.1. Corpus

Nowadays, no parallel Spanish–Basque corpora are available for research purposes. Research groups have to develop their own corpus in order to carry out contrastive

research in these two languages. For this reason, we had to create a specific corpus to perform our analysis. There are no previous studies comparing rhetorical structures in Spanish and Basque. As mentioned, our corpus contains 20 abstracts in Spanish and Basque included in medical research articles from the *Gaceta Médica de Bilbao* written by medical specialists between the years 2000 and 2008.

The first reason to choose this corpus was that this journal requests that authors submit the articles in Spanish and the corresponding abstracts in Spanish, Basque and English. As most of the authors of the texts of our corpus are Basque and a relevant portion of the Basque population is bilingual, we assume that they themselves wrote both the abstracts in Spanish and Basque. Nevertheless, in some cases, the author may have asked for some help to write the Basque abstract. We think this fact is not really relevant, because the journal gives the authors very detailed guidelines about the information that they have to include in their abstracts (in the three mentioned languages). Authors are asked to use in their abstracts the IMRD structure (Swales, 1990): Introduction, Methods, Results and Discussion:

> The summary must contain approximately 150 words and it must include:
> a) the purpose of the study,
> b) the used procedures and the principal findings,
> c) the most relevant conclusions, with emphasis on what is new or relevant in the article.[2]

We think these two facts (bilingualism and journal guidelines) guarantee that both abstracts (Spanish and Basque) include the same information and a similar structure.

The second reason to choose this corpus is to analyze the relations among macrostructures and genres and, in this way, to highlight a rather open question of RST. As Taboada and Mann (2006) state: 'A more exhaustive study of different genres would throw light on the relationship between macrostructures or genres and RST structures.' We have selected a specialized corpus that contains medical texts with a very specific genre: the research article. In the future, we plan to analyze a general corpus to compare it with this specialized corpus.

Appendix Table 1 shows the information of the corpus texts (title, author[s] and year of publication).

2.2. Departure criteria

In order to avoid circularities as much as possible, we first define what is an EDU (Elementary Discourse Unit) in an abstract way and, second, we segment all the text only focusing on syntactic clues (see section 2.2.1.) before carrying out the rhetorical analysis.

2.2.1. EDU segmentation. Mann and Thompson (1988) proposed a definition of discourse unit based on a theory-neutral classification. Their motivation was to describe a theoretical frame for RST. To this end, they proposed an abstract definition and they escaped from a circular definition:

> Unit size is arbitrary but the division of the text into units should be based on some theory-neutral classification. That is, for interesting results, the units should have independent functional integrity. In our analyses, units are essentially clauses, except that clausal subjects and complements and restrictive relative clauses are considered parts of their host clause units rather than separate units. (Mann and Thompson, 1988: 6)

Although Marcu (1999) uses RST as well, his definition of discourse unit has a different motivation: the conformation of a corpus of tagged documents for the research community. Thus, the annotation should offer all the possible information. As he states:

> One (probably) uncontroversial choice would be to take sentences as the elementary units of discourse. Unfortunately, if we do so, we leave lots of rhetorical information outside the scope of our analysis. (Marcu, 1999: 9)

Marcu's definition of unit can be controversial in some aspects because of its circular nature, but for Marcu this is a secondary question given that it does not interfere with his main motivation.

Our goal is far from both Mann and Thompson's (1988) and Marcu's (1999) proposals because, first, we want to compare the rhetorical structure of translations at a propositional level and, second, we want to analyze some problems that appear during the annotation process. Therefore, in this work, we do not consider it necessary to carry out such a detailed analysis as Marcu.

With regard to EDU segmentation, we follow more or less the most common set of guidelines for segmenting text in RST. Carlson and Marcu (2001) departed from them in some aspects and we have revised some questions from their manual. Some specifications were made so that we would be able to clearly differentiate syntactic and discursive levels. In this work, we consider that EDUs must include a finite verb (that is, they have to constitute a sentence or a clause) and must show, strictly speaking, a rhetorical relation. These established specifications are the following ones:[3]

a) In Carlson and Marcu (2001), complements of attribution verbs (speech acts and other cognitive acts) are treated as EDUs, as example 1a shows:[4]

1a. [Bush indicated] [there might be 'room for flexibility' in a bill] [. . .]

In contrast, our approach does not consider these complements of attribution verbs as EDUs, and we would segment the same passage as example 1b shows:

1b. [Bush indicated there might be 'room for flexibility' in a bill] [. . .]

The clause 'there might be ''room for flexibility'' in a bill' constitutes a direct object (from a traditional grammar-oriented approach) or an actant II (from a dependency grammar-oriented approach) of the verb 'to indicate' and, because of that, we consider it only at this level (syntactic).

We do not consider the Attribution relation for three types of reasons: a) a definitional reason: it does not make explicit any kind of writer's intention, so Attribution does not

have the same status as other RST relations (Stede, 2008); b) a language level reason: it can be identified only by syntax rules (Skadhauge and Hardt, 2005); and c) a procedural reason: it implies circularity in EDU definition. As Stede (2008: 316) states:

> *Attribution* thus does not have the same status as, say, relations of causality or contrast: The relationship between an event of saying and the specific contents of that saying is different from a coherence relation linking two complete propositions.

b) Carlson and Marcu (2001) specify that the clauses that depend to 'so that their clients can' are treated as various EDUs and these are considered as satellites in a Purpose relation. In turn, the satellite constitutes a multinuclear List of coordinated clauses, as we can see in example 2a:

2a. [Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] [so that their clients can find temporary living quarters,] [buy food,] [replace lost clothing,] [repair broken water heaters,] [and replaster walls.]

In contrast, we would treat all these clauses as a single EDU:

2b. [Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] [so that their clients can find temporary living quarters, buy food, replace lost clothing, repair broken water heaters, and replaster walls.]

c) In Carlson and Marcu (2001), relative clauses, nominal postmodifiers and clauses that break up other legitimate EDUs are treated as embedded discourse units, while we do not consider these units as such. Several examples follow:

Relative clauses:

3a. [A separate inquiry by Chemical cleared Mr. Edelson of allegations] [*that* he had been lavishly entertained by a New York money broker.]

3b. [A separate inquiry by Chemical cleared Mr. Edelson of allegations *that* he had been lavishly entertained by a New York money broker.]

Nominal postmodifiers with non-finite clause:

4a. [The results underscore Sears's difficulties] [*in implementing* the 'everyday low pricing' strategy] [that it adopted in March, as part of a broad attempt] [*to revive* its retailing business.]

4b. [The results underscore Sears's difficulties *in implementing* the 'everyday low pricing' strategy that it adopted in March, as part of a broad attempt *to revive* its retailing business.]

Appositives:

5a. [The fact] [*that* this happened two years ago] [and there was a recovery] [gives people some comfort] [*that* this won't be a problem.]

5b. [The fact *that* this happened two years ago and there was a recovery gives people some comfort *that* this won't be a problem.]

Parentheticals:

6a. [The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles] [*($US693.4 billion)*] [and expenditures of 489.9 billion rubles] [*($US790.2 billion)*.]

6b. [The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles *($US693.4 billion)* and expenditures of 489.9 billion rubles *($US790.2 billion)*.]

In this work, we only segment units appearing in parentheses when they clearly constitute an EDU, or an element maintaining some discourse relation with another element and containing a finite verb.

Coordinated clauses in embedded units:

7a. [She signed up,] [starting as an 'inside' adjuster,] [who settles minor claims] [*and* does a lot of work by phone.]

7b. [She signed up,] [starting as an 'inside' adjuster, who settles minor claims *and* does a lot of work by phone.]

d) In Carlson and Marcu (2001), phrases that begin with a strong discourse marker, such as *because*, *in spite of*, *as a result of*, *according to*, are treated as EDUs, as examples 8a and 9a show:

8a. [But some big brokerage firms said] [they don't expect major problems] [*as a result of* margin calls.]

9a. [Today, no one gets in or out of the restricted area] [*without* De Beers's stingy approval.]

In this work, we consider that sentences starting by these markers are EDUs only if a finite verb also exists. Therefore, we would segment the previous examples as follows:

8b. [But some big brokerage firms said they don't expect major problem *as a result of* margin calls.]

9b. [Today, no one gets in or out of the restricted area *without* De Beers's stingy approval.]

e) Carlson and Marcu (2001) establish several criteria to determine EDUs' boundaries. In this work, we only use these criteria if the marked EDU contains a finite verb. Some examples are offered below:

Parenthesis:

10a. [If the government can stick with them,] [it will be able to halve this year's 120 billion ruble] [*(US$193 billion)*] [deficit.][5]

10b. [If the government can stick with them,] [it will be able to halve this year's 120 billion ruble *(US$193 billion)* deficit.]

Dashes:

11a. [This will require us to define] [– *and redefine* –] [what is 'necessary' or 'appropriate' care.]

11b. [This will require us to define – *and redefine* – what is 'necessary' or 'appropriate' care.]

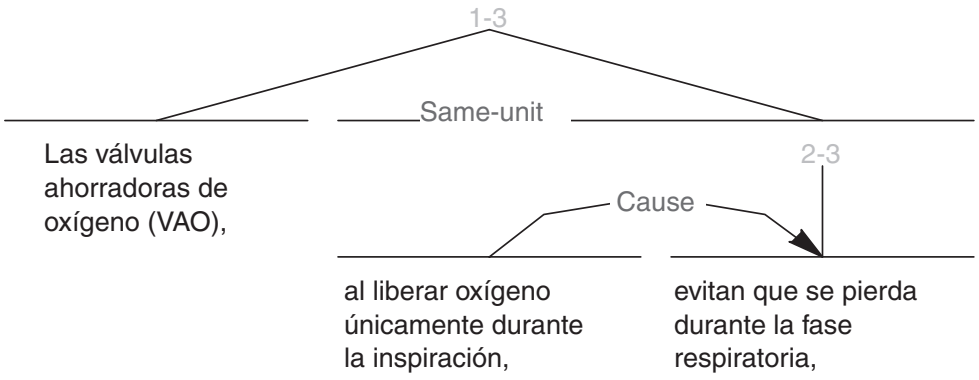


Figure 1. Rhetorical tree showing a Same-unit relation

With regard to the utilization of other punctuation marks (comma, full-stop, semicolon, etc.) like boundary marks, we agree with Carlson and Marcu (2001: 30):

> Commas and periods are not independent justification for an EDU boundary. If a unit is a legitimate EDU and it ends with a comma or period, the punctuation is included as part of that EDU.

Finally, it is important to highlight that an EDU can be truncated by another one (that is, it can include another EDU). If this occurs in our work, as in Carlson and Marcu (2001), the two fragments of the first EDU are segmented and they are linked later with a Same-unit relation, which is not a relation but a convention. For example, Figure 1 would be labeled as follows:

12. *[Las válvulas ahorradoras de oxígeno (VAO),] [al liberar oxígeno únicamente durante la inspiración,] [evitan que se pierda durante la fase respiratoria,]* […]

ENGLISH TRANSLATION: [Oxygen Conserving Valves (OCV),] [because of their release of oxygen only during inhalation,] [avoid losing oxygen during the breathing phase,] […]

2.2.2. Rhetorical relations. Concerning the detection of rhetorical relations and nuclearity (that is, with regard to the decision of considering a segment as nucleus or satellite), the following tasks were carried out:

a) The list of rhetorical relations of the RST was determined. There are various classifications of rhetorical relations: the classic one by Mann and Thompson of 24 relations (Mann and Thompson, 1988), the extended one by Mann and Thompson of 30 relations (Mann, 2005) and Marcu's classification of 136 relations (Carlson et al., 2001), among others. The extended classification (Mann, 2005) was chosen for the annotation of the parallel corpus. As Marcu et al. (1999: 55) point out, reduction in the relations' taxonomy does not have a significant impact on annotators' agreement:

> The results [. . .] show that a significant reduction in the size of the taxonomy of relations may not have a significant impact on agreement ($k_{\gamma\gamma}$ is only about 4% higher than k_{γ}). This suggests that choosing one relation from a set of rhetorically similar relations produces some, but not too much, confusion.

b) We looked for a real representative example of each relation and nuclei and satellites were marked. Examples are taken from the corpus used in da Cunha (2008), containing Spanish medical articles that were extracted from the journal *Medicina Clínica* ('Clinical Medicine').[6] Once the Spanish examples were selected, they were translated into Basque and their nuclei and satellites were marked.

Appendix Table 2 includes the list of relations used in this work, specifying if they are multinuclear relations (N-N) or nuclear relations (N-S). For each relation, an example in Spanish and Basque is provided, where its nuclei (N) and satellites (S) are marked.

2.3. Rhetorical annotation

Once departure criteria were established, both annotators labeled the 20 texts of the corpus with RST relations (one in Spanish [A1] and another one in Basque [A2]). The annotation was divided into two main stages: EDU segmentation and rhetorical analysis.

2.3.1. EDU segmentation. In this stage, each annotator segmented the 20 abstracts of the corpus into EDUs by using the RSTTool (O'Donnell, 2000).[7] This task was done separately and without any contact among annotators.

Once the data on the agreement of the performed segmentations by both annotators was collected, we carried out a small discussion in order to homogenize the segmentation of Spanish and Basque abstracts. This homogenization was carried out in order to minimize the noise that could arise from a different segmentation. By these means, we aimed at obtaining, first, a more detailed quantification of the nuclearity and of the relations of rhetorical trees and, secondly, an evaluation of the factors affecting the structure. This comparison was performed manually (measuring precision and recall), due to the current lack of automatic tools comparing rhetorical trees in different languages. Mazeiro and Pardo (2009) have developed the RSTeval tool, which does compare rhetorical trees but in the same language, so it could not be used in this study.

Since our comparison had to be manually done, we considered it appropriate to carry out this task of EDU homogenization so that annotators could label the same segments, establish relations among them, build the rhetorical trees and, finally, carry out the comparison among them in a more accurate way.

2.3.2. Rhetorical analysis. In this stage, each annotator labeled the homogenized segmentation of the studied abstracts, marking rhetorical relations among EDUs and determining which of these EDUs were nuclei or satellites. To this end, the RSTTool and the extended classification of rhetorical relations were used.

2.4. Quantitative analysis

After the annotation, a quantitative analysis about the two aspects detailed in the previous section was performed.

2.4.1. EDU segmentation. The contrast between the EDU segmentation of both annotators was carried out by evaluating precision and recall. To measure precision, we observed the coincidence between the selected EDUs by A2 and the selected EDUs by A1. To

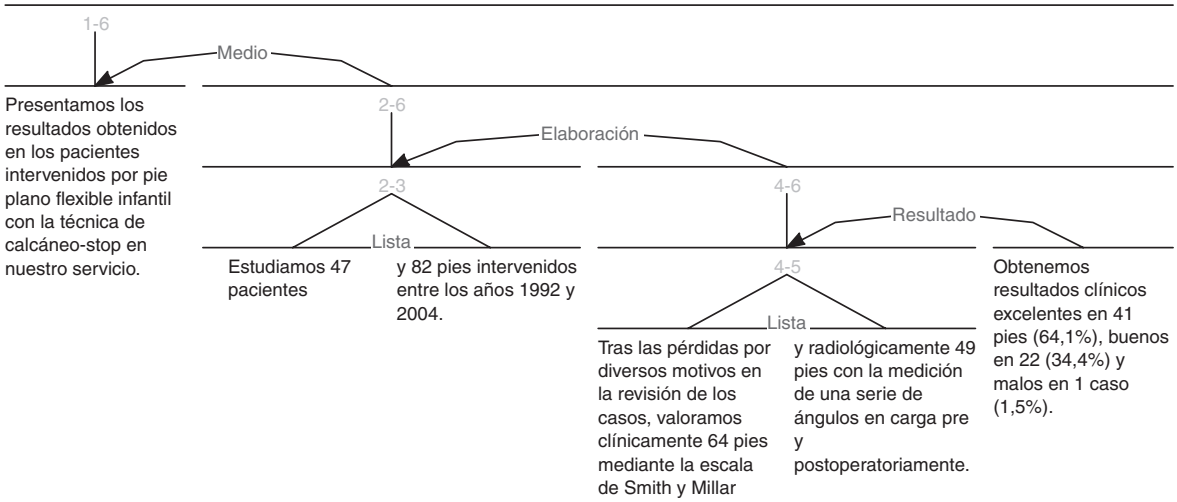


Figure 2. Rhetorical tree in Spanish by A1

measure recall, we compared the number of detected EDUs by A2 with the number of detected EDUs by A1. This analysis was carried out, on the one hand, for each individual text and, on the other hand, for the set of texts of our corpus.

2.4.2. Rhetorical analysis. To quantify the agreement between the rhetorical analyses by both annotators, we used Marcu's (2000b) method. Specifically, we obtained data concerning detected spans (i.e. sets of related EDUs), nuclearity and rhetorical relations.

To compare both rhetorical analyses, precision and recall were measured again. To measure precision, we counted the number of detected spans, nuclei and satellites, and rhetorical relations marked by A2 coinciding with the ones selected by A1. To measure recall, we counted the total number of the same elements detected by A2, with regard to the total number detected by A1. Once again, this analysis was performed for each text and for the texts of our corpus taken together. For instance, Figure 2 shows a rhetorical tree fragment in Spanish carried out by A1, whereas Figure 3 shows the rhetorical tree of the same passage in Basque, carried out by A2. The English abstract passage of the author that corresponds with this text is provided in here, in order to make the example more understandable to the reader:[8]

English translation:

Unit 1: [We report our experience and the results obtained with surgical treatment of infantile flexible flan foot using the calcaneus-stop technique.]

Unit 2: [From 1992 through 2004, 47 patients]

Unit 3: [and 82 feet were studied.]

Unit 4: [After our revision, 64 feet were evaluated clinically using the Smith and Millar scale]

Unit 5: [and 49 feet were evaluated radiologically by several preoperative and postoperative radiological variables.]

Unit 6: [The clinical results were excellent in 41 feet (64.1%), good in 22 feet (34.4%) and bad in only case (1.5%).]

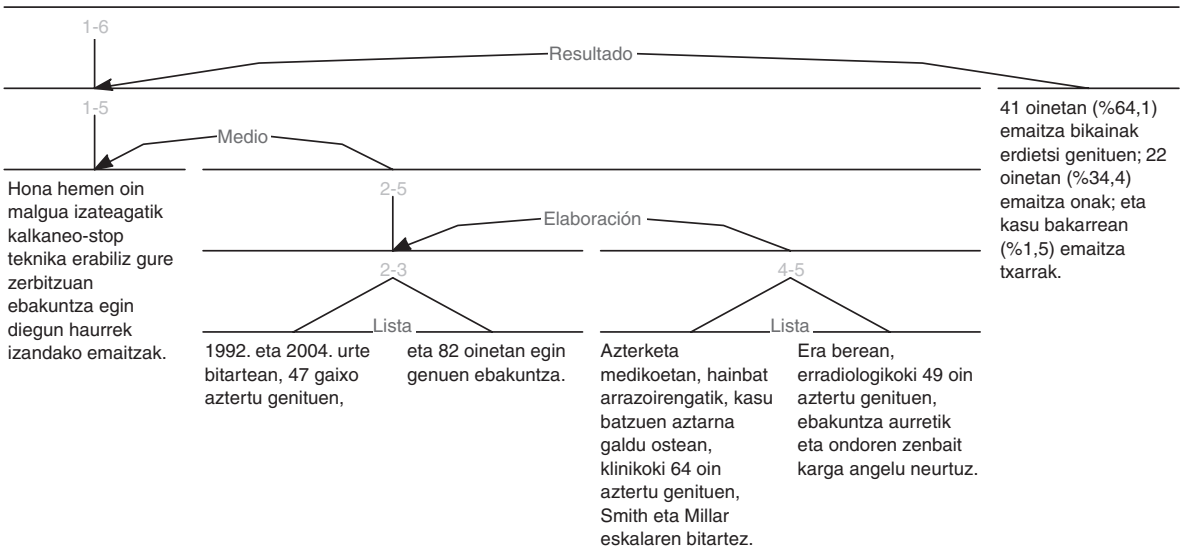


Figure 3. Rhetorical tree in Basque by A2

Table 1 below exemplifies Marcu's (2000b) evaluation methodology. It includes a comparison of detected spans, nuclearity and relations annotated by A1 and A2. We have used the NUCLEUS[9] label to refer to the nuclei of nuclear relations, and the relation name (e.g. Result, Elaboration, Means, List, etc.) to refer either to the satellites of nuclear relations or to the nuclei of multinuclear relations. It is necessary to take into account that, since we homogenized the EDUs in the segmentation stage (see section 2.3.1.), the detected EDUs by A1 and A2 always coincided. In Table 1 we have indicated in grey the differences between both annotators, where nuclei are denoted by 'N' and satellites by 'S'.

Table 1. Quantitative evaluation using Marcu's (2000b) method

	EDU		Span		Nuclearity		Relation	
Element	A1	A2	A1	A2	A1	A2	A1	A2
1–1	X	X	X	X	N	N	NUCLEUS	NUCLEUS
2–2	X	X	X	X	N	N	LIST	LIST
3–3	X	X	X	X	N	N	LIST	LIST
4–4	X	X	X	X	N	N	LIST	LIST
5–5	X	X	X	X	N	N	LIST	LIST
6–6	X	X	X	X	S	S	RESULT	RESULT
4–5	–	–	X	X	N	S	NUCLEUS	ELABORATION
4–6	–	–	X	–	S	–	ELABORATION	–
2–3	–	–	X	X	N	N	NUCLEUS	NUCLEUS
2–6	–	–	X	–	S	–	MEANS	
2–5	–	–	–	X	–	S	–	MEANS
1–5	–	–	–	X	–	N	–	NUCLEUS

Table 2. Quantitative evaluation results of rhetorical trees showed in Figures 2 and 3

	Recall	Precision
Spans	100%	80%
Nuclearity	100%	70%
Relations	100%	70%

After the data were formalized with this method, we measured precision and recall, in the way explained above. Table 2 shows the results of this evaluation. The three factors obtain 100 percent of recall, whereas precision oscillates between 80 percent (spans) and 70 percent (nuclearity and rhetorical relations).

2.5. Qualitative analysis

As for qualitative analysis, we also focused on questions concerning EDU segmentation and rhetorical analysis.

2.5.1. EDU segmentation. After we quantified the differences of EDU segmentation by both annotators, we observed the specific cases on which they differed and we investigated the possible reasons for disagreement.

We observed that, when homogenizing EDUs, some aspects contradicted the established guidelines of segmentation. This is due to the fact that translation strategies also affect segmentation. For instance, some passages are considered as a single EDU in Spanish, but they have been segmented into two units in order to carry out the homogenization:

13a. [*Se realiza el estudio de la proteína 14–3-3, que resulta ser positivo*.]
ENGLISH TRANSLATION: [The study of 14–3-3 protein is carried out, which obtains positive results.]
13b. [*14–3-3 proteinaren azterketa egin zaio,] [eta emaitza positiboak lortu dira*.]
ENGLISH TRANSLATION: [The study of 14–3-3 protein is carried out,] [and its results are positive.]

Example 13a above shows that A1 annotated the Spanish passage as a single EDU, since relative clauses are not considered as EDUs. However, in example 13b, we observe that in Basque this relative clause was translated like a main sentence, related to the previous one by means of a discourse marker, the coordinative conjunction *eta* ('and'). In order to homogenize the segments, we decided to divide the Spanish EDU into two EDUs, as follows:

13c. *[Se realiza el estudio de la proteína 14–3-3,] [que resulta ser positivo.]*
ENGLISH TRANSLATION: [The study of 14–3-3 protein is carried out,] [which obtains positive results.]

Table 3. Qualitative partial evaluation of spans and nuclearity[a]

Element		Span		Nuclearity	
A1	A2	A1	A2	A1	A2
4–5	4–5	X	X	S	S
2–3	2–3	X	X	N	N
2–6	2–5	X	X	S	S
1–6	1–5	X	X	N	N
4–6	1–6	–	X	S	S

[a]The nuclei and the satellites are denoted by N and S, respectively.

Table 4. Qualitative partial evaluation of relations

Annotated relations	
A1	A2
Elaboration	Elaboration
List	List
Means	Means
List	List
Result	Result

Both annotators marked the same relation for this passage: the Result relation. This is due to the fact that there is the verb 'result' into the second EDU, and it produces more effect than the syntactic structure or the discourse marker. Probably, if there was another verb, the Elaboration relation would be considered in Spanish because of the relative clause, and the List relation would be considered in Basque because of the conjunction.

2.5.2. Rhetorical analysis. Though the evaluation method of Marcu (2000b) exemplified in section 2.4.2 is considered to be valid, the method only considers the absolute agreement in all factors. Thus, a disagreement on the segmentation or a disagreement on the lower spans will affect significantly the agreement on the upper rhetorical relations of a tree. For example, if we follow Marcu's (2000b) method, disagreement with regard to spans, nuclearity and relations is observed. However, the five relations that were marked by both annotators coincide. In fact, there are differences concerning the detected nodes, but not with regard to the detected relations. We consider it necessary to also carry out this type of approach, more optimistic in a certain way and that we call 'qualitative partial evaluation', because we believe this approach to be necessary in order to detect and analyze the linguistic differences in rhetorical structure that are originated by translation strategies. Tables 3 and 4 include the data of this evaluation, concerning, in the first place, spans and nuclearity and, in the second place, relations.[10]

Table 5. Qualitative partial evaluation results of rhetorical trees showed in Figures 2 and 3

	Recall	Precision
Spans	100%	80%
Nuclearity	100%	100%
Relations	100%	100%

Table 5 shows the qualitative partial evaluation results of the example. We notice that precision and recall are 100 percent in all cases, except for precision in spans, which is 80 percent.

Since we could obtain quantitative results concerning spans and nuclearity with Marcu's (2000b) method, we only focused on the qualitative partial evaluation of rhetorical relations. We think this qualitative evaluation is an effective way to detect the linguistic differences affecting rhetorical structure.

In the qualitative partial evaluation we systematically analyzed the causes of the disagreement between annotators. On the one hand, we observed the phenomena that could cause differences concerning the annotation agreement, mentioned by Mann and Thompson (1988): ambiguity of text structure, simultaneous analyses and analytic mistakes, among others. On the other hand, we analyzed the phenomenon reflected in Marcu et al. (2000: 10), consisting of changing the type of rhetorical relation when translating:

> Hence, the mappings in (4) provide an explicit representation of the way information is reordered and re-packaged when translated from Japanese into English. However, when translating text, it is also the case that the rhetorical rendering changes. What is realized in Japanese using a CONTRAST relation can be realized in English using, for example, a COMPARISON or a CONCESSION relation.

In this way, we detected the possible causes of discrepancies among annotators and the influence that translation strategies have on rhetorical structure (as explained in section 3.2.).

In order to count all the relations, we decided to consider each nuclear relation as one relation, while we considered multinuclear relations as binary ones. For example, a List relation with four nuclei is represented by joining its nuclei in a binary way, obtaining three multinuclear relations, each one with two nuclei. Figures 4 and 5 show respectively the Same-level annotation and the binary annotation of this List relation.

By these means, apart from correctly counting multinuclear relations, we could compare, for example, a) three units or spans of a List relation with three nuclei (by A1) with b) a List relation with two nuclei and one Elaboration relation (by A2). If we had not done it in that way, we would not have been able to compare a List relation by A1 with a List relation and an Elaboration relation by A2, and the evaluation could have lost precision. Moreover, it would not be correct to count as relations all the nuclear elements of a List relation, since multinuclear relations would then be more relevant than the others in the qualitative partial evaluation.

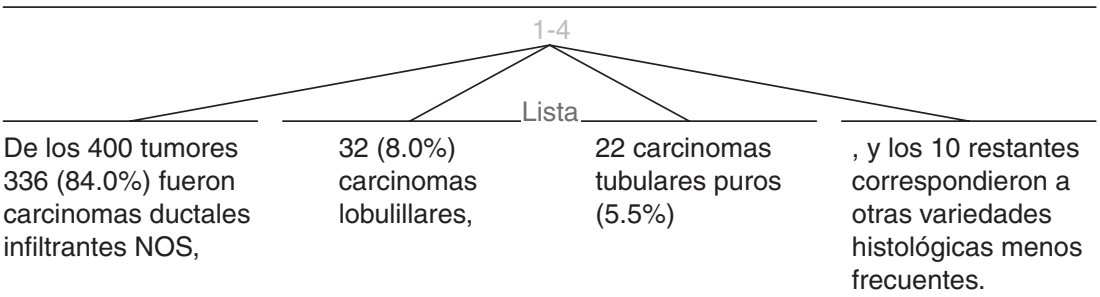


Figure 4. Same-level annotation of List relation

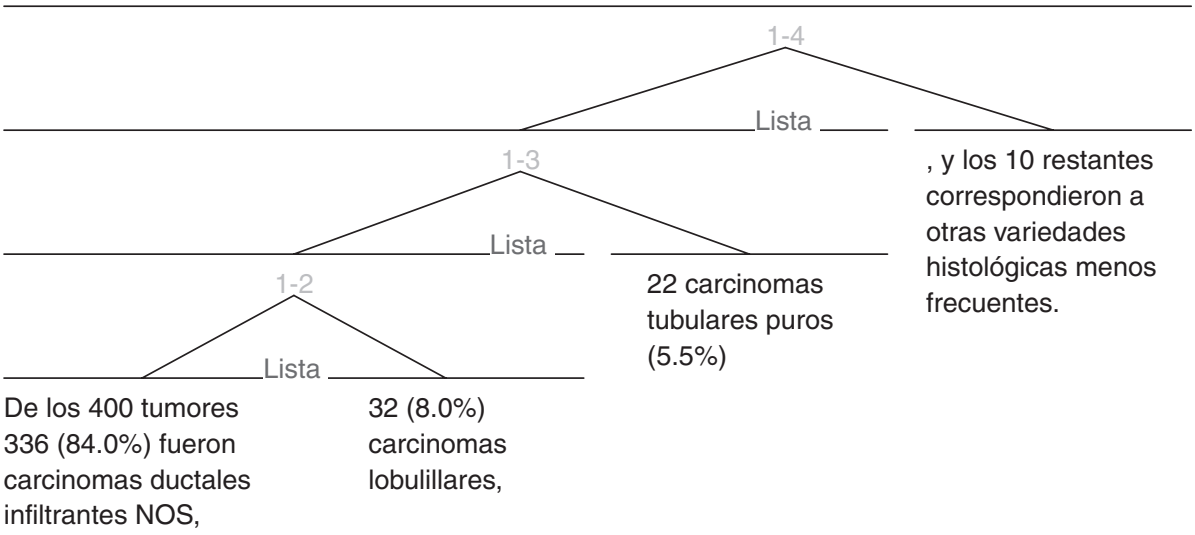


Figure 5. Binary interpretation of List relation

3. Results

In the previous sections the methodology of our experiment was presented. In this section we present segmentation and nucleus-satellite issues, with their corresponding results of agreement, and a discussion of the used translation strategies.

3.1. Segmentation issues

The number of segmented EDUs by A1 in Spanish texts is 206, while the number of segmented EDUs by A2 in Basque texts is 238. We think there are more EDUs in Basque than in Spanish because Basque nominalization and subordination work with different syntactic procedures (Arakama et al., 2005). Arakama et al. (2005) state that some comprehension problems arise with literal translations of Spanish relatives. To avoid this problem, there is more than one translation strategy, one of them being the splitting of sentences. Language typology has an influence when nominalization is done, because Basque typology uses more verbs than nominalization, given that the ellipsis of verbal arguments is common in Basque (due to verb concordance). Thus, literal translation has no sense or comprehension problems arise.

Both annotators agreed on 152 EDUs. Following the explained methodology in section 2.4.1., we obtained precision (63.9%) and recall (86.6%) of the performed segmentation. The sources of disagreement are linguistic differences, being mainly motivated by translation strategies (85 cases) from Spanish to Basque, which we explore in detail in this section.

We noticed that, sometimes, linguistic differences between texts in Basque and Spanish cause a different segmentation of the same passage by annotators (see example 14).

14a. *[Hemos estudiado retrospectivamente 23 infecciones protésicas de rodilla tratadas en nuestro hospital entre el año 1996 y el 2004 de las cuales hemos excluido 6 por diferentes motivos.]*
ENGLISH TRANSLATION: [We retrospectively have studied 23 prosthetic knee infections that were treated in our hospital between 1996 and 2004 of which we have excluded 6 for different reasons.]
14b. *[1996. eta 2004. urteen bitartean gure ospitalean izandako 23 infekzio protesiko aztertu ditugu.] [Horien artean, 6 kasu baztertu ditugu hainbat arrazoiengatik.]*
ENGLISH TRANSLATION: [We have studied 23 prosthetic knee infections that were treated in our hospital between 1996 and 2004.] [Of these, we have excluded 6 for different reasons.]

In example 14a, we observe that A1 has established a single EDU in Spanish while, in example 14b, we notice that A2 has segmented the same passage in two EDUs. This disagreement on the segmentation phase is due to two facts: a) the relative clause is not considered as an EDU and b) the syntactic structure of the relative clause has been translated into Basque as a different sentence by using punctuation.

When the evaluation of the segmentation was carried out, the same difficulty mentioned by Carlson and Marcu (2001: 2) was found: they declare that the boundary between discourse and syntax can be very blurry. We think this fact is more prominent when structures of two languages are compared:

> The first step in characterizing the discourse structure of a text in our protocol is to determine the elementary discourse units (EDUs), which are the minimal building blocks of a discourse tree. Mann and Thompson (1988, p. 244) state that 'RST provides a general way to describe the relations among clauses in a text, whether or not they are grammatically or lexically signalled.' Yet, applying this intuitive notion to the task of producing a large, consistently annotated corpus is extremely difficult, because the boundary between discourse and syntax can be very blurry.

Indeed, translation strategies are one of the causes influencing segmentation decisions. Consider example 15 below:

15a. *[Se han estudiado un total de 442 cánceres de mama unifocales de 2 cm o menos en la pieza histológica (pT1)* operados *entre enero de 1993 y diciembre de 2005.]*
ENGLISH TRANSLATION: [We have studied a total of 442 unifocal breast cancers of 2 cm or less in the histological part (pT1) *operated between* January 1993 and December 2005.]

15b. *[Guztira, foku bakarreko 442 bularreko minbizi aztertu dira, pieza histologikoan (pT1) 2 cm edo gutxiago dituztenak.] [Guztiak 1993ko urtarrilaren eta 2005eko abenduaren artean* operatu ziren.]

ENGLISH TRANSLATION: [We have studied a total of 442 unifocal breast cancers of 2 cm or less in histological part (pT1).] [All of them *underwent surgery* between January 1993 and December 2005.]

In this example, the non-finite verb (the participle form *operado* ['operated']) was translated into Basque like a finite verb (*operatu ziren* ['underwent surgery']). Besides, the sentence was separated by a full stop. These two facts strongly affect the segmentation in both languages.

We observe various translation strategies affecting the performed segmentation by both annotators, which we explore in detail in section 3.3. It is noteworthy that there is almost a total segmentation agreement concerning EDUs that were not influenced by translation strategies. Segmentation errors of annotators were minimal in these cases.

3.2. Nucleus-satellite issues

Disagreement with regard to the choice of nucleus and satellite is an interesting point of RST. On the one hand, the choice depends on the way the information is presented or the linguistic forms are employed (Marcu, 1999). On the other hand, the choice also depends on the context or the point of view of the whole text (Bateman and Rondhuis, 1997). Stede (2008: 317) criticizes RST because trees do not make the source of the choice explicit:

> The final RST tree does not indicate whether some relation at the level of minimal units is there because its definition is optimally fulfilled or because text global factors make it seem advantageous to select one particular nucleus, which is incidentally performed by that particular relation.

As described in section 2.4.2. above, we measured precision and recall to assess the agreement between the two annotators on spans, nuclearity and rhetorical relations. Table 6 shows an overall result for the 20 texts of the corpus. We noted that results in terms of recall are similar, which is due to EDU homogenization, explained in section 2.3.1. However, results regarding precision vary. Despite this fact, the precision achieved is substantially high in all cases: the agreement between the annotated spans is 92.5 percent, the agreement on nuclearity is 82.1 percent and the agreement regarding the relations is 68.3 percent.

Table 6. Results of the quantitative evaluation

	Recall	Precision
Spans	98.6%	92.5%
Nuclearity	98.6%	82.1%
Relations	98.6%	68.3%

Concerning rhetorical analysis, we mainly observed two types of situations:

1) Ambiguity or different interpretations when choosing relations: Annotators labeled differently some relations that could be ambiguous. For instance, in example 16, while A1 annotated a relation of Background, A2 annotated a relation of Elaboration for the same passage.

16a. *[Han participado 92 pacientes ingresados en un Área Médica del Hospital de Basurto (Bilbao).]N [Todos los pacientes fueron entrevistados para elaborar la historia patopsicobiográfica necesaria para aplicar la Clasificación Psicosomática de Pierre Marty.]S_ Elaboración*

ENGLISH TRANSLATION: [92 patients admitted in a Medical Area Hospital de Basurto (Bilbao) have been involved.]N [All these patients were interviewed to develop the patopsicobiographic history that is needed to apply the Psychosomatic Classification of Pierre Marty.]S_ Elaboration

16b. *[Basurtoko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute.]S_Fondo [Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, elkarrizketa egin zitzaien gaixo guztiei.]N*

ENGLISH TRANSLATION: [92 patients admitted in a Medical Area Hospital of Basurto (Bilbao) have been involved.]S_Background [All these patients were interviewed to develop the patopsicobiographic history that is needed to apply the Psychosomatic Classification of Pierre Marty.]N

In this case, a disagreement regarding the nuclearity of the relation entails a different interpretation about the existing relation between two EDUs. In the example above the nucleus of the Spanish text is the first EDU (the participants of study) (16a), whereas the nucleus of the Basque text is the second EDU (the research methodology) (16b).

Consider other examples:

17a. *[Se estima que el 80% de los usuarios acuden por iniciativa propia a los servicios de urgencia]N_Lista [y que el 70% de las consultas son consideradas leves por el personal sanitario.]N_Lista*

ENGLISH TRANSLATION: [It is calculated that 80% of visitors come to emergency services by their own initiative]N_List [and that 70% of consultations are considered like mild by the health staff.]N_List

17b. *[Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea]N [eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.] S_Elaboración*

ENGLISH TRANSLATION: [80% of visitors come to emergency services by their own initiative]N [and 70% of consultations are considered like mild by the health staff.]S_Elaboration

In example 17 there was also a disagreement concerning nuclearity. However, in this case, the disagreement affects the nature of the relation: A1 annotated a paratactic relation of List (17a), while A2 annotated a hypotactic relation of Elaboration (17b).

18a. *[Por lo demás existen buenos indicadores de proceso]S_Antítesis [pero se aprecia un escaso registro de la capacidad funcional del paciente al alta, que dificulta la comparación de los resultados de la atención sanitaria.]N*

English translation: [In addition, there are good indicators of the process]S_Antithesis [but we see a poor record of the patient's functional ability to discharge, which makes the comparison of health care results difficult.]N

18b. *[Gainerakoan, prozesu adierazle egokiak daude,]N [baina altan dagoen gaixoaren lanen funtzionalaren erregistro urria antzematen da, eta horrek osasun arretaren emaitzen alderaketa zailtzen du.]S_Concesión*

English translation: [In addition, there are good indicators of the process]N [but we see a poor record of the patient's functional ability to discharge, and this makes the comparison of health care results difficult.]N_Concession

In example 18 the disagreement is due to the different meanings of the relation. Both annotators selected a hypotactic relation of presentation but, while A1 annotated an Antithesis relation (18a), A2 annotated a Concession relation (18b).

In this example, the disagreement is not due to the translation, since linguistic forms involved in the relation are identical, including the translation of the discourse marker 'but' (*pero* in Spanish and *baina* in Basque). Thus, we wonder which the source of the disagreement is: is it really a problem of relations definition or maybe a more general problem? This situation was considered by Stede (2008: 318):

> Consider as one example the definitions of Antithesis and Concession. The constraints on the nucleus and the intentions of the writer (i.e., the 'effect') are identical. Antithesis has no constraint on the satellite, whereas Concession offers the constraint that 'writer is not claiming that satellite does not hold'. (Since Antithesis has no constraint here, does it properly subsume Concession?) Finally, the constraints on the nucleus/satellite combinations are largely paraphrastic with the one exception that Antithesis adds that 'one cannot have positive regard for both situations' (in nucleus and satellite). In total, the differences are not very restrictive, so that in many contexts both definitions are equally applicable. But, in the presentational/subject-division of the relations suggested by Mann and Thompson, Antithesis appears in the former, and Concession in the latter, despite their effects being identical. So it is not clear on what grounds the grouping is made in this case.

2) Differences regarding Spanish–Basque translation strategies: the linguistic differences between these two languages sometimes imply that annotators interpret the same passage differently (see examples 19 and 20).

19a. *[Escogiendo la especialidad más barata existente en el mercado]S_Circunstancia [podríamos alcanzar un ahorro de 6.463.400,35€.]N*

English translation: [Choosing the cheapest specialty in the market]S_Circumstance [we could achieve a saving of 6,463,400.35€.]N

19b. *[Merkatuak eskaintzen digun espezialitate merkeena aukeratuko bagenu]S_Condición [6.463.400,35€-ko aurrezpena lortuko genuke.]N*

English translation: [If we chose the cheapest specialty in the market]S_Condition [we would achieve a saving of 6,463,400.35€.]N

The gerund form (*escogiendo* ['choosing']) may indicate the relation of Circumstance in Spanish. But in Basque no gerund is included in the sentence; the conditional mark (*ba-* ['if']) in the verb (*bagenu* ['(we) chose']) justifies the annotation of the relation of Condition.

20a. *[En los 7 ítems se han encontrado diferencias estadísticamente significativas entre el grupo de pacientes oncológicos con los pacientes afectos de otro tipo de patologías (p < 0.05).]N [*Estos ítems *diferencian a los pacientes con neoplasias de otro tipo de pacientes, y permiten una valoración global de los mismos, ofreciendo una idea de las expectativas del proceso.]S_Elaboración*

ENGLISH TRANSLATION: [In the 7 items we have found statistically significant differences between the group of cancer patients and patients suffering from other pathologies (p < 0.05).]N [These items differentiate patients with tumors from other patients, and they allow an overall assessment of the patients, providing an idea of the process prospects.]S_Elaboration

20b. *[7 itemak aztertuta, estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p < 0.05).]N_Unión [Horrez gain, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte, horiei buruzko balorazio orokorra egiteko aukera ematen dute, eta prozesuaren igurkapenen gaineko argibideak ematen dizkigute.]N_Unión*

ENGLISH TRANSLATION: [Having studied the 7 items, we have found statistically significant differences between the group of cancer patients and patients suffering from other pathologies (p < 0.05).]N_Joint [In addition, these items differentiate patients with tumors and other patients, they allow an overall assessment of the patients, and they provide an idea of the process prospects.]N_Joint

In Spanish, the relation of Elaboration was annotated due to the presence of the anaphora. The semantic relation between both EDUs shows an elaboration of the same topic. Nevertheless, in Basque, the additive connector *horrez gain* ('in addition') does not allow inclusion of both EDUs in the same argumentative scale (Cuartero, 1995), since it introduces a new topic in the speech. This fact causes A2 to select a multinuclear relation. Therefore, it is evident that a different translation strategy affects the rhetorical analysis of the text.

We studied this phenomenon systematically, which we explain in detail in section 3.3.

3.3. Discussion of translation strategies

As we have said in section 3.1, translation strategies are one of the causes influencing segmentation decisions. We observe various translation strategies affecting the performed segmentation by both annotators. Specifically, the authors of the texts used two main strategies to translate from Spanish into Basque. These two strategies constitute the 74.28 percent of all the translation strategies.

- Relative subordinate clauses in Spanish have been translated as separate sentences in Basque.
- Missing elements from ellipsis and anaphors in Spanish are retaken in Basque, forming new sentences.

The consequences of these translation strategies are:

- There are more EDUs in Basque than in Spanish. Specifically, in our corpus, there are 13.45 percent more EDUs in Basque than in Spanish.

- This difference between EDUs in the two languages significantly affects the agreement on the segmentation, and therefore it affects in a gradual way the other annotation levels and evaluated factors (spans, nuclearity and relations) as well. This fact makes quantitative and qualitative evaluation more difficult to perform.

As we have said in section 3.2, translation strategies may be the cause of a different rhetorical analysis. We include in Table 7 the used strategies to translate from Spanish into Basque, with their frequencies.

Three of these translation strategies are mentioned in Arakama et al. (2005): completing ellipsis and/or dividing sentences, using a finite verb and deleting relative clauses. Another of these strategies is used when the translator wants to provide more coherence to the translation: using discourse markers (Zabala, 1996).

We provide some examples herein:

a) Completing ellipsis and/or dividing sentences:

21a. *[Todos los pacientes presentaban una insuficiencia ventilatoria, en 10 casos de tipo obstructivo y en los restantes de tipo no obstructivo o mixto.]*

English translation: [All patients had ventilatory failure, 10 cases of obstructive type and the remaining of non-obstructive or mixed type.]

21b. *[Gaixo guztiek zeukaten aireztapen gutxiegitasuna;] [hamar kasutan butxaketa-motakoa zen] [eta gainerakoetan ezbutxaketakoa edo mistoa zen.]*

English translation: [All patients had ventilatory failure;] [10 cases were of obstructive type] [and the remaining were of non-obstructive or mixed type.]

In this example, the translation strategy was in Basque to complete the ellipsis of verbs describing the cases of 'ventilatory failure'.

b) Using a finite verb:

22a. *[Estudiamos 47 pacientes y 82 pies* intervenidos *entre los años 1992 y 2004.]*

English translation: [We studied 47 patients and 82 feet *undergoing surgery* between 1992 and 2004.]

22b. *[1992. eta 2004. urte bitartean, 47 gaixo aztertu genituen,] [eta 82 oinetan* egin genuen ebakuntza.]

English translation: [Between 1992 and 2004, we studied 47 patients] [and we *operated* 82 feet.]

Table 7. Translation strategies determining different rhetorical relations

Translation strategies	Spanish	Basque	Total
a) Completing ellipsis and/or dividing sentences	1	5	6
b) Using a finite verb	0	5	5
c) Using discourse markers	2	7	9
d) Deleting relative clauses	0	6	6
e) Other strategies	0	5	5
Total	3	28	31

The Spanish participle (*intervenidos* ['undergoing surgery']) was translated into Basque by a structure with a finite verb and its direct object (*ebakuntza egin genuen* ['(we) operated']).

23a. *[Nuestros resultados sugieren la presencia de alteraciones respiratorias crónicas con el resultado de un déficit ventilatorio, varias décadas después del tratamiento con colapsoterapia;* comprobando *una buena respuesta al tratamiento con ventilación domiciliaria.]*
ENGLISH TRANSLATION: [Our results suggest the presence of chronic respiratory disorders with the result of a ventilatory deficit, several decades after treatment with Collapse Therapy; *proving* a good response to treatment with home ventilation.]
23b. *[Gure emaitzek iradokitzen dute kolapsoterapiarekin egindako tratemendutik hamarkada batzuk gerago arnas alterazio kronikoak daudela aireztapen déficit baten emaitzarekin;] [eta* egiaztatu *da etxeko aireztapenarekin egindako tratamenduak erantzun ona izan duela.]*
ENGLISH TRANSLATION: [Our results suggest the presence of chronic respiratory disorders with the result of a ventilatory deficit, several decades after treatment with Collapse Therapy;] [and a good response to treatment with home ventilation *has been proved*.]

In this example, the Spanish gerund (*comprobando* ['proving']) was translated into Basque by the finite verb (*egiaztatu da* ['(it) has been proved']).

c) Using discourse markers:

24a. *[Como cirugía primaria presenta una mortalidad del 0,5%] [y un 8,8% de complicaciones perioperatorias, destacando la hemorragia (4,8%) y la dehiscencia anastomótica (1,7%).]*
ENGLISH TRANSLATION: [As primary surgery, it presents a mortality of a 0.5%] [and a 8.8% of perioperative complications, standing out hemorrhages (4.8%) and dehiscence of anastomosis (1.7%).]
24b. *[Kirurgia mota honetan, heriotza tasa % 0,5ekoa da,] [eta ebakuntza osteko arazoak, berriz, % 8,8koak dira: odoljarioa (% 4,8) eta dehiszentzia anastomotikoa (% 1,7).]*
ENGLISH TRANSLATION: [In this type of surgery, the mortality rate is 0.5%] [while the perioperative complications are 8.8%: haemorrhages (4.8%) and dehiscence of anastomosis (1.7%).]

The use of the Basque counterargument connector *berriz* ('while') shows a contrast, not a contradiction. This connector means that A2 labels this passage with a Contrast relation, while A1 labels the same passage with List relation, because he did not have any discourse marker.

d) Deleting relative clauses:

25a. *[Creemos que es importante dar a nuestros pacientes una información previa a la exploración lo más precisa posible, que sea capaz de resolver todas las posibles dudas que les plantee y que les permita afrontarla con tranquilidad.]*
ENGLISH TRANSLATION: [We think that it is important to give our patients a pre-scan information as accurate as possible, being able to resolve all the possible doubts raised by it and allowing them to deal with it peacefully.]
25b. *[Garrantzitsua iruditzen zaigu azterketa egin baino lehen, gaixoei informazio zehatza aurreratzea.] [Horrela, bere zalantzak argituz, hobeto egingo diote aurre azterketari.]*
ENGLISH TRANSLATION: [We think that it is important to give our patients a pre-scan information as accurate as possible.] [In this way, resolving their doubts, they will deal better with the medical examination.]

Table 8. Data of the partial qualitative evaluation

	Absolute data	%
Total relations	224	100%
Agreement on relations	157	71%
Disagreements on relations	65	29%
Translation source	31	13.8%
Interpretation source	34	15.2%

In this example, the literal translation of the relative clause used in Spanish was avoided in Basque and it was translated by an independent sentence with a finite verb (*aurre egingo diote* ['(they) will deal with']).

Once all the cases have been described, we conclude that the use of the detected translation strategies is due to the fact that Basque sentences have the semantic load at the end of the sentence, since it is an SOV language. In order to facilitate the understanding, the translator has to locate the semantic load earlier in the sentence or has to reduce the size of it. In this corpus more sentences in Basque than in Spanish were used to facilitate the understanding of the semantic content. Precisely for this reason (to shorten sentences), some translation strategies were used in Basque. The use of these strategies definitely increases the linguistic differences that affect the rhetorical structure, changing the relations among EDUs and, thus, changing sometimes the meaning of the text or, at least, the presentation of the information. If the meaning of the text is different, it is normal that the disagreement between the annotators increases and, thanks to the partial qualitative evaluation, this great increase in the disagreement becomes an indicator of translation techniques.

Table 8 shows the data of the partial qualitative evaluation that we performed in this work.

Finally, Table 9 provides recall and precision of the quantitative evaluations, and recall of the qualitative evaluation. It is noticed that the precision of both evaluations is very similar (68.3% in the quantitative evaluation and 71% in the qualitative evaluation).

As it is shown in Table 9, the precision of the qualitative evaluation from the comparison of the 20 rhetorical trees of the corpus is more optimistic than the quantitative one, but not too much (only 2.7% more). However, this situation is not constant, since in some trees the difference between evaluations ranges approximately from –10% to +10%.

Although the use of translation strategies definitely affects rhetorical structures, it does not seem to affect the texts' superstructure, since both annotators have constructed a very similar superstructure for both languages. The macrostructure of a text is, according to van Dijk (1980, 1989), an abstract representation which tends to the overall understanding of the meaning of the text, while the superstructure is the organizational structure of the text, which can vary depending on the type of the text. Van Dijk (1989) described the superstructure of various types of texts, for example scientific texts, and he stated that:

> *En los discursos científicos se presenta una variante especial de las superestructuras argumentativas [. . .]. La estructura básica del discurso científico no (sólo) consiste en una CONCLUSIÓN y su JUSTIFICACIÓN, sino también en un PLANTEO DEL PROBLEMA y una SOLUCIÓN*. (van Dijk, 1989: 164)
>
> English translation: Scientific discourse provides a special variant of argumentative superstructures [. . .]. The basic structure of scientific discourse is not (only) a CONCLUSION and its JUSTIFICATION, but also a PROBLEM STATEMENT and a SOLUTION. (van Dijk, 1989: 164)

Table 9. Final results of quantitative evaluation and partial qualitative evaluation

	Quantitative		Qualitative
	Recall	Precision	Precision
Relations	98.6%	68.3 %	71%

For example, van Dijk (1989) analyzed the superstructure of the *Experimental Report*, finding in it some *observations*, an *explanation*, a *hypothesis*, an *experiment*, etc. In this work we also analyze a scientific discourse but, as we have already discussed, our corpus of analysis includes abstracts of original articles, specifically from the medical field. These abstracts maintain the same superstructure of the articles that are related to them and, therefore, they have four main sections: *Introduction*, *Patients and methods*, *Results* and *Discussion*. This structure was labeled exactly by both annotators, by means of RST relations as Background, Means, Result and Interpretation. Figure 6 shows a diagram of this structure.

4. Conclusions

To conclude, we think that this work represents a new contribution concerning RST, since it extends our understanding about the comparison of rhetorical trees in various languages, specifically the comparison between Spanish and Basque, that had not been made before. We have mentioned some problems of quantitative evaluation, and an original qualitative evaluation has also been presented. Our work shows that, though there are differences regarding rhetorical analysis performed over the same corpus (with parallel texts in two languages) by two annotators, these are mainly due to the translation strategies being used. However, these strategies do not affect the superstructure of medical abstracts in a decisive way.

Another conclusion of this work is that translation strategies influence the interpretation of RST rhetorical relations. The translator did sometimes not use the same linguistic structures when translating from one language into another. Since the rhetorical structures were not maintained, the two annotators of our study interpreted differently a same passage written in two languages.

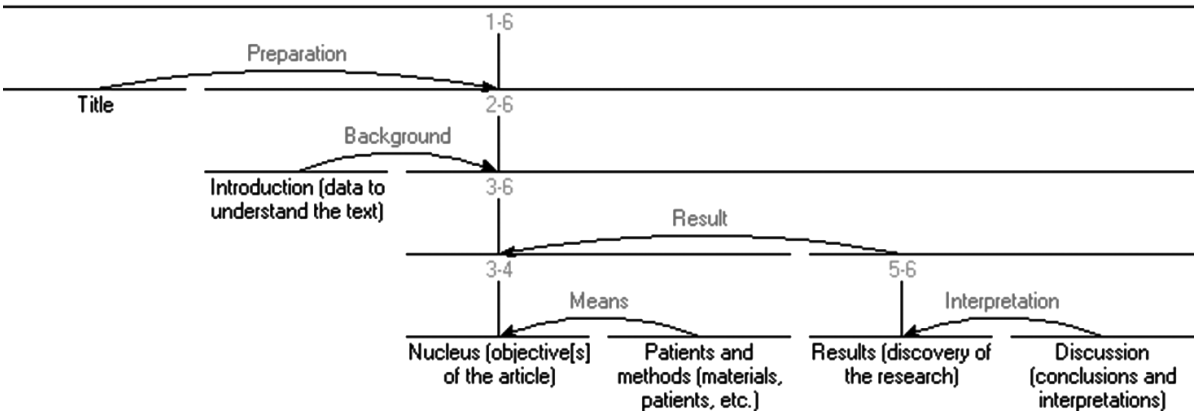


Figure 6. Main superstructure labeled by both annotators

Likewise, the comparison of rhetorical trees of parallel texts has allowed us to observe two situations: a) when translating an abstract, its rhetorical structure is not taken into account as much as its syntactic structure, and b) in the cases where it is not convenient to translate syntactic structures literally, the used translation strategies provide some clues about how languages usually structure their discourse (which is an issue to take into account for automatic translation of rhetorical structures).

As future work, we would like to compare the top spans of rhetorical structures in order to determine the level of agreement concerning the superstructure, and to analyze the linguistic factors determining the disagreement on rhetorical structure. Although the abstracts are quite short, we think their length is enough to evaluate the agreement of the annotators. Furthermore, we would like to study the reasons for the oscillations between the quantitative and qualitative evaluations, and to also add to this study a third language, English, since, as we have already mentioned, *Gaceta Médica de Bilbao* also includes the abstracts of the authors in that language. We consider that it is important to observe which types of translation strategies have been used and the existing differences among them. As English and Spanish are linguistically more similar, the applied translation strategies should be reduced and, therefore, this variable would decrease when comparing closer languages. In addition, we would like to confirm if medical abstracts in English have the same superstructure. Moreover, we plan to carry out a compilation of discourse markers in Spanish, Basque and English, starting from an empirical analysis of medical abstracts written in these three languages. The main goal of this last study would be to analyze the correlations among rhetorical relations and discourse markers, in the same way that Iruskieta et al. (in press) have done.

Notes

1. http://www.gacetamedicabilbao.org/web/es/.
2. The English translation is ours (see http://www.gacetamedicabilbao.org/web/es/autores.php).
3. The following examples are proposed by Carlson and Marcu (2001).
4. Throughout this article, examples marked with ‘a’ show the segmentation included in Carlson and Marcu (2001), and examples marked with ‘b’ show the segmentation that we would establish in our work.
5. ‘Deficit’ is part of the unit ‘it will be able to halve this year’s 120 billion ruble’.
6. http://dialnet.unirioja.es/servlet/revista?tipo_busqueda=CODIGO&clave_revista=2426.
7. http://www.wagsoft.com/RSTTool/.
8. For the purpose of this article, we have tried to do, for the English translation, the EDU segmentation as similar as possible with regard to the one proposed in Spanish and Basque.
9. Marcu (2000b) names them ‘spans’.
10. Note that numerical elements are included in one column in Table 1, while in Table 3 these elements are included in the first two.

References

Abelen, E., Redeker, G. and Thompson, S.A. (1993) ‘The Rhetorical Structure of US-American and Dutch Fund-Raising Letters’, *Text* 13(3): 323–350.

Arakama, J.M., Arrieta, A., Lozano, J., Robles, J. and Urrutia, R.M. (2005) *IVAPeko Estilo Liburua*. Zarautz: IVAP.

Bateman, J.A. and Rondhuis, K.J. (1997) 'Coherence Relations: Towards a General Specification', *Discourse Processes* 24: 3–50.

Bouayad-Agha, N. (2000) 'Using an Abstract Rhetorical Representation to Generate a Variety of Pragmatically Congruent Texts', in *Proceedings of the 38th Meeting of the Association for Computational Linguistics*. *Student Workshop*, 16–22.

Burstein, J. and Marcu, D. (2003) 'A Machine Learning Approach for Identification of Thesis and Conclusion Statements in Student Essays', *Computers and the Humanities* 37(4): 455–467.

Carlson, L. and Marcu, D. (2001) *Discourse Tagging Reference Manual*. ISI Technical Report ISITR-545. Los Angeles, CA: University of Southern California.

Carlson, L., Marcu, D. and Okurowski, M.E. (2001) 'Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory', in *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*. 1–10.

Cuartero, J.M. (1995) 'El estatuto categorial de *además* y sus propiedades distribucionales', *Dicenda* 13: 103–118.

Cui, S. (1986) 'A Comparison of English and Chinese Expository Rhetorical Structures', Unpublished Master's thesis, UCLA.

da Cunha, I. (2008) *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Barcelona: IULA. [CD-ROM] (Sèrie Tesis; 23).

da Cunha, I. and Torres-Moreno, J.-M. (2010) 'Automatic Discourse Segmentation: Review and Perspectives', in *Proceedings of the International Workshop on African Human Languages Technologies*. Djibouti: Institute of Sciences and Information Technologies.

da Cunha, I., Wanner, L. and Cabré, M.T. (2007) 'Summarization of Specialized Discourse: The Case of Medical Articles in Spanish', *Terminology* 13(2): 249–286.

Delin, J., Hartley, A. and Scott, D. (1996) 'Towards a Contrastive Pragmatics: Syntactic Choice in English and French Instructions', *Language Sciences* 18(3–4): 897–931.

Ghorbel, H., Ballim, A. and Coray, G. (2001) 'ROSETTA: Rhetorical and Semantic Environment for Text Alignment', in P. Rayson, A. Wilson, A.M. McEnery, A. Hardie and S. Khoja (eds) *Proceedings of Corpus Linguistics 2001*, pp. 224–233.

Haouam, K. and Marir, F. (2003) 'SEMIR: Semantic Indexing and Retrieving Web Document using Rhetorical Structure Theory', *Lecture Notes in Computer Science*: 596–604.

Iruskieta, M., Diaz de Ilarraza, A. and Lersundi, M. (in press) 'Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso', *Proceedings of 27th AESLA International Conference: Ways and Modes of Human Communication*. Ciudad Real: Universidad de Castilla-La Mancha.

Kong, K.C.C. (1998) 'Are Simple Business Request Letters Really Simple? A Comparison of Chinese and English Business Request Letters', *Text* 18(1): 103–141.

Mann, W.C. (2005) *RST Web Site*. Available at: www.sfu.ca/rst (accessed 15 August 2009).

Mann, W.C. and Thompson, S.A. (1988) 'Rhetorical structure theory: Toward a functional theory of text organization', *Text* 8(3): 243–281.

Marcu, D. (1998) 'The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts', PhD thesis, University of Toronto.

Marcu, D. (1999) *Instructions for manually annotating the discourse structure of texts*. Available at: http://www.isi.edu/~marcu.

Marcu, D. (2000a) *The Theory and Practice of Discourse Parsing Summarization*. Cambridge, MA: Massachusetts Institute of Technology.

Marcu, D. (2000b) 'The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach', *Computational Linguistics* 26(3): 395–448.

Marcu, D., Amorrortu, E. and Romera, M. (1999) 'Experiments in Constructing a Corpus of Discourse Trees', in *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*: 48–57.

Marcu, D., Carlson, L. and Watanabe, M. (2000) 'The Automatic Translation of Discourse Structures', *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 9–17.

Mazeiro, E. and Pardo, T.A.S. (2009) 'Metodologia de avaliação automática de estruturas retóricas', in *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*. São Carlos, São Paulo.

O'Donnell, M. (2000) 'RSTTOOL 2.4 – A markup tool for rhetorical structure theory', in *Proceedings of the International Natural Language Generation Conference*: 253–256.

Pardo, T.A.S. and Nunes, M.G.V. (2008) 'On the Development and Evaluation of a Brazilian Portuguese Discourse Parser', *Journal of Theoretical and Applied Computing* 15(2): 43–64.

Pardo, T.A.S., Nunes, M.G.V. and Rino, L.H.M. (2004) 'DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese', *Lecture Notes in Artificial Intelligence*: 224–234.

Ramsay, G. (2000) 'Linearity in Rhetorical Organisation: A Comparative Cross-Cultural Analysis of Newstext from the People's Republic of China and Australia', *International Journal of Applied Linguistics* 10(2): 241–258.

Ramsay, G. (2001) 'What are they Getting At? Placement of Important Ideas in Chinese Newstext: A Contrastive Analysis with Australian Newstext', *Australian Review of Applied Linguistics* 24(2): 17–34.

Salkie, R. and Oates, S.L. (1999) 'Contrast and Concession in French and English', *Languages in Contrast* 2(1): 27–56.

Scott, D., Delin, J. and Hartley, A. (1998) 'Identifying Congruent Pragmatic Relations in Procedural Texts', *Languages in Contrast* 1(1): 45–82.

Skadhauge, P. and Hardt, D. (2005) 'Syntactic Identification of Attribution in the RST Treebank', in *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*. Jeju Island, 57–62.

Stede, M. (2008) 'Disambiguating Rhetorical Structure', *Journal of Research in Language and Computation* 6: 311–332.

Sumita, K., Ono, K., Chino, T., Ukita, T. and Amano, S. (1992) 'A Discourse Structure Analyzer for Japanese Text', in *Proceedings of the International Conference on Fifth Generation Computer Systems*, 1133–1140.

Swales, J. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taboada, M. and Mann, W.C. (2005) 'Applications of Rhetorical Structure Theory', *Discourse Studies* 8(4): 567–588.

Taboada, M. and Mann, W.C. (2006) 'Rhetorical Structure Theory: Looking Back and Moving Ahead', *Discourse Studies* 8(3): 423–459.

van Dijk, T.A. (1980) *Macro-Structures. An Interdisciplinary Study of Global Structures in Discourse, Cognitions and Interaction*. Hillsdale, NJ: Lawrence Erlbaum.

van Dijk, T.A. (1989) *La ciencia del texto*. Barcelona: Paidós.

Zabala, I. (1996) 'Testu-lotura: lotura tematikoa eta erreferentzia-sareak testu teknikoetan', in *Testu-loturarako baliabideak: euskara teknikoa*, pp. 15–44. Bilbao: EHU.

Appendix Table 1. Information about the analyzed corpus[a]

Reference	Title	Author(s)	Year
Text 1	Pharmacoepidemiologic and pharmacoeconomic study of arterial hypertension	L.C. Abecia	2008
Text 2	Serious psychomatic criteria in oncology	R. Ruiz, A. Aljelani, U. Shelick, U. Usobiaga, J. Muro, J. Bilbao, F. Franco	2007
Text 3	The 'basal-like' (c-erb-B2 -, ER - and PR - negative) tumour phenotype defines a biologically highly aggressive subgroup of surgical pT1 stage breast cancers	J. Schneider, A. Tejerina, C. Perea, A. Tejerina R. Lucas, J. Sánchez	2007
Text 4	Real incidence of axillar nodal invasion in T1 breast cancer among our population	J. Schneider, A. Tejerina, J. Sánchez, J. Lucas	2007
Text 5	Prosthetic infection of knee	O. Sáez-de-Ugarte-Sobrón, I. Gutiérrez-Sánchez, A. Cruchaga-Celada, F. Labayru-Etxebarria, I. Garcia Sánchez, A. Álvarez-González	2008
Text 6	Recurrent aphthous stomatitis (I): Epidemiologic, ethiologic and clinical features	A. Eguía, R. Saldón, J. M. Aguirre	2003
Text 7	The surgery of the carotid bifurcation in cerebral ischemia of extracranial origin: A 10 year experience	L. Estallo, A. Barba, L. Rodríguez, S. Gimena, A. G. Alfageme	2000
Text 8	Uncommon clinical features in Whipple's disease: An assay of four cases	E. Ojeda, A. Cosme, J. Lapaza, J. Torrado, I. Arruabarrena, L. Alzate.	2005
Text 9	Evolution of the anthropometric measures in children's feet: Correlation indices with other variables	R. De los Mozos, A. Alfageme, E. Ayerdi	2002
Text 10	Evolution of the anthropometric measures in children's feet: A stratified descriptive study	R. De los Mozos, A. Alfageme, E. Ayerdi	2002
Text 11	Evolution of the anthropometric measures in children's feet: An overall descriptive study	R. De los Mozos Bozalongo, A. Alfageme Cruz, E. Ayerdi Salazar	2003
Text 12	Stroke acute care and improvement possibilities	J. Pérez-de-Arriba, G. Achutegui, L. Epelde, G. Viñegra, J.L. Elexpuru.	2005
Text 13	Morbidity and tolerance of the ultrasound-guided prostatic biopsy punction in 392 patients	J. A. López-Lendoiro, P. Aísa, X. Aguirre, E. Añorbe, M. Paraíso	2002

Appendix Table 1. (Continued)

Reference	Title	Author(s)	Year
Text 14	Surgical treatment of infantile flexible flan using the calcaneus-stop technique	I. Etxebarria-Foronda, I. Garmilla-Iglesias, A. Gay-Vitoria, J. Molano-Muñoz. D. Izal-Miranda, E. Esnal-Baza, A. Ruiz-Sánchez.	2006
Text 15	The profile of the users from the emergency department from Galdakao's Hospital	I. Bengoetxea Martínez	2004
Text 16	Fast progression dementia and myoclonus	I. Villamil-Cajoto, A, M. J. González-Quintela, V. Villacian-Vicedo	2005
Text 17	Surgical and ultrasound correlation in full thickness tears of the shoulder rotator cuff	J. de la Fuente-Ortiz-de-Zárate, J. Kutz-Peyroncelli, J. L. Imizcoz-Barriola	2004
Text 18	Surgical treatment for morbid obesity	I. Díez-del-Val, C. Martínez-Blázquez, V. Sierra-Esteban, J. M. Vitores-López, J. Valencia-Cortejoso	2005
Text 19	Progress of patients undergoing collapsotherapy due to pulmonary tuberculosis	K. Abu-Shams, J. Ardanaz, M. Murie, A. Sebastián, G. Tiberio, A. Arteche.	2000
Text 20	Pseudomonas aeruginosa infection-colonization in patients with bronchiectasias or COPD. Clinical features, microbiology and outcome	J. Garrós Garay, E. Ruiz de Gordejuela, G. Martín Saco, L. Gallego, J. Pérez Escajadillo, F. García Cebrián	2002

[a]The titles in English have been extracted from the original articles, except for the titles of texts 7 and 19; we have translated these from Spanish into English.

Appendix Table 2. List of relations used in this study following the extended version and with representative examples in Spanish and Basque[a]

Relation		Example
CONTRAST (N-N)	S	*[Los antecedentes de primer grado se relacionan con un mayor riesgo de aparición del tumor,]$_N$ [mientras que los antecedentes familiares de segundo grado no influyen de manera importante.]$_N$*
	B	*[Lehen graduko aurrekariak tumorearen agertze arrisku handiagoekin lotzen dira;]$_N$ [bigarren graduko aurrekari familiarrak, ordea, ez dute modu garrantzitsuan eragiten]$_N$*
	E	[First-degree medical history is associated with an increased risk of developing the tumour,]$_N$ [while second-degree family medical history did not influence significantly.]$_N$
JOINT (N-N)	S	*[En todos los pacientes se realizó un seguimiento radiológico]$_N$ [y fueron dados de alta tras una radiografía del abdomen sin evidencia de cuerpos extraños.]$_N$*
	B	*[Paziente guztiei erradiologiako jarraipena egin zaie]$_N$ [eta gorputz arrotzen ebidentzia gabeko sabelaldearen erradiografien ostean guztiei alta eman zitzaien]$_N$*
	E	[All the patients underwent radiological monitoring]$_N$ [and were discharged after a scan of the abdomen without evidence of strange bodies.]$_N$
LIST (N-N)	S	*[El 68% de los pacientes eran varones.]$_N$ [El 92% procedían de Colombia.]$_N$ [El 65% ingirieron fármacos antidiarreicos.]$_N$*
	B	*[Pazienteen % 68a gizonezkoak ziren.]$_N$ [% 92ak kolonbiar jatorria zuen.]$_N$ [% 65ak beherakoaren kontrako botika irentsi zuen.]$_N$*
	E	[68% of patients were male.]$_N$ [92% came from Colombia.]$_N$ [65% ingested anti-diarrhea medication.]$_N$
SEQUENCE (N-N)	S	*[A todos ellos se les realizaron una historia clínica y un examen físico.]$_N$ [Se les preguntó por el país de procedencia.]$_N$ [Se registraron la frecuencia cardíaca, la temperatura y la presión arterial.]$_N$*
	B	*[Horiei guztiei egin zitzaien historia klinikoa eta azterketa fisikoa.]$_N$ [Jatorriko herrialdeaz galdetu zitzaien.]$_N$ [Bihotz-maiztasuna, tenperatura eta presio arteriala erregistratu ziren.]$_N$*
	E	[We carried out a medical history and a physical examination to all of them.]$_N$ [We asked them their country of origin.]$_N$ [We registered their heart rate, temperature and blood pressure.]$_N$
DISJUNCTION (N-N)	S	*[La mayoría de los pacientes que han perdido peso de forma apreciable roncan menos]$_N$ [o han dejado de hacerlo por completo.]$_N$*
	B	*[Pisua nabarmen galdu duten pazienteen gehiengoak zurrunga gutxiago egiten dute]$_N$ [edo zurrunga egiteari utzi diote]$_N$*
	E	[Most of the patients who have lost weight appreciably snore less]$_N$ [or they have stopped completely.]$_N$
CONJUNCTION (N-N)	S	*[Mendel no sabía que los genes se localizan en cromosomas]$_N$ [ni que los genes localizados uno cerca del otro en el mismo cromosoma se transmiten juntos.]$_N$*
	B	*[Mendelek ez zekien geneak kromosometan kokatzen zirela]$_N$ [ezta elkarrekin transmititzen zirela ere kromosoma batean bata bestetik hurbil kokaturiko geneak.]$_N$*
	E	[Mendel did not know that genes are located in chromosomes]$_N$ [nor that genes that are located near each other in the same chromosome are transmitted together.]$_N$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
BACKGROUND (N-S)	S	*[A los portadores de cuerpos extraños intraabdominales que contienen cocaína, con fines de contrabando, se les conoce con el síndrome del body packer.]$_S$ [Hemos estudiado la aparición de complicaciones en el seguimiento de individuos que ingieren estos paquetes de droga, con el fin de poder dar unas normas de actuación en estos casos.]$_N$*
	B	*[Kokainadun sabelalde barneko gorputz arrotzen eramaileak, kontrabando helburudunak, "body packer" sindromea izenaz ezagutzen dira.]$_S$ [Droga pakete hauek irensten dituzten norbanakoen jarraipenean konplikazioen agerpenak ikertu ditugu.]$_N$*
	E	[Persons who transport strange bodies containing cocaine by internal concealment for smuggling purposes are referred to body packer syndrome.]$_S$ [We have analyzed the monitoring complications of persons that consume these packets of drug, with the objective of giving rules of conduct in these cases.]$_N$
CIRCUMSTANCE (N-S)	S	*[Parece necesario propiciar algún tipo de campaña informativa para sensibilizar a la población femenina ante el cáncer de mama,]$_N$ [mientras no se diluciden las incógnitas que plantean las costosas campañas de detección temprana.]$_S$*
	B	*[Bularreko minbiziaren aurrean beharrezkoa dirudi emakumezko biztanleriari zuzendutako nolabaiteko informazio-kanpainari bide ematea,]$_N$ [goiz antzemate kanpaina garestien auzia argitzen ez den bitartean behintzat.]$_S$*
	E	[It seems necessary to carry out some sort of information campaign to sensitize the population to the female breast cancer,]$_N$ [until the factors of costly campaigns of early detection are not adequately considered.]$_S$
CONCESSION (N-S)	S	*[El porcentaje de curación fue algo menor en los obesos que en los no obesos,]$_N$ [aunque esta diferencia no ha sido estadísticamente significativa.]$_S$*
	B	*[Sendatze-portzentajea zerbait hobeagoa izan da pertsona gizenetan ez-gizenetan baino,]$_N$ [nahiz eta diferentzia hori ez den estatistikoki esanguratsua izan.]$_S$*
	E	[The cure rate was slightly lower in obese people than in non-obese people,]$_N$ [although this difference was not statistically significant.]$_S$
CONDITION (N-S)	S	*[A efectos del presente estudio consideramos que ha habido acceso a la mamografía]$_N$ [si la mujer se ha realizado al menos una prueba en los 2 años previos a la realización del estudio.]$_S$*
	B	*[Ikerketa honen xedeetarako mamografia egin izan dela kontsideratu dugu]$_N$ [baldin eta emakumeak gutxienez froga bat egin izan badu ikerketa egin baino 2 urte lehenago]$_S$*
	E	[In this study, we consider that there has been access to mammography]$_N$ [if the woman has had at least one test in the 2 years preceding the survey.]$_S$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
ELABORATION (N-S)	S	*[Los pacientes suicidas que padecían una enfermedad orgánica eran 45.]$_N$ [La edad media de estos pacientes fue de 58,3 años (varones 57,6 años y mujeres 59,2 años) con unos límites de 16 a 90 años.]$_S$*
	B	*[Gaixotasun organikoa zuten pazienteak 45 izan dira]$_N$ [16 eta 90 urte bitarteko paziente hauen bataz besteko adina 58,3 urtekoa izan zen (gizonezkoak 57,6 urte eta emakumezkoak 59,2 urte)]$_S$*
	E	[Suicidal patients suffering from organic disease were 45.]$_N$ [The average age of these patients was 58.3 years (men 57.6 years and women 59.2 years) with a range of 16 to 90 years.]$_S$
JUSTIFICATION (N-S)	S	*[Se realizó cirugía en 7 pacientes (3.3%),]$_N$ [en cinco de ellos porque presentaban obstrucción, en uno por rotura de uno de los paquetes y en otro por ausencia de progresión de dos de los paquetes que eran de tamaño superior al resto.]$_S$*
	B	*[7 pazientengan (% 3,3a) kirurgia burutu zen,]$_N$ [haietako bostek buxadura zutelako, beste bati paketeetako bat apurtu zitzaiolako eta beste bati handiagoak ziren 2 paketeren kanporaketan garapenik agertzen ez zelako.]$_S$*
	E	[Surgery was performed in 7 patients (3.3%),]$_N$ [in five of them because they had obstruction, in one due to the breakage of one package and in another one because of lack of progression of two packages that were larger than the rest.]$_S$
PURPOSE (N-S)	S	*[Para que puedan cumplir su función con eficacia,]$_S$ [los SUH precisan que exista un equilibrio apropiado entre la demanda asistencial y su capacidad de respuesta.]$_N$*
	B	*[Eraginkortasunez haren funtzioa bete dezan,]$_S$ [SUHak laguntza-eskaeraren eta haren erantzun-gaitasunaren arteko oreka egokia eduki behar du.]$_N$*
	E	[In order to fulfil their role effectively,]$_S$ [ED needs a proper balance between care demand and its responsiveness.]$_N$
REFORMULATION (N-S)	S	*[Se incluyeron sólo pacientes que se consideraba que estaban estables,]$_N$ [es decir, que no habían precisado cambiar su medicación habitual en los últimos 15 días y clínicamente no referían un empeoramiento importante.]$_S$*
	B	*[Egonkor zeudela kontsideratzen ziren pazienteak bakarrik sartu genituen,]$_N$ [hau da, azkeneko 15 egunetan ohiko medikazioa aldatu behar izan ez zutenak eta klinikoki okerrera egin ez zutenak.]$_S$*
	E	[We have included only patients who were considered as stable,]$_N$ [that is, patients who did not need to change their regular medication in the last 15 days and who reported no significant worsening clinically.]$_S$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
RESULT (N-S)	S	*[Se practicó una radiografía simple del abdomen en todos los enfermos.]$_N$ [Se observaron cuerpos extraños intra-abdominales en el 98,6% de los enfermos.]$_S$*
	B	*[Gaixo guztietan sabelaldearen erradiografia sinplea praktikatu da.]$_N$ [Sabelalde barneko gorputz arrotzak gaixoen % 98,6gan hauteman ziren.]$_S$*
	E	[All patients underwent normal radiographs of the abdomen.]$_N$ [Intra-abdominal strange bodies were detected in 98.6% of the patients.]$_S$
SUMMARY (N-S)	S	*[Se realizó una radiografía simple.]$_N$ [También se llevó a cabo una radiografía combinada mediante varias técnicas.]$_N$ [En resumen, se han aplicado diferentes pruebas radiológicas.]$_S$*
	B	*[Erradiografia sinplea egin zen.]$_N$ [Zenbait teknika bidezko erradiografia konbinatua ere egin zen.]$_N$ [Laburtuz, froga erradiologiako desberdinak aplikatu izan dira.]$_S$*
	E	[A normal X-ray was performed.]$_N$ [We also carried out a combined X-ray by several techniques.]$_N$ [In short, we have applied various radiological tests.]$_S$
EVIDENCE (N-S)	S	*[Presentaron datos clínicos de obstrucción intestinal 11 pacientes.]$_N$ [En todos ellos se observaron signos radiológicos de obstrucción.]$_S$*
	B	*[11 pazienteren hesteetako buxaduraren datu klinikoak aurkeztu ziren.]$_N$ [Horietan guztietan buxaduraren zeinu erradiologiakoak hauteman ziren.]$_S$*
	E	[11 patients presented clinical data of intestinal obstruction.]$_N$ [Radiological signs of obstruction were detected in all of them.]$_S$
INTERPRETATION (N-S)	S	*[La utilización de técnicas como el lavado gástrico, la endoscopia, la extracción manual transanal o el uso de laxantes por vía rectal para intentar extraer los paquetes aumenta el riesgo de rotura de los mismos,]$_N$ [por lo que se desaconseja su uso.]$_S$*
	B	*[Urdail-garbiketak, endoskopioak, ondeste-bideko eskuzko erauzketak edo ondeste-bideko laxanteen erabilerak paketeak apurtzeko arriskua handitzen dute.]$_N$ [zeinarengatik ez dira horien erabilera gomendatzen.]$_S$*
	E	[The use of techniques such as gastric lavage, endoscopy, manual transanal removal, or the use of rectal laxatives to try to extract the packages are factors that increase the risk of breaking them,]$_N$ [so we advise against their use.]$_S$
OTHERWISE (N-S)	S	*[Consideramos que el programa tenía cobertura total si incluía a todos los municipios;]$_N$ [si no, la cobertura del programa era considerada parcial.]$_S$*
	B	*[Programak kobertura osoa zuela kontsideratu dugu herri guztiak barnean biltzen bazituen;]$_N$ [bestela, programaren estaldura partzialtzat hartu izan da.]$_S$*
	E	[We consider that the program had full coverage if it included all municipalities;]$_N$ [if not, the program's coverage was considered as partial.]$_S$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
ANTITHESIS (N-S)	S	*[Uno de los factores que se asocian al suicidio es, precisamente, la enfermedad física.]*$_N$ *[Sin embargo, la existencia de una enfermedad física no constituye una evidencia incontrovertible de que éste sea el factor único, ni siquiera el más importante, en determinar el acto suicida.]*$_S$
	B	*[Buru-hiltzeari lotutako eragile bat, hain zuzen ere, gaixotasun fisikoa izaten da.]*$_N$ *[Hala ere, gaixotasun fisikoa ez da ez halabeharrezko arrazoia ez faktore bakarra, ezta garrantzitsuena ere buru-hiltzearen ekintza determinatzeko.]*$_S$
	E	[One of the factors that is associated with suicide is precisely the physical illness.]$_N$ [However, the existence of a physical illness is not an incontrovertible proof that this is the only factor, nor even the most important, for determining the suicidal act.]$_S$
ENABLEMENT (N-S)	S	*[Al paciente no solo se le ha de diagnosticar y tratar la infección.]*$_N$ *[Es necesario ofrecerle pautas para que dicha infección no vuelva a aparecer.]*$_S$
	B	*[Pazienteari diagnostikatzea eta infekzioa tratatzea ez da nahikoa.]*$_N$ *[Beharrezkoa da jarraibideak eskaintzea infekzioa berriz ager ez dadin.]*$_S$
	E	[It is not enough to diagnose and treat the infection of patients.]$_N$ [It is necessary to offer them guidelines in order to avoid the reappearance of this infection.]$_S$
CAUSE (N-S)	S	*[La psiconeuroinmunología es un nuevo campo de la ciencia que está emergiendo]*$_N$ *[debido a un número cada vez mayor de datos que demuestran interrelaciones entre funciones inmunes y psiconeurales.]*$_S$
	B	*[Psikoneuroinmunologia garatzen ari den zientziaren eremu berria da.]*$_N$ *[Izan ere, gero eta datu gehiagok frogatzen dute funtzio immuneen eta psikoneuralen arteko erlazioak.]*$_S$
	E	[Psychoneuroimmunology is a new field of science that is emerging]$_N$ [due to an increasing number of data that show interrelationships between immune functions and psychoneural functions.]$_S$
EVALUATION (N-S)	S	*[Hay trabajos que demuestran una mejoría en la distancia recorrida en la prueba de marcha debido al aprendizaje, sobre todo cuando las pruebas se repiten en un corto espacio de tiempo.]*$_N$ *[Teniendo esto en cuenta, puede considerarse que las pruebas de marcha son adecuadas para este tipo de estudios y reflejan el esfuerzo que el paciente hará en la vida cotidiana.]*$_S$
	B	*[Ikasketaren ondorioz ibilketa-proban ibilitako distantzian hobekuntza frogatzen duten lanak daude, batez ere denbora laburrean errepikatzen diren frogetan.]*$_N$ *[Hau kontuan izanik, pentsa daiteke ibilketa-probak ikasketa tipo hauentzat egokiak direla eta pazienteak eguneroko bizitzan egingo duen ahalegina erakusten dutela.]*$_S$
	E	[There are works that show that there is an improvement regarding the distance that is covered in walking tests due to a learning process, especially when the tests are repeated in a short space of time.]$_N$ [Bearing this in mind, we consider that walking tests are adequate for this type of study and they show the effort that patients would make in their daily living.]$_S$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
MOTIVATION (N-S)	S	*[En contraste con las numerosas propuestas terapéuticas, sorprende que la pérdida de peso, mediante una dieta alimentaria hipocalórica, aparezca en un segundo o tercer plano y sean muy escasas las publicaciones dedicadas, exclusivamente, a los resultados de la misma, máxime cuando la gran mayoría de los pacientes son obesos.]$_S$ [Por este motivo, nos hemos decidido a comunicar nuestra experiencia con la dieta hipocalórica como tratamiento único en pacientes afectos de OSAS.]$_N$*
	B	*Makina bat proposamen terapeutikorekin kontrastean, harrigarria da dieta hipokalorikoa bigarren edo hirugarren maila batean agertzea eta hain publikazio gutxi egotea proposamen horien datuei buruz; batez ere pazienteen gehiengoa pertsona gizenak direnean.]$_S$ [Zio horregatik, dieta hipokalorikoa tratamendu bakar gisa OSAS duten pazienteentzat izan dugun esperientzia komunikatzea erabaki dugu.]$_N$*
	E	[In contrast to the many therapeutic proposals, it is surprising that weight loss, by a hypocaloric diet, appears in second or third place and that there are very few publications dealing exclusively with its results, especially since most of the patients are obese.]$_S$ [For this reason, we have decided to report our experience with hypocaloric diet as monotherapy in patients with OSAS.]$_N$
PREPARATION (N-S)	S	*[Pacientes y métodos.]$_S$ [Los 257 pacientes estudiados constituyen el 5% seleccionado de un total de 4.850 que se visitaron en la unidad de interconsulta psiquiátrica del Hospital Clínic i Provincial (HCP) de Barcelona desde junio de 1984 a junio de 1990.]$_N$*
	B	*[Pazienteak eta metodoak.]$_S$ [1984ko ekainetik 1990eko ekainerarte Bartzelonako Hospital Clínic i Provincial (HCP) psikiatria sail arteko unitatean bisitatu ziren 4.850 pazientetik % 5 osatzen dute azterturiko 257 pazienteak.]$_N$*
	E	[Patients and methods.]$_S$ [The 257 studied patients constitute the 5% of 4850 that visited the consultation-*liaison psychiatry* unit of the Hospital Clinic i Provincial (HCP) in Barcelona from June 1984 to June 1990.]$_N$
SOLUTION (N-S)	S	*[Además de los problemas de infraestructura y de su mayor coste otro inconveniente de las fuentes portátiles es su corta autonomía.]$_N$ [En este sentido, se han diseñado diversos dispositivos destinados a economizar oxígeno manteniendo un aporte de gas suficiente.]$_S$*
	B	*[Azpiegitura arazoez eta hauen kosteez gain iturri eramangarrien beste eragozpen bat autonomia eskasia da.]$_N$ [Hori dela eta, gas hornikuntza nahikoa mantentzen duten oxigenoa aurrezteko zenbait gailu diseinatu dira.]$_S$*
	E	[In addition to infrastructure problems and their greater cost, another disadvantage of portable sources is their short autonomy.]$_N$ [In that sense, various devices have been designed to save oxygen and maintain an adequate gas supply.]$_S$

(Continued)

Appendix Table 2. (Continued)

Relation		Example
MEANS (N-S)	S	*[Las tasas de mortalidad por muerte cardíaca súbita pueden reducirse,]$_N$ [entre otros factores, por la correcta identificación de los pacientes con riesgo de sufrirla, por la rapidez con que se realicen las maniobras de reanimación y por la calidad del traslado a centros especializados.]$_S$*
	B	*[Bat-bateko heriotza kardiakoaren heriotza-tasak murritz daitezke,]$_N$ [beste faktore batzuen artean, sufritzeko arriskua duten pazienteen identifikazio zehatzari esker, suspertze eragiketak buruturiko bizkortasunari esker eta gune espezializatuetarako lekualdaketa kalitateari esker.]$_S$*
	E	[Mortality rates due to sudden cardiac death can be reduced,]$_N$ [among other factors, by the correct identification of patients at risk of suffering it, by the speed of the resuscitation and by the quality of the move to specialized centers.]$_S$
UNCONDITIONAL (N-S)	S	*[Parece que la administración de este medicamento tiene efectos adversos,]$_N$ [aun incluso si se administra la dosis mínima.]$_S$*
	B	*[Botika hau hartzeak aurkako eraginak dituela dirudi,]$_N$ [nahiz eta dosi txikiena emanda ere.]$_S$*
	E	[It seems that the administration of this drug has adverse effects]$_N$ [even if the minimum dose is given.]$_S$
UNLESS (N-S)	S	*[Los terapeutas deben admitir a cualquier paciente en el grupo,]$_N$ [a no ser que éste presente signos claros de actitud violenta que puedan perjudicar el correcto desarrollo de la terapia.]$_S$*
	B	*[Terapeutek edozein paziente onartu behar dute taldean,]$_N$ [non eta honen jarrera bortitzak ez duen terapiaren garapen zuzena kaltetzen.]$_S$*
	E	[Therapists must accept any patient in the group]$_N$ [unless he presents clear signs of violent behaviour that could harm the therapy success.]$_S$

[a]In the second column, 'S' means Spanish, 'B' means Basque and 'E' means English.

Iria da Cunha Fanego holds a Hispanic Philology degree at the University of Santiago de Compostela, Spain and a PhD on Applied Linguistics at the Pompeu Fabra University (UPF), Spain. She was Assistant Professor at the UPF and researcher of the Institute for Applied Linguistics until 2008. At present, she holds a postdoctoral grant awarded by the Spanish Ministry of Science and Innovation to work at the Laboratoire Informatique d'Avignon, France. Her research fields are automatic summarization, discourse parsing and analysis of specialized discourse.

Mikel Iruskieta holds a Basque Philology degree at the University of the Basque Country (UPV/EHU). Since 2004, he has been a member of the IXA Research Group (Natural Language Processing Group) at the Faculty of Informatics (UPV/EHU), where he is doing a PhD on Applied Linguistics. He has also been professor of Basque at the same university since 2008. His research fields are semantic, syntactic and discourse parsing, and development of linguistic resources for Basque.

9

A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora

Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

A Qualitative Comparison Method for Rhetorical Structures

Identifying different discourse structures in multilingual corpora

Mikel Iruskieta · Iria da Cunha · Maite Taboada

Received: date / Accepted: date

Abstract Explaining why the same passage may have different rhetorical structures when conveyed in different languages remains an open question. Starting from a trilingual translation corpus, this paper aims to provide a new qualitative method for the comparison of rhetorical structures in different languages and to specify why translated texts may differ in their rhetorical structures. To achieve these aims we have carried out a contrastive analysis, comparing a corpus of parallel English, Spanish and Basque texts, using Rhetorical Structure Theory (RST). We propose a method to describe the main linguistic differences among the rhetorical structures of the three languages in the two annotation stages (segmentation and rhetorical analysis). We show a new type of comparison that has important advantages with regard to the quantitative method usually employed: it provides an accurate measurement of inter-annotator agreement, and it pinpoints sources of disagreement among annotators. With the use of this new method, we show how translation strategies affect discourse structure.

Keywords Annotation Evaluation · Discourse Analysis · Rhetorical Structure Theory · Translation Strategies

M. Iruskieta
Teacher Training College Bilbao, Sarriena auzoa z/g, 48940 Leioa (Basque Country), Spain
Tel.: +0034-94601-7569
E-mail: mikel.iruskieta@ehu.es

I. da Cunha
University Institute for Applied Linguistics, Universitat Pompeu Fabra, C/ Roc Boronat 138, 08018, Barcelona, Spain
Tel.: +34-93-542-1187
E-mail: iria.dacunha@upf.edu

M. Taboada
Department of Linguistics, Simon Fraser University, 8888 University Dr., Burnaby, B.C., V5A 1S6, Canada
Tel.:+1-778-782-5585
E-mail: mtaboada@sfu.ca

1 Introduction

Translation or parallel corpora on the one hand and comparable corpora on the other are useful in many tasks, in applied linguistics and in Natural Language Processing. Compiling such corpora can provide insight into translation strategies, can help validate or disprove intuitions about differences across languages, and can be useful in computational applications such as machine translation or terminology extraction.

Translation corpora have been useful in testing hypotheses about language contrasts. Granger [2003], for instance, using translation corpora, put into question the over-generalization that "French favors explicit linking while English tends to leave links implicit". Translation corpora also help identify strategies used in the translation process, such as the strategy that Xiao [2010] found in translated Chinese texts, where there was an increased use of discourse markers, presumably to more clearly identify the rhetorical structure of the text (although introducing discourse markers may lead to subtle changes in rhetorical structure as well, in cases when the translator interprets a different relation than that intended by the original author).

Most contrastive corpus-based studies emphasize surface-level aspects of language, such as differences in terminology in general [Gomez and Simoes, 2009; Morin et al, 2007; Fung, 1995; Wu and Xia, 1994] and specific lexical items in particular [Fetzer and Johansson, 2010; Flowerdew, 2010]; differences in aspects of modality [Kanté, 2010; Usoniene and Soliene, 2010]; or the use of discourse markers [Mortier and Degand, 2009]. There exists, however, a sizeable body of work on differences in the rhetorical structure of texts across languages, in particular within the framework of Rhetorical Structure Theory (RST), a theory of text structure proposed by Mann and Thompson [1988]. The first contrastive RST study comparing one European language and one Asian language was carried out by Cui [1986], who compared English and Chinese expository rhetorical structures. Kong [1998] and Ramsay [2000, 2001] studied the same pair of languages, in both cases examining specific genres (business request letters and news texts). Other pairs of languages studied within RST include Arabic and English [Mohamed and Omer, 1999], Japanese and English [Marcu et al, 2000], or a range of European languages, such as Dutch-English [Abelen et al, 1993], Finnish-English [Sarjala, 1994], French-English [Delin et al, 1996; Salkie and Oates, 1999], Spanish-English [Taboada, 2004a,b], and Spanish-Basque [da Cunha and Iruskieta, 2010].

Contrastive studies comparing the rhetorical structures of more than two languages are not very common, although we can mention the study in Portuguese-French-English by Scott et al [1998]. They show a methodology to carry out RST contrastive analysis of instructional texts in different languages, and they present the results of an empirical cross-lingual experiment based on this methodology. More information about contrastive RST studies or studies about other languages can be found in Taboada and Mann [2006a,b].

One observation in RST-based work is that the same passage, when conveyed in two different languages, may have different underlying rhetorical structures [Bateman and Rondhuis, 1997; Delin et al, 1994]. An explanation for such differences is that translation strategies reorganize the structure of the discourse, with the resulting underlying structures being different. The translation literature deals with many as-

pects of this phenomenon, one being differences in explicitness, which in some cases result in different underlying structures [House, 2004].

This proposal (that translation strategies lead to different structures) is often presented on the basis of individual examples, with no unifying principle for the representation of underlying structure. In this paper, we present a new method for the evaluation of discourse structures across multiple languages to analyze which translation strategies affect rhetorical structure.

The first aim of this paper is to provide a new qualitative method to compare rhetorical structures in different languages and/or by different annotators. Existing work comparing different annotations uses a quantitative methodology [Marcu, 2000a]. The main comparison methodology consists of quantifying the agreement between the rhetorical analyzes by annotators, in terms of Elementary Discourse Units (EDUs), spans (sets of related EDUs), nuclearity (nucleus or satellite role of a span) and rhetorical relations (set of hypotactic and paratactic relations). To compare rhetorical analyzes, typical precision and recall measures are used. Work by da Cunha and Iruskieta [2010] and van der Vliet [2010] presents some criticisms of Marcu's methods, arguing that this quantitative method amalgamates agreement coming from different sources, because decisions at one level in the tree structure affect decisions and factors at other levels, with the result that the factors are not independent. Disagreement on segmentation or attachment point at lower levels in the tree significantly affects agreement on the upper rhetorical relations in a tree, and should be accounted for separately. Mitocariu et al [2013] have proposed an evaluation method (for RST and Veins Theory) which checks the inner nodes[1] (attachment point), nuclearity of the relation (nuclearity) and the vein expressions or constitution of the units (constituent [Marcu, 2000a]) but excludes the names of relations as a comparison criterion. In our evaluation method we consider Mitocariu et al's factors (attachment point, constituent and nuclearity) and the rhetorical relations. We believe that the qualitative method that we present here addresses the deficiencies in previous proposals and provides a qualitative description of annotation dispersion, while at the same time allowing for quantitative evaluation.

The second aim of this paper is to propose this method and to test this. In order to detect differences among rhetorical structures and study the origin of such differences, we analyze a corpus of parallel texts in three different languages: English, a Germanic language; Spanish, a Romance language; and Basque, a non-Indo-European language. We investigate whether differences are motivated by different translation strategies or by the choice of one relation over another in a group of similar relations, as Stede [2008b] proposes. Our corpus, albeit small, is comparable to the only other trilingual comparative corpus [Scott et al, 1998, 11], and it is rich enough to allow the development and evaluation of a qualitative comparison method for rhetorical relations.

Our study is useful from a theoretical point of view, because it will help us understand how the rhetorical structures of texts in different languages are constructed. Moreover, the study provides rhetorical analyzes of a less-commonly studied lan-

[1] Soricut and Marcu [2003, pg. 152] use the term "attachment point" or "dominance set".

guage,[2] Basque, the only pre-Indo-European language of Western Europe [Trask, 1997] and one of the four official languages of Spain (together with Catalan, Galician and Spanish), spoken in the Basque country. From an applied point of view, this work supports the development of computational linguistics systems (such as summarization, information extraction and retrieval systems), where accurate annotation is of paramount importance. In addition, our methodology can be useful in research on automatic compilation of specialized corpora, and can help professional translators and machine translation researchers.

The paper is organized as follows: Section 2 presents the methodology and theoretical background of our study. Section 3 describes our methodological proposal and provides the results of the discourse analysis of our corpus. Section 4 provides conclusions and proposals for future work.

2 Methodology

Our work consisted of three stages. First, we decided on the theoretical framework of our study, RST. Second, we built the corpus. Finally, we carried out the analysis, including a comparison of the three different RST structures for each text, using both: a quantitative methodology and our proposed new qualitative methodology.

2.1 Theoretical Framework

In this study, we use RST, since it is a language-independent theory. RST is a descriptive theory for textual organization that characterizes text structure using relations among the discourse or rhetorical elements a text contains. These elements are called spans, and they can be nucleus (if the element is more essential to the speaker's purpose) or satellite (if it provides some rhetorical information about the nucleus). The relations can be: a) nuclear relations (e.g., ANTITHESIS, CAUSE, CIRCUMSTANCE, CONDITION, ELABORATION, EVIDENCE, JUSTIFICATION, MOTIVATION, PURPOSE), that is, hypotactic relations between nuclei and satellites, and b) multinuclear relations (e.g., CONTRAST, JOINT, LIST, SEQUENCE), that is, paratactic relations among nuclei, where more than one unit is central with regard to the author's purposes. For a more detailed explanation of RST, see Mann and Thompson [1988] and the RST web site by Mann and Taboada [2010].

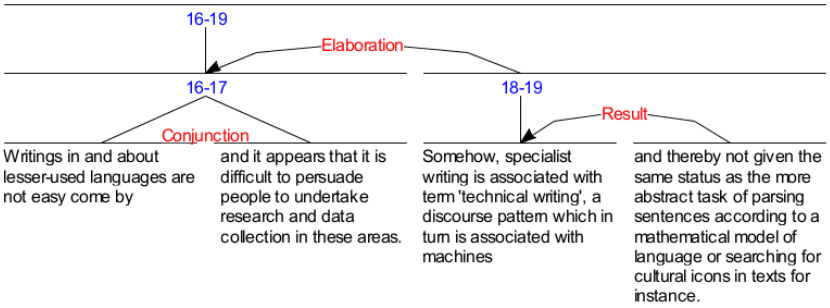
RST relations are typically represented as trees. Figure 1 shows a fragment of an RST tree,[3] with one multinuclear relation (CONJUNCTION) and two multinuclear relations (RESULT and ELABORATION). The annotator recognized that spans 16 and 17

[2] Although great efforts have been made to stimulate Machine Translation studies for different language pairs, non-official languages that are typologically different and could be interesting are not considered. For example Koehn [2005] presents a 30 million word corpus translated to the 11 official of the European Union: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish to study different language pairs translations, but less common languages spoken in the EU are not included.

[3] The source of the text (TERM#_original language) is shown in square brackets at the end of the figures, tables or examples.

are conjoined, forming another span where each item has a comparable role (moreover, each span has a verb *are* and *appears*, and they are linked by the connector *and*). The annotator also found a RESULT relation, since she understood that span 18 could be the cause for the situation explained into the span 19 (again, each unit has a finite verb: *is associated* and [*is*] *given*, and they are linked by the double connector *and thereby*). It is important to observe that rhetorical relations are applied recursively, i.e., spans that stand in a relation: such as 18 and 19 in Figure 1 form a new span (18-19) that can enter into new relations, such as the ELABORATION relation. In this case, the annotator labelled this relation as such because the span made up of units 18-19 (satellite) provides additional information about the previous span (16-17), which constitutes the nucleus of the relation. Following Marcu's [2000b] strong compositionality criteria, the most important units for the 16-19 span are 16 and 17. For the span 18-19 the most important unit is 18.

Fig. 1 Example of an RST tree, TERM30_ENG



In the literature on RST, there is agreement that the most important unit of the tree is the "central unit(s)" [Stede, 2008b] and the most important unit of a span is the "central subconstituent" [Egg and Redeker, 2010]. So following this framework we will use the term "Central Unit(s)" (CU) for the most important unit of an RS-tree and "Central Proposition(s)" (CP) for the most important unit of a span.

Table 1 provides a representation of this example.

Relation	Left Span	Right Span	CP	Nuclearity
RESULT	18	19	-	NS
CONJUNCTION	16	17	16-17	NN
ELABORATION	16-17	18-19	18	NS

Table 1 Formalization of Figure 1, TERM30_ENG

There are several classifications of RST relations: the classic one by Mann and Thompson of 24 relations [Mann and Thompson, 1988], the extended one by Mann

and Thompson of 30 relations, available on the RST site [Mann and Taboada, 2010], and Marcu's classification of 78 relations [Carlson et al, 2003], among others. We have chosen the extended classification for the annotation of our trilingual corpus. Space constraints preclude an extensive discussion of its merits over other approaches [see Taboada and Mann, 2006a, for a discussion].

2.2 Corpus

As Granger [2003] proposes, a multilingual translation corpus is:

> [...] the most obvious meeting point between CL (Contrastive Linguistics) and TS (Translation Strategies). Researchers in both fields use the same resource but to different ends: uncovering differences and similarities between two (or more) languages for CL and capturing the distinctive features of the translation process and product for TS.
>
> [Granger, 2003, pg. 22]

In translation studies where the intention is to study similarities and differences in large corpus studies it is difficult to find a balanced corpus in size and similar composition of genres [Baker, 2004]. Our problem was to find a balanced multidirectional corpus of such size that allowed for a manual comparison of all the rhetorical structures by language pair. One of our aims, as we said, is to propose a methodology to describe when a different RST relation can be attributed to annotator interpretation or to different language forms.

As far as we know, no multilingual corpus with English, Spanish and Basque texts exists. Our corpus was then compiled specifically for this work.[4] It is a multidirectional translation corpus which contains abstracts of research papers published in the proceedings of the International Conference about Terminology that took place in Donostia and Gasteiz in 1997 [UZEI and HAEE-IVAP, 1997]. In this conference, authors were allowed to send full papers in English, French, Spanish or Basque, but they had to provide titles and abstracts in the four languages. In order to have a multidirectional and trilingual balanced corpus, we have chosen abstracts for which the original paper was written in English (five texts), Spanish (five texts) and Basque (five texts). Thus, we have analyzed 15 abstracts (the same ones for each language), written by different authors, constituting three subcorpora. Table 2 summarizes the statistics of the subcorpora.

In order to find correlations between translation strategies and rhetorical relations, a methodology that can compare parallel rhetorical structures is needed. We built our corpus in order to develop such a methodology, and consider that the number of texts is sufficient for the design of the qualitative method that we present. This qualitative method applies to any type of text,[5] since the principles on which it is based are

[4] A problem with work in the framework of RST is that there is no annotated bilingual or trilingual corpus to study the effects of translation strategies on rhetorical structure. As a consequence, a researcher in such situation first needs to learn RST and perform annotations, as Maxwell [2010] suggests.

[5] It was used also to evaluate the RST Basque TreeBank [Iruskieta et al, 2013a], available at: http://ixa2.si.ehu.es/diskurtsoa/en/.

Subcorpus	Annotators	Texts	Words	Sentences	EDUs
ENG	A1	15	5706	201	318
SPA	A2	15	6324	193	318
BSQ	A3	15	4800	197	318

Table 2 Corpus statistics

general RST-based principles. We believe that the analysis is general enough and the method applicable across genres. We also discuss some examples detected with the qualitative evaluation in this parallel corpus that show how translation strategies could be related to rhetorical structures (see Subsection 3.2.2).

After the corpus compilation, we carried out the analysis. This analysis had two main phases: discourse segmentation and rhetorical analysis.

2.3 Discourse Segmentation

The first step in analyzing texts under RST consists of segmenting the text into spans. Exactly what a span is, under RST, and more generally in discourse, is a well-debated topic. RST [Mann and Thompson, 1988] proposes that spans, the minimal units of discourse —later called elementary discourse units (EDUs) [Marcu, 2000a]— are clauses, but that other definitions of units are possible.

From our point of view, adjunct clauses stand in clear rhetorical relations (cause, condition, concession, etc.). Complement clauses, however, have a syntactic, but not discourse, relation to their host clause. Complement clauses include, as Mann and Thompson [1988] point out, subject and object clauses, and restrictive relative clauses, but also embedded report complements, which are, strictly speaking, also object clauses.

Other possibilities for segmentation exist; one of the better-known ones is the proposal by Carlson et al [2003] for segmentation of the RST Discourse Treebank [Carlson et al, 2002]. Carlson et al [2003] propose a much more fine-grained segmentation, where report complements, relative clauses and appositive elements constitute their own EDUs.

In our work three annotators segmented the EDUs of each corpus (A1 segmented English texts, A2 segmented Spanish texts, and A3 segmented Basque texts).[6]

These annotators are experts in RST, having carried out research in this field for a number of years, and they have participated in several projects related to the design

[6] When a corpus is annotated only with one annotator per language, the results may yield subjective idiosyncrasies. This is not a problem for the aim of this paper, because we do not want to provide a reliable annotated corpus in three languages, but we do provide a qualitative way to compare annotation in different languages. Comparisons have been done manually and by pairs of languages following two different evaluations: a) Marcu's quantitative method and b) a new qualitative-quantitative method. So even the corpus is small, comparison work is huge. The aim to provide reliable corpora has been achieved in other papers by the authors (English SFU corpus [Taboada and Renkema, 2008], Spanish RST TreeBank [da Cunha et al, 2011] and Basque RST TreeBank [Iruskieta et al, 2013a]).

and elaboration of RST corpora in the three languages under consideration. Annotators performed this segmentation task separately and without contact among them. In our segmentation, we follow the general guidelines proposed by Mann and Thompson [1988] which we have operationalized for this paper. We detail the principles below.

Every EDU Should Have a Verb. In general, EDUs should contain a (finite) verb. The main exception to this rule is the case of titles, which are always EDUs, whether they contain a verb or not. Non-finite verbs form their own EDUs only when introducing an adjunct clause (but not a modifier clause; see APPENDIX A for a detailed explanation).

Coordination and Ellipsis. Coordinated clauses are separated into two segments, including cases where the subject is elliptical in the second clause. In Spanish and Basque, both pro-drop languages, this is in fact the default for both first and second clause, and therefore we see no reason why a clause with a pro-drop subject cannot be an independent unit. We follow the same principle for English.

Coordinated verb phrases (VPs) or verbs do not constitute their own EDUs. We differentiate coordinated clauses from coordinated VPs because the former can be independent clauses with the repetition of a subject; the latter, in the second part of the coordination, typically contain elliptical verbal forms, most frequently a finite verb or modal auxiliary.

Relative, Modifying and Appositive Clauses. We do not consider that relative clauses (whether restrictive or non-restrictive), clauses modifying a noun or adjective, or appositive clauses constitute their own EDUs. We include them as part of the same segment together with the element that they are modifying. This departs from RST practice, where (restrictive) relative clauses are often independent spans, as seen in many of the examples in the original literature and the analyzes on the RST web site [Mann and Thompson, 1988; Mann and Taboada, 2010]. We found that relative clauses and other modifiers often lead to truncated EDUs, resulting in repeated use of the SAME-UNIT label,[7] and thus decided that it was best to not elevate them to the status of independent segments.

Parentheticals. The same principle applies to parentheticals and other units typographically marked as separate from the main text (with parentheses or dashes). They do not form an individual span if they modify a noun or adjective, but they do if they are independent units, with a finite verb.

Reported Speech. We believe that reported and quoted speech do not stand in rhetorical relations to the reporting units that introduce them, and thus should not constitute separate EDUs, also following clear arguments presented elsewhere [da Cunha and

[7] See the paragraph on Truncated EDUs in this section.

Iruskieta, 2010; Stede, 2008a]. This is in contrast to the approach in the RST Discourse Treebank [Carlson et al, 2003], where reported speech (there named ATTRIBUTION) is a separated EDU. There are, in any case, no examples of reported speech in our corpus.

Truncated EDUs. In some cases, a unit contains a parenthetical or inserted unit, breaking it into two separate parts, which do not have any particular rhetorical relation between each other. In those cases, we make use of a non-relation label, Same-unit, proposed for the RST Discourse Treebank [Carlson et al, 2003].

Once our segmentation criteria were established and the three annotators carried out the segmentation, the three segmentations were compared in terms of F-measure and Kappa. In this way, we quantified agreement and disagreement across segmentations. Moreover, we analyzed the main causes of the disagreements. Results are shown in Subsection 3.1. After the segmentation agreement evaluation, we harmonized the segmentation, ensuring that units were comparable across the languages. At this point, we also calculated linguistic distance between the pairs of languages, by calculating which language required the most changes in the harmonization process. This harmonization process was necessary to start out the analysis with similar units, and to avoid confusing analysis disagreement and segmentation agreement. Marcu et al [2000] and Ghorbel et al [2001] also align (which we termed harmonize) their texts, decreasing the granularity of their segmentation to avoid complexity. With this decision, we lose some rhetorical information at the most detailed level of the tree. This does not, however, affect higher levels of tree structure. The results of this harmonization are shown in Subsection 3.1.1.

2.4 Rhetorical Analysis

Starting from the same discourse segmentation, we carried out the discourse annotation of our corpus. Once again, A1 annotated English texts, A2 annotated Spanish texts and A3 annotated Basque texts, using the mentioned extended discourse relations set and RSTTool [O'Donnell, 2000]. We compared the resulting rhetorical trees using two different evaluation methods. One of them, which we characterize as a quantitative evaluation, was proposed by Marcu [2000a], and the other one, which we describe as qualitative evaluation, was developed by our research team.

A qualitative comparison method for rhetorical structures in multilingual corpora should quantify data, but also (and more importantly) should show linguistic features affecting rhetorical structure. The quantitative/qualitative distinction is due to the fact that the first method only gives us an approximate measure of agreement, whereas the second method provides a qualitative description of annotation dispersion. The qualitative evaluation, in addition to its use as a measure of inter-annotator agreement, can also be deployed to evaluate discourse structures built by a parser.

2.4.1 Quantitative Evaluation

In this section we present the quantitative method of Marcu [2000a] and its limitations, already pointed out in other works [van der Vliet, 2010; da Cunha and Iruskieta, 2010; Iruskieta et al, 2013b]. The main limitations are:

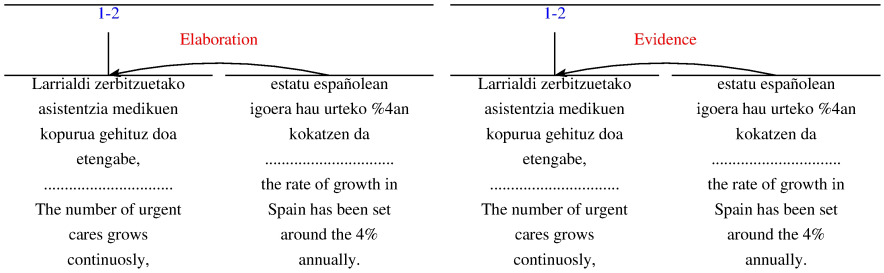
i) Two of the factors evaluated, nuclearity and relation, are not independent of each other: factor conflation.

ii) The description of comparison and weight given to the agreement in certain rhetorical relations could be improved: deficiencies in the description.

Marcu [2000a] presented a method to evaluate the correctness of discourse trees, comparing automatically-built trees with manually-built ones. This method measures recall and precision according to four factors: elementary discourse units (EDU), units linked with relations (Span), nuclear or satellite position (Nuclearity) and rhetorical meaning of units (Relation). We refer to this method as the quantitative method, because it uses exclusively numerical measures.

i) *Factor conflation: nuclearity and relations.* When measuring the relation factor, the quantitative method conflates the label SPAN with a relation. Thus, the SPAN label carries the same weight as any other relation. As we can see in Figure 2, one of the annotators has labelled the relation as ELABORATION, and the other as EVIDENCE.

Fig. 2 Quantitative evaluation: factor conflation [Iruskieta et al, 2013a, GMB0401]



If we describe such disagreement with the quantitative method, we can see that there is a degree of agreement with respect to the relation in the Figure 3, when in fact the agreement captured is simply the agreement in nuclearity, that is, in SPAN. Figure 3 shows the results obtained after the comparison of the two rhetorical structures included in Figure 2 by using the quantitative evaluation. These results have been obtained automatically by using RSTeval, which is an implementation of Marcu's comparison method.[8]

As we have pointed out in Section 1, RSTeval applies this method automatically and does not take into account the language of the rhetorical structures; however, it

[8] This evaluation method has been automated by Maziero and Pardo [2009] and it can be used in four languages: English, Spanish, Portuguese and Basque. Available at http://www.nilc.icmc.usp.br/rsteval/

eliminates the stopwords from the text, which are not used to build the EDUs and Spans. In the first table of Figure 3, absolute matches between structures can be observed (e.g. Units: Matches = 2 of 2), as well as percentages (e.g. Units: Recall = 1 / Precision = 1), for the four mentioned factors. The second table of Figure 2 shows the detailed comparison process, where all the constituents of the structures are included. In this case, the first constituent corresponds to the first EDU, that is, words from "1 to 8" in the text; the second constituent corresponds to the second EDU, that is, words from "9 to 13"; and the third constituent corresponds to the Span formed by the two mentioned EDUs, that is, words from "1 to 13" (the exclamation point means that the constituent is a Span). The symbol "x" indicates that a Unit or Span is included in the corresponding rhetorical structure; "n" means nucleus; "s" means satellite, and "r" refers to the biggest span, that is, the span including the complete text. In the Relations factor, if there is a nucleus, the category "span" is included when a nuclear relation is under consideration or the name of relation when a multinuclear relation is under consideration, while, if there is a satellite, the name of the corresponding rhetorical relation is included.

Fig. 3 Quantitative evaluation of Figure 2 with RSTeval

RSTeval Tool for discourse parsing evaluation

This tool provides an automatic method to compare two RST structures, one made by a human being (the ideal structure) and another made by an automatic system.

Evaluation ID: Euskara

Text	Units			Span			Nuclearity			Relation		
ID	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision	Matches	Recall	Precision
ex-la	2 of 2	1	1	3 of 3	1	1	3 of 3	1	1	2 of 3	0.666666666666667	0.666666666666667

Evaluation Table

Constituent	Units		Spans		Nuclearity		Relations	
	Manual	Auto	Manual	Auto	Manual	Auto	Manual	Auto
1 to 8 (Larrialdi_zerbitzuetako...etengabe)	x	x	x	x	n	n	span	span
9 to 13 (españolean_igoera...da)	x	x	x	x	s	s	elaborazioa	ebidentzia
!1 to 13 (Larrialdi_zerbitzuetako...da)			x	x	r	r	span	span

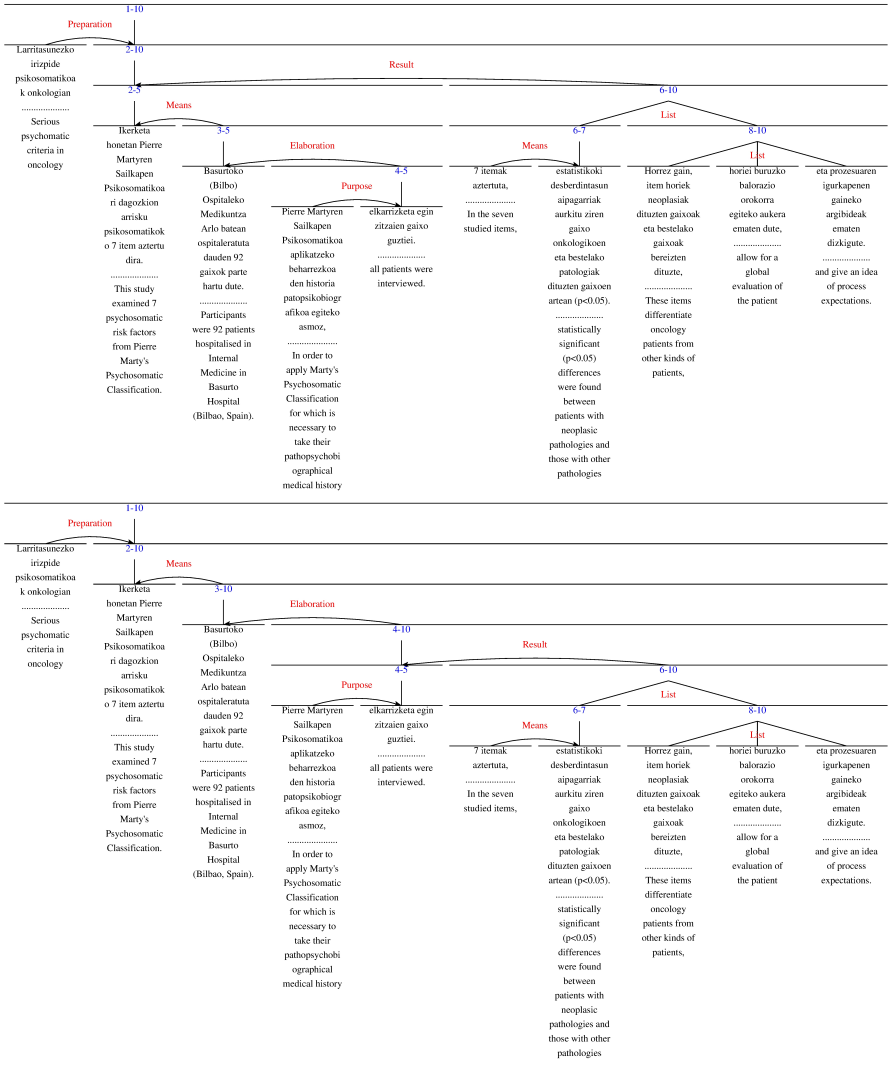
We can see another example, now of a real text in Figure 4.

In Table 3 we can see how *RSTeval* describes the agreement. The agreement levels are shown in Table 4. For ease of reference, we have highlighted the disagreements in gray.

When examining the rhetorical relations factor, we can see that the SPAN label plays a role in the description of agreement levels in Table 4: F-measure: 0.842 (16 agreements out of 19). If we describe the agreement without the SPAN label, however, the degree of agreement changes, as we can see in Table 5: F-measure: 0.778 (7 agreements out of 9).

ii) *Deficiencies in the description.* When annotators decide that a relation has an attachment point at different levels in the tree structure [da Cunha and Iruskieta, 2010],

Fig. 4 Interpretations of text GMB0701 [Iruskieta et al, 2013a]



the method proposed by Marcu [2000a] is not able to compare the relations where constituents has changed. Observe the following in Figure 4:

1) In Table 3 the agreement in the ELABORATION relation cannot be included, because the relation has different *spans*: in A3 '23 to 31' and in A4 '!23 to 65' both attachments are referred as the same constituent, '23 to 31'.
2) The MEANS constituent of A3 '!16 to 35' and in A4 of '!16 to 65', both attach to the same EDU (EDU2 or '5 to 15'), but since the constituents do not coincide, the two MEANS relations cannot be compared.

EDU	Constituent	Units		Spans		N/S		Relations	
		A3	A4	A3	A4	A3	A4	A3	A4
1	1 to 4 (Larritasunezko_irizpide... onkologian)	x	x	x	x	s	s	preparation	preparation
2	5 to 15 (Ikerketa_Pierre... aztertu)	x	x	x	x	n	n	span	span
3	16 to 22 (Basurtoko_Ospitaleko... gaixok)	x	x	x	x	n	n	span	span
4	23 to 31 (Pierre_Martyren... asmoz)	x	x	x	x	s	s	purpose	purpose
5	32 to 35 (elkarrizketa_zitzaien... guztiei)	x	x	x	x	n	n	span	span
4-5	!23 to 35 (Pierre_Martyren... guztiei)			x	x	s	n	elaboration	span
6	36 to 38 (7_itemak... aztertuta)	x	x	x	x	s	s	means	means
7	39 to 50 (estatistikoki_desberdintasun... 05)	x	x	x	x	n	n	span	span
6-7	!36 to 50 (7_itemak... 05)			x	x	n	n	list	list
8	51 to 57 (Horrez_item... bereizten)	x	x	x	x	n	n	list	list
9	58 to 60 (horiei_balorazio... orokorra)	x	x	x	x	n	n	list	list
8-9	!51 to 60 (Horrez_item... orokorra)			x	x	n	n	list	list
10	61 to 65 (prozesuaren_igurkapenen... dizkigute)	x	x	x	x	n	n	list	list
8-10	!51 to 65 (Horrez_item... dizkigute)			x	x	n	n	list	list
6-10	!36 to 65 (7_itemak... dizkigute)			x	x	s	s	result	result
4-10	!23 to 65 (Pierre_Martyren... dizkigute)				x		s		elaboration
3-10	!16 to 65 (Basurtoko_Ospitaleko... dizkigute)				x		s		means
2-10	!5 to 65 (Ikerketa_Pierre... dizkigute)			x	x	n	n	span	span
1-10	!1 to 65 (Larritasunezko_irizpide... dizkigute)			x	x	r	r	span	span
3-5	!16 to 35 (Basurtoko_Ospitaleko... guztiei)			x		s		means	
2-5	!5 to 35 (Ikerketa_Pierre... guztiei)			x		n		span	

Table 3 Qualitative method for text GMB0701

Units			Spans			N-S			Relations		
Match	R	P	Match	R	P	Match	R	P	Match	R	P
10 of 10	1	1	17 of 19	0.895	0.895	16 of 19	0.842	0.842	16 of 19	0.842	0.842

Table 4 Quantitative method: agreement level for text GMB0701

Relations		
Match	R	P
7 of 9	0.778	0.778

Table 5 Agreement level according to rhetorical relations in GMB0701

Following da Cunha and Iruskieta [2010], Iruskieta et al [2013b] and Mitocariu et al [2013], we think that a qualitative method should describe the six factors involved in all rhetorical relations independently: EDU and Span (segmentation), nucleus-satellite function (Nuclearity), and attachment point, constituent and rhetorical meaning (Relation). When parallel texts are compared, a qualitative method should take in account whether the language form is parallel, as explained in the next section.

2.4.2 Qualitative Evaluation

The qualitative evaluation method that we propose considers both type of agreement and source of disagreement, which results in a better explanation of the dispersion in annotator interpretations about text structure. When analyzing rhetorical structures using Marcu's method, we observed that similar structures at the intermediate level of a tree structure spans could not be compared, because the constituents did not coincide. Such structures had, however, the same rhetorical relation, and the fact that the relation is the same should be reflected in a measure of agreement. If we accept that constituents do not need to coincide in their (span size) entirety to be compared,

the issue is whether we can state that there is agreement with respect to the rhetorical relation, but disagreement about the constituents.

In our evaluation method it is not necessary for the constituents to be compared to be identical, like Marcu's [2000b]; only the central proposition (CP) has to be the same.[9] With such restriction we are able to compare rhetorical relations, using four independent criteria: constituent, attachment point, the direction of the relation (nuclearity) and effect of the relation.

When comparing RST structures with independent factors, we do not use typical nucleus and satellite terms to describe the extension of spans, because our method assesses independently nuclearity and unit size. The comparison in our method is based on rhetorical relations and no in spans of relations as Marcu's [2000b] method does. In our method we have a line for each relation, while in Marcu's [2000b] method there are two lines for each relation. The term constituent (C) refers to the length of the constituents, and the term attachment point (A) refers at the height of the tree where the constituent is linked (in Marcu's [2000b] evaluation method this factor is not considered, because what is compared are spans of relations). Because we are comparing relations and no spans of relations, in our comparison also nuclearity has a different meaning; while in Marcu's [2000b] method nuclearity has two possible values (S or N, where S means satellite and N means nucleus) for each span, in our method nuclearity has three values (SN, NN and NS) for each relation.

First of all, we present the types of agreement, and the two sources of disagreement in the qualitative evaluation by comparing annotators' RST trees.

We measure the agreement in rhetorical relations based on the following factors: constituent (C), attachment point (A) and the name of relation (R), checking some agreement types:

1. Agreement in relation, constituent and attachment point (**RCA**).
2. Agreement in relation and constituent (**RC**).
3. Agreement in relation and attachment point (**RA**).
4. Agreement only in relation (**R**).

A decision tree from Figure 5 formalizes the method to check the agreement types in rhetorical relations. As we mentioned before, to check agreement in rhetorical relation, the constituent of this relation must have the same central proposition (CP). If this condition is fulfilled then we check if both relation name (R), constituent (C) and attachment point (A) are exactly the same.

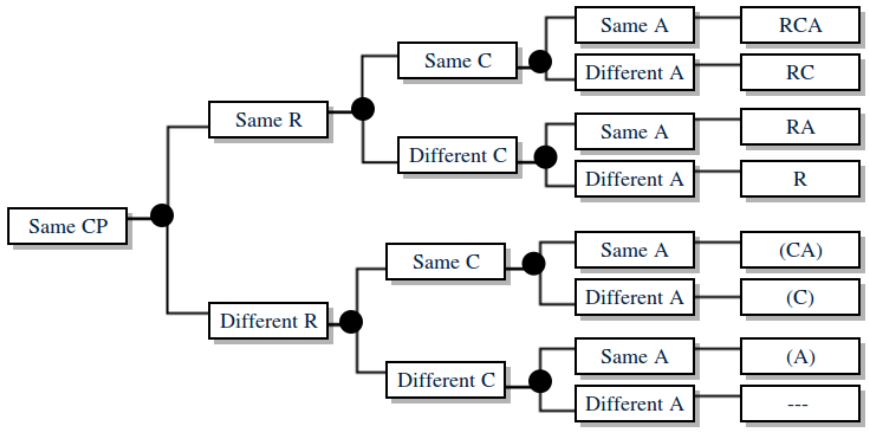
We distinguish two sources of disagreement, disagreements of type A and type L, for Annotator or Language disagreements:

Disagreements of type A (Annotator): No significant linguistic differences in the text, but distinct relations labelled by two annotators (marked with an [A] in column Disagree of Table 7, and in corpus results in Table 17 under Annotation Discrepancies). We have found seven sources of such disagreement:

1. Different choice in nuclearity entailed a N/N-N/S mix-up **(N/N-N/S)**.

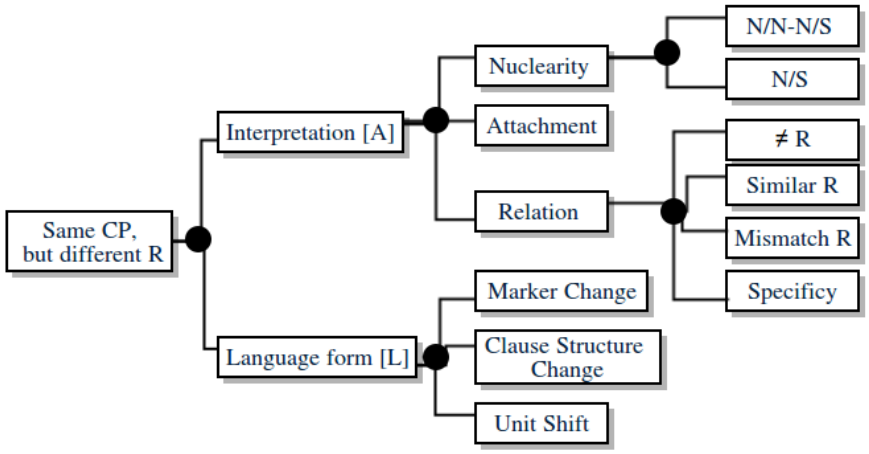
[9] If there is more than one CP (because there is a multinuclear relation constituting the relation) at least one has to be the same for N/S-N/N mix-up.

Fig. 5 Decision tree based on C AND ACP to establish the agreement types about R



2. Different choice in nuclearity entailed discrepancy in N/S relations **(N/S)**.
3. A relation has the same constituent and attachment point, but not the same relation label ($\neq$ **R**).
4. Relations chosen are similar in nature **(Similar R)**.
5. Relations with mismatched RST trees **(Mismatch R)**.
6. A relation is more specific than the other **(Specificity)**.
7. Different choice in attachment entailed a different relation **(Attachment)**.

Fig. 6 Decision tree to establish the sources of agreement and disagreement about R



Disagreements of type L (Language): Two annotators labelled distinct relations because there is a significant difference in the linguistic form (marked with an [L] in column Disagree of Table 7 and in corpus results in Table 20 under Translation Strategies). We have found three different sources. These are in fact translation strategies, and are sensitive to corpus and language. Studies in other corpora, genre or languages may reveal different strategies and sources of disagreement:

1. A relation is signalled with a different discourse marker **(Marker Change or MC)**.
2. A different organization of constituent phrases is used, mostly from non-finite verb phrase to finite verb phrase **(Clause Structure Change or CSC)**.
3. A change in unit level (phrase−clause−sentence) is done **(Unit Shift or US)**.

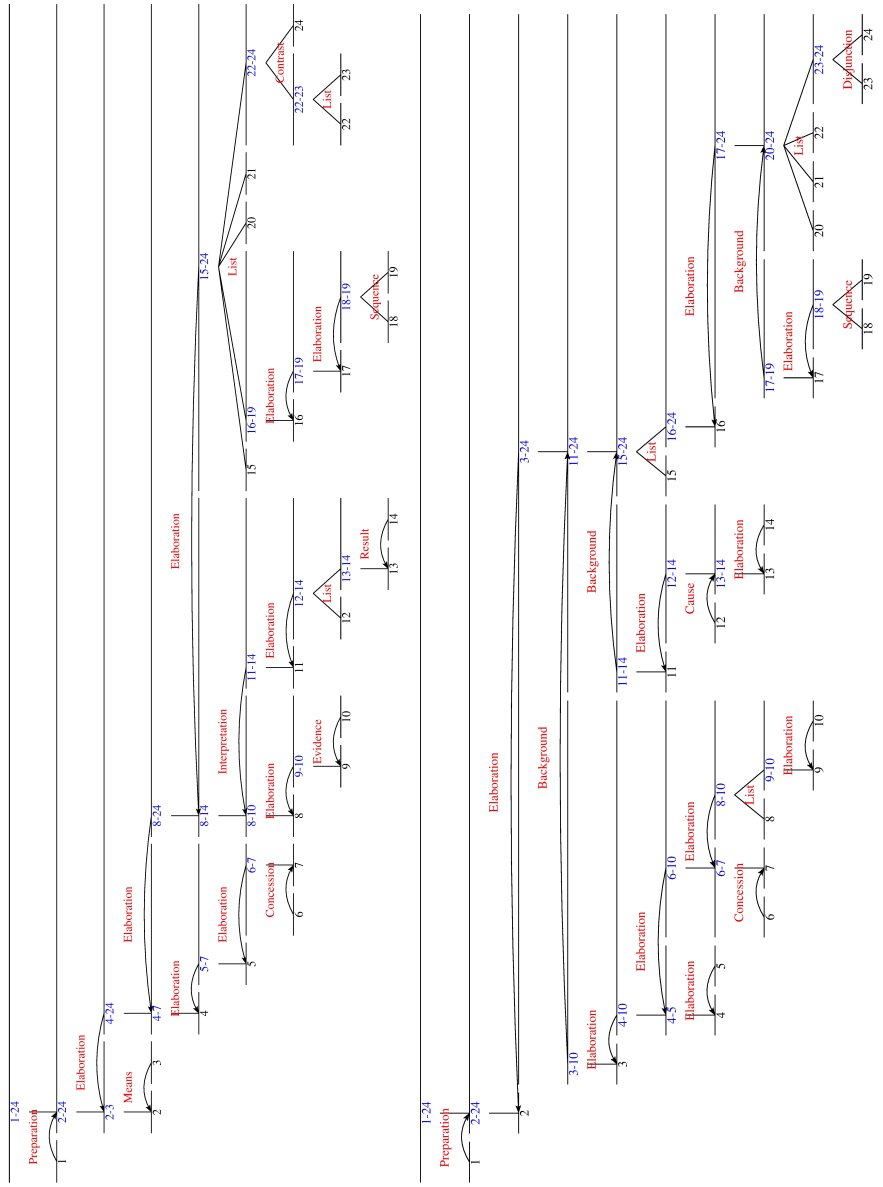
In Table 6 we show an example extracted from the corpus of text TERM38_SPA wich was segmented and harmonized in Spanish (A2) and in English (A1) (Figure 7) to illustrate the qualitative method (Table 7).[10]

[10] Basque (A3) was also harmonized, but space constraints preclude us to align with Spanish and English. Anyway, the harmonization of TERM38_SPA in the three languages could be consulted at http://ixa2.si.ehu.es/rst/segmentuak_multiling.php?bilatzekoa=TERM38%

Tables		Languages	
7	9	Spanish	English
1	1 to 6	**La neología contrarreloj: Internet**	**Neology against the clock: the Internet**
2	7 to 22	El propósito de esta comunicación es hacer una reflexión sobre los retos a que se está enfrentando la neología terminológica en la realidad actual,	This paper is intended to look at the challenges faced by neology in terminology at the present time.
3	23 to 38	para lo cual vamos a abordar diversos aspectos que influyen en la creación neológica en el ámbito de Internet.	I will do this by discussing various points which influence neology in the field of the Internet.
4	39 to 67	Los términos referidos a Internet nacen y se difunden a una velocidad y con una amplitud tal que constituye una verdadera carrera contrarreloj en las distintas lenguas.	Terms referring to the Internet are coined and spread at such speed and to such an extent that they have turned into a race against the clock in different languages.
5	68 to 92	Efectivamente, la formación de nuevos términos está sometida a un ritmo trepidante, paralelo al avance e innovación tecnológica en el sector de la informática y, en general, de las telecomunicaciones.	The formation of new terms goes on at a dizzy speed, parallel to technological advances and innovations in the field of computer science and telecommunications in general.
6	93 to 105	Si bien este aspecto es común al progreso científico y técnico y, por lo tanto, característico de la neología terminológica,	This is common in all scientific and technological progress, and therefore characteristic of neology in terminology,
7	106 to 123	la especificidad del área tratada confiere a la neología que le es propia unas particularidades que cabe tener en cuenta.	but the specific nature of this area confers particular features on neology which must be taken into account.
8	124 to 164	En primer lugar, el canal por el que se dan a conocer los términos de Internet, la misma red, no sólo supone una rápida difusión de la terminología —la información en Internet es de acceso (casi) inmediato—, sino también un alcance muy vasto —llega a cualquier parte del mundo—.	First of all the channel through which Internet terms are made known is the net itself. This means that they not only spread rapidly (information on the internet can be accessed almost immediately) but also reach vast areas (all over the world).
9	165 to 173	Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados;	Furthermore, terms can be compiled, discussed and assessed anywhere:
10	174 to 196	de ahí, por ejemplo, los apartados que encontramos en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar.	many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.
11	197 to 203	Esto nos lleva a una cuestión fundamental:	This leads us to the fundamental point:
12	204 to 224	la terminología de Internet traspasa los límites del área de especialidad (a la que se circunscribe por definición el léxico científico y técnico)	Internet terminology extends beyond the bounds of its specialist field (which by definition is part of the lexicon of science and technology)
13	225 to 229	e irrumpe en la lengua de uso general,	and breaks into general language.
14	230 to 256	siendo utilizada tanto por los usuarios heterogéneos de la red (de cualquier o ninguna especialidad) como por las personas que leen la prensa o están atentas a los medios de comunicación.	It is used both by a wide variety of net users (from any or no specialist fields) and by people who read the press or follow the media.
15	257 to 262	¿Qué tipo de terminología se está creando?	What type of terminology is being created?
16	263 to 267	¿Qué sistemas de creación léxica predominan?	What lexical creation systems predominate?
17	268 to 273	Un único denominador común existe para todas las lenguas:	There is a common denominator in all languages:
18	274 to 278	los términos se generan en inglés	terms are generated in English
19	278 to 281	y penetran como préstamos en aquellas.	and come in as loanwords.
20	282 to 289	¿Cómo responden las lenguas receptoras?	How do the receiving languages respond to this?
21	290 to 296	¿Cómo tratan la terminología de Internet?	How do they deal with Internet terminology?
22	297 to 307	¿Son términos todos los que lo parecen,	Are all those words which seem to be terms actually terms?
23	308 to 314	responden a necesidades reales de denominación,	Do they meet actual needs for names
24	315 to 320	o abundan las creaciones léxicas sensacionalistas y efímeras?	or do sensationalist, ephemeral terms abound?

Table 6 TERM38_SPA segmented and harmonized in Spanish and English

Fig. 7 Rhetorical tree by A2 (Spanish) and A1 (English), TERM38_SPA



L	ENG				SPA				Qualitative evaluation					
	CP(s)	R	C	A	CP(s)	R	C	A	N	R	C	A	Agree	Disagree
1	1	Preparation→	1S	2-24N	1	Preparation→	1S	2-24N	✓	✓	✓	✓	RCA	
2	3	Means←	3S	2N	3	Background→	3-10S	11-24N	✗	✗	✗	✗		N/S $[A]$
3	4	Elaboration←	4-24S	2-3N	4	Elaboration←	4-10S	3N	✓	✓	✗	✗	R	
4	5	Elaboration←	5-7S	4N	5	Elaboration←	5S	4N	✓	✓	✗	✓	RA	
5	7	Elaboration←	6-7S	5N	7	Elaboration←	6-10S	4-5N	✓	✓	✗	✗	R	
6	6	Concession→	6S	7N	6	Concesion→	6S	7N	✓	✓	✓	✓	RCA	
7	9	Elaboration←	9-10S	8N	8|9	List↔	9-10N	8N	✗	✗	✓	✓	(CA)	N/N vs N/S $[A]$
8	10	Evidence←	10S	9N	10	Elaboration←	10S	9N	✓	✗	✓	✓	(CA)	MC $[L]$
9	11	Interpretation←	11-14S	8-10N	11	Background→	11-14S	15-24N	✗	✗	✓	✗	(A)	N/S $[A]$
10	12	List↔	12N	13-14N	12	Cause←	12S	13-14N	✗	✗	✓	✓	(CA)	N/N vs N/S $[A]$
11	14	Result←	14S	13N	14	Elaboration←	14S	13N	✓	✗	✓	✓	(CA)	CSC $[L]$
12	15|16-24	List↔	15N	16-24N	15|16-24	List↔	15N	16-24N	✓	✓	✓	✓	RCA	
13	15|16|20-24	Elaboration←	15-24S	8-14N	15|16|20-24	Elaboration←	3-24S	2N	✗	✓	✗	✓	R	
14	16|20|21|22-24	List↔	20-24N	16-19N	20|21|22-24	Elaboration←	17-24S	16N	✗	✗	✗	✗		N/S $[A]$
15	17	Elaboration←	17-19S	16N	17	Background→	17-19S	20-24N	✗	✗	✓	✗	(A)	N/S $[A]$
16	18-19	Elaboration←	18-19S	17N	18-19	Elaboration←	18-19S	17N	✓	✓	✓	✓	RCA	
17	18|19	Sequence↔	18N	19N	18|19	Sequence↔	18N	19N	✓	✓	✓	✓	RCA	
18	20|21-24	List↔	20N	21-24N	20|21-24	List↔	20N	21-24N	✓	✓	✓	✓	RCA	
19	21|22-24	List↔	21N	22-24N	21|22-24	List↔	21N	22-24N	✓	✓	✓	✓	RCA	
20	22|23-24	List↔	22N	23N	22|23-24	List↔	22N	23-24N	✓	✓	✓	✓	RCA	
21	22-23|24	Contrast↔	22-23N	24N	23|24	Disjunction↔	23N	24N	✓	✗	✗	✓	(C)	$\neq$ R $[A]$
22	8	Elaboration←	8-24S	4-7N	8|9	Elaboration←	8-10S	6-7N	✓	✓	✗	✗	R	
23	12|13	Elaboration←	12-14S	11N	13	Elaboration←	12-14S	11N	✓	✓	✓	✓	RCA	

Table 7 Qualitative evaluation matrix TERM38_SPA

Table 7 includes the analyzed factors for Figure 7: nuclearity (N), relation (R), constituent (C) and attachment point (A). These factors compare A2 (Spanish) and A1 (English). In the Qualitative Evaluation columns, we mark with a "✓" an instance of agreement, and with an "✗" a disagreement. The last two columns summarize the type of agreement (Agree) or the disagreement source (Disagree).

If there is a multinuclear relation inside of a constituent of another relation (see lines 22 and 23 in Table 7) comparing CPs is not trivial, because multinuclear relations have more than one CP. The line 23 is representative of this problem. If we look at this line we can see that the problem is not the relation that we are comparing, but the problem comes from a lower level, since there is full agreement (RCA) between annotators (on R: ELABORATION, on C: 11N and on A: 12-14S). When this is the case there are two choices: a) do not compare relations and leave it as "no-match"[11] and b) compare first non-ambiguous CPs and leave problematic comparisons (lines 22 and 23) for the end. Following the last choice there is not any ambiguous CP in Table 7, because the other CP candidate (CP 12 in line 10) was used in other structure. Because of that, when we have to compare relations with more than one CP with another that has only one CP, at least one of the CPs has to be identical. If still there were cases in which we can not compare structures we have used the no-match label. This problem was found also in text summarization by Marcu [2000b], since the most important unit can be formed by more than one EDU.[12]

In Table 8 we present the results of our evaluation method for the example in Figure 7.

Nuclearity		**Relation**		**Composition**		**Attachment**	
Matches	**F1**	**Matches**	**F1**	**Matches**	**F1**	**Matches**	**F1**
16 of 23	0.6957	14 of 23	0.6087	15 of 23	0.6522	16 of 23	0.6957

Table 8 Qualitative evaluation results for the example in Figure 7, TERM38_SPA

In order to better highlight the differences between the quantitative method and our qualitative proposal, we have kept the rhetorical structure, but have used one of the languages to compare using RSTeval in contingency Table 9.

[11] If we follow this decision, we could not compare structures that contain a N/N-N/S mix-up inside the relation.

[12] As the evaluation has been done manually, there have been some problematic cases that have not counted as an agreement. For cases in which some structures cannot be compared, no-match label has been used, which represents no more than 0.06% of all relations (53 No Match / 900 relations), about 1.18 relations per text on average (53 No Match / 45 texts).

Constituent	Units		*Spans*		Nuclearity		Relation		Constituent	Units		*Spans*		Nuclearity		Relation	
	A1	A2	A1	A2	A1	A2	A1	A2		A1	A2	A1	A2	A1	A2	A1	A2
1 to 6	x	x	x	x	s	s	preparation	preparation	!268 to 281			x	x	s	s	background	elaboration
7 to 22	x	x	x	x	n	n	span	span	!263 to 281				x		n		list
23 to 38	x	x	x	x	n	s	span	means	257 to 262	x	x	x	x	n	n	list	list
!7 to 38				x		n		span	282 to 289	x	x	x	x	n	n	list	list
39 to 67	x	x	x	x	n	n	span	span	!257 to 289				x		n		list
68 to 92	x	x	x	x	s	n	elaboration	span	290 to 296	x	x	x	x	n	n	list	list
93 to 105	x	x	x	x	s	s	concession	concession	!257 to 296				x		n		list
106 to 123	x	x	x	x	n	n	span	span	297 to 307	x	x	x	x	n	n	list	list
!93 to 123			x	x	n	s	span	elaboration	308 to 314	x	x	x	x	n	n	disjunction	list
!68 to 123				x		s		elaboration	!297 to 314				x		n		contrast
!39 to 123				x		n		span	315 to 320	x	x	x	x	n	n	disjunction	contrast
124 to 164	x	x	x	x	n	n	list	span	!297 to 320				x		n		list
165 to 173	x	x	x	x	n	n	span	span	!257 to 320			x	x	n	n	span	list
174 to 196	x	x	x	x	s	s	elaboration	evidence	!263 to 320			x	x	n	s	list	elaboration
!165 to 196			x	x	n	s	list	elaboration	!124 to 320				x		s		elaboration
!124 to 196			x	x	s	n	elaboration	span	!39 to 320				x		s		elaboration
197 to 203	x	x	x	x	n	n	span	span	!7 to 320			x	x	n	n	span	span
204 to 224	x	x	x	x	s	n	cause	list	!1 to 320			x	x	r	r	span	span
225 to 229	x	x	x	x	n	n	span	span	!39 to 92			x		n		span	
230 to 256	x	x	x	x	s	s	elaboration	result	!93 to 196			x		s		elaboration	
!225 to 256			x	x	n	n	span	list	!39 to 196			x		s		elaboration	
!204 to 256			x	x	s	s	elaboration	elaboration	!23 to 196			x		s		background	
!197 to 256			x	x	s	s	background	interpretation	!282 to 296			x		n		list	
!124 to 256				x		n		span	!282 to 307			x		n		list	
263 to 267	x	x	x	x	n	n	span	span	!308 to 320			x		n		list	
268 to 273	x	x	x	x	n	n	span	span	!282 to 320			x		n		span	
274 to 277	x	x	x	x	n	n	sequence	sequence	!268 to 320			x		s		elaboration	
278 to 281	x	x	x	x	n	n	sequence	sequence	!197 to 320			x		n		span	
!274 to 281			x	x	s	s	elaboration	elaboration	!23 to 320			x		s		elaboration	

Table 9 Contingency table for text TERM38_SPA with quantitative method, using *RSTeval*

Units		Span		Nuclearity		Relation	
Match	***F*1**	**Match**	***F*1**	**Match**	***F*1**	**Match**	***F*1**
24 of 24	1	36 of 47	0.766	29 of 47	0.617	20 of 47	0.425

Table 10 Quantitative method results for text TERM38_SPA

Both methods measure the similar factors: i) EDUs and spans (constituent and attachment), ii) nuclearity (of each unit, or direction of the relation) and rhetorical relations (of each unit: relation plus span, or relation as a whole). Thus, in Table 11 we can compare how each method accounts for these factors.

	Units		Spans				Nuclearity		Relation	
Quanti.	24 of 24	1	37 of 46		0.8043		29 of 46	0.6304	21 of 46	0.4565
	Units		**Composition**		**Attachment**		**Nuclearity**		**Relation**	
Quali.	24 of 24	1	15 of 23	0.6522	14 of 23	0.6087	17 of 23	0.7391	13 of 23	0.5652

Table 11 Comparison using both methods, TERM38_SPA

In Table 11 both methods describe total agreement in segmentation. This is of course due to the fact that segmentation was harmonized before the analysis was undertaken. The *span* factor of the quantitative method is described using factors C and A, this factor being more positive in the quantitative method. In terms of nuclearity and rhetorical relations, we can see that the qualitative method is able to describe more agreements in the evaluation of text TERM38.

In Table 12 we can observe further detail on how both methods describe agreement in relations, and the weight given to each relation in the calculation of agreement. To better understand the table, we have highlighted in gray the most important differences.

Relation	**Quantitative method**				**Qualitative method**			
	A1	**A2**	**Match**	%	**A1**	**A2**	**Match**	%
Background	3				3			
Cause	1				1			
Concession	1	1	1	2,17	1	1	1	4,35
Contrast		2				1		
Disjunction	2				1			
Elaboration	10	9	2	4,35	10	9	6	26,09
Evidence		1				1		
Interpretation		1				1		
List	10	12	6	13,04	5	6	4	17,39
Means		1				1		
Preparation	1	1	1	2,17	1	1	1	4,35
Result		1				1		
Sequence	2	2	2	4,35	1	1	1	4,35
Span	16	15	9	19,57	–	–	–	–
Total	**46**	**46**	**21**	**45,65**	**23**	**23**	**13**	**56,52**

Table 12 Comparison, description of agreement under both methods for text TERM38

As we can see in Table 12, an important part of the agreement in quantitative evaluation method is capture in the SPAN label (which is not an RST relation). In addition, the contingency table shows that the relation with most agreement is the LIST relation, followed by ELABORATION and SEQUENCE. Thanks to the qualitative evaluation, however, we can see that the ELABORATION relation actually has a higher degree of agreement, followed by LIST. In contrast, SEQUENCE has little importance, the same as CONCESSION and PREPARATION. We would like to point out that the difference is more striking when describing agreement (Match: columns 4 and 8), rather than when describing how often the annotator has used such relation (A1: columns 2 and 6, and A2: columns 3 and 7). For instance, in both methods we can see that A1 has used 10 ELABORATION relations, whereas A2 has used 9. The quantitative method captures an agreement of 4.35%; the qualitative method throws a much higher agreement, reaching 26.09%.

The root of this difference can be found in the fact that the quantitative evaluation does not evaluate nuclearity and rhetorical relations in an independent way. When creating relation pairs, the pairs do not have well-formed members (in particular because of the use of the SPAN label). This is the reason why in the quantitative method, out of 10 ELABORATION relations, only two show agreement.

Advantages of the qualitative evaluation method. The formalization of qualitative evaluation (Table 7) describes the annotation agreement (Agree) in a more complete way than quantitative evaluation (Table 9): the relation factor (R) is compared in an isolated manner, that is, nuclearity is not reanalyzed in the relation factor. This fact has methodological implications and some the advantages show in contingency Table 7:

i) Independent factors are evaluated. A different attachment point of a relation only implies disagreement in attachment point (disagreement described at the same line) and in constituent (disagreement described at higher level tree structure) and not in relation as quantitative method does. Moreover, the qualitative method accounts for the source of disagreement (Disagree).

ii) Only rhetorical relations are compared. The description allows for a full coincidence in structure (RCA), or a partial match (RA, RC or R).

iii) Reasons for annotator disagreement are captured: *a*) because of differences in the linguistic expression [L] or *b*) because of interpretation [A].

iv) Relation pairs in the contingency table are able to better describe agreement and disagreement (confusion patterns, [Marcu, 2000a]).

For example, in Table 7 we can observe the following types of information on the relation agreement:

1. Match in relation, constituent and attachment point (RCA) in the following nine lines: 1, 6, 12, 16, 17, 18, 19, 20 and 23. We observe that in these lines there was total agreement in the three factors observed, that is, for example, in line 1 an agreement in all factors: same CP (1), relation (PREPARATION), constituent (1S) and attachment point (2-24N).
2. Match in relation and attachment point (RA) in line 4. A partial agreement, but now in CP (5), relation (ELABORATION) and attachment point (4N), whereas in constituent there was a slight disagreement (A2: 5-7S but A1: 5S).

3. Match only in relation (R) in four lines: 3, 5, 13 and 22. For example, in line 3 there was an agreement only in CP (4) and relation (ELABORATION), whereas there were discrepancies in constituent (A2: 4-24S but A1: 4-10S) and attachment point (A2: 2-3N but A1: 3N).

On the relation disagreement, we can observe the following types of information in Table 7:

1. A different choice in nuclearity (N/S [A]) in four lines: 2, 9, 14 and 15.
2. A N/N-N/S mix-up (N/N-N/S [A]) in two lines: 7 and 10.
3. A different relation label ($\neq$ R [A]) in a line: 21.
4. A Marker Change (MC [L]) in a line: 8.
5. A Clause Structure Change (CSC [L]) in a line: 11.

3 Results

In this section, we first present the results of segmentation, and then we compare the results of rhetorical structure based on two evaluation methods: quantitative method [Marcu, 2000a] and our new proposal for a qualitative evaluation method.

3.1 Discourse Segmentation Results

The initial round of segmentation led to the following number of EDUs: 330 in English, 318 in Spanish, and 323 in Basque. We calculated agreement using F-score and Kappa, in a pairwise manner. First of all, we calculated the total coincidence of EDUs, using the verb of the main clause and its principal arguments (VP). If the main verb was the same in both EDUs, then we tabulated it as a match. As we stated in page 8, one of our segmentation principles is that every EDU should contain a finite verb. The main verb of an EDU indicates the principal action, process, state, condition, etc., in relation to the subject of the clause. Therefore, if two EDUs in different languages contain the same verb (that is, both verbs are translation equivalents), they are expressing the same event and we consider that there is coincidence between EDUs. Thus, in this sense, syntax has an important role to play in the detection of the EDUs to be compared, since we take the main verb of the clausal syntactic structure in each language to carry out the comparison. In this work, we have not used a syntactic parser to perform the analysis. We have done the analysis manually, because it was feasible to do it over our corpus and we also wanted to avoid possible mistakes in the harmonization work.[13] In future work, however, we plan to automate our methodology to compare discourse structures, and, in this case, we could integrate a syntactic parser in the system. We then calculated F-measure and Kappa as presented in Table 13.[14]

[13] This harmonization work can be found at http://ixa2.si.ehu.es/rst/segmentuak_multiling.php.

[14] For Kappa segment candidates were calculated automatically by counting verbs.

Language	Correct	Match	Wrong	Missing	Candidates	F-measure	Kappa
ENG-SPA	330	230	88	12	731.4	70.99	0.7139
ENG-BSQ	330	226	97	7	742.9	69.22	0.7057
BSQ-SPA	323	230	88	5	731.4	71.76	0.7333

Table 13 Segmentation agreement

3.1.1 Discourse Segmentation Harmonization

In our segmentation, it was often the case that one language used a finite verb, whereas the other language used a non-finite verb or other expression, leading to differences in segmentation. Another source of disagreement was the interpretation of ellipsis, where one annotator decided there was more than subject ellipsis in coordination, and did not break up the two VPs, whereas the other annotator decided to break them up. Two other sources of disagreement were different texts in the two languages (not different formulations, but a completely different text, with one sentence deleted or inserted), and simple human error. The latter accounts for no more than two disagreements per language pair.

Harmonization led to joining or separating EDUs in one of the languages, contravening our general principles for segmentation. The main changes in this harmonization were:

1. When two parallel passages share the same structure and the third passage does not, then we harmonize the segmentation of the third language taking into account the segmentation of the two coincident languages.
2. When the segmentations of the three parallel passages are different, then we harmonize the segmentation taking into account the structure of the simplest passage.

In Example (1) a Basque conjunct was translated as a clause in both English and Spanish. In the English example there are three finite verbs (all three of them instances of the verb *is*), as is the case in Spanish (*es*, '[it] is'; *se ubica,* '[it] is located'; and *va*, '[it] goes'). In Basque, however, there are only two finite verbs (*estrapolatuko du,* '[it] will extrapolate [it]'; and *jartzen du,* '[it] places [it]'). The third part of the conjunct contains no verb (*eta hizkuntza erromanikoek ezkerraldean,* 'and the Romance languages on the left side'). In the harmonization we inserted a new segment in Basque, reinterpreting not as coordinated NP, but as a juxtaposed clause with an elided verb.[15]

(1) a. [Our hypothesis is that a syntactic characteristic of Basque and the romance languages is extrapolated to their morphology,] [so that in Basque derivations the core of the structure is on the right,] [while in the romance languages it is on the left.]

b. [Nuestra hipótesis es que una característica sintáctica del euskera y de las lenguas románicas se extrapola hasta la morfología,] [de manera que en euskera, también en derivación, el núcleo de la estructura se ubica a la derecha,] [mientras que en las lenguas románicas va a la izquierda.]

[15] In the example, the original segmentation is marked with square brackets and the segmentation after harmonization with curly brackets.

c. [Gure hipotesiak, euskararen eta hizkuntza erromanikoen ezaugarri sintaktiko bat morfologiaraino estrapolatuko du:] [eratorpenean ere euskarak egituraren burua edo gunean eskuinaldean jartzen du,} {eta hizkuntza erromanikoek ezkerraldean.] TERM50_BSQ

In Example (2) the translation from Spanish into English has led to two separate clauses. The Spanish original segmentation contained only one span, since the first idea (*un aumento cuantitativo de la terminología especializada,* 'an increase in the number of specialist terms') is embedded in a non-finite clause (*además de provocar,* 'in addition to leading to'). The English translation splits the ideas into two coordinated clauses (*factors lead to an increase* and *but also* [*factors*] *call into question*). Basque also has two clauses to express these two ideas. Since two of the languages divided this sentence into two clauses, in the harmonization we inserted a new boundary in Spanish.

(2) a. [All these factors lead to an increase in the number of specialist terms which enrich terminology] [but also call into question some of its basic concepts, such as the one to one relationship between ideas and names, the concept of mastery of a specialist field and the role of standardization in terminology.]

b. [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,} {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos, como la univocidad noción-denominación, el concepto de dominio de especialidad o el papel mismo de la normalización en terminología.]

c. [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;] [eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan jarri ditu terminologiaren oinarrizko zenbait kontzeptu: kontzeptu-izendapen bikotearen adierabakartasuna, espezialitateko eremuen kontzeptua, eta normalizazioak terminologian duen eginbeharra.] TERM19_SPA

We quantified the changes necessary to harmonize the segmentations by counting how many times a change was necessary, per language. Table 14 summarizes those changes (the typical actions are "join" or "break up"), and the number of affected EDUs. To compute the number of affected EDUs, we counted, in the cases where we needed to break down a unit, how many new units were necessary (+). In the cases where we needed to join, we counted how many original units were integrated (−). In the table, "initial spans" refers to the spans proposed by the individual annotator for each language, and "affected spans", to the number of spans that underwent a change, whether to join, or to break up. "Harmonized spans" represents the final agreed upon spans across all three languages, for each text.

	Initial Spans			Harmon.	Affected Spans		
Text	**ENG**	**SPA**	**BSQ**	**Spans**	**ENG**	**SPA**	**BSQ**
TERM18_ENG	8	11	14	8	0	−3	−6
TERM19_SPA	14	12	13	14	0	+2	+1
TERM23_ENG	15	14	14	14	−1	0	0
TERM25_BSQ	10	11	8	10	0	+1	+2
TERM28_BSQ	16	14	12	15	−1	+1	+3
TERM29_SPA	14	14	13	14	0	0	+1
TERM30_ENG	26	27	33	28	+2	+1	−5
TERM31_BSQ	53	52	44	52	−1	0	+8
TERM32_ENG	13	13	18	13	0	0	−5
TERM34_BSQ	50	45	44	46	−4	+1	+2
TERM38_SPA	27	25	28	24	−3	−1	−4
TERM39_ENG	7	8	9	9	+2	+1	0
TERM40_SPA	8	8	8	8	0	0	0
TERM50_BSQ	34	35	30	30	−4	−5	0
TERM51_SPA	35	29	35	31	−4	+2	−4
Total	**330**	**318**	**323**	**316**	**±22**	**±18**	**±41**
Change rate					**6.67%**	**5.66%**	**12.69%**

Table 14 Segmentation changes

We can see from the table that the language with more changes is Basque.[16] We found that the linguistic expression of the same or similar concepts required different syntactic constructions in Basque. This makes sense, given that Basque is a non-Indo-European language, showing considerable typological distance from both Spanish and English [Cenoz, 2003]. Note that whereas Spanish and Basque were affected in the same proportion in both directions (when breaking down SPA: 44.44% and BSQ: 41.46%; when joining SPA: 55.56% and BSQ: 58.54%), harmonization in English involved breaking down in a much lower proportion (when breaking down ENG: 18.18%; when joining ENG: 81.82%). This seems to indicate the corpus abstracts in English (whether translated or original) express clauses as separate units, either as simple sentences or as clear (finite) adjunct clauses, without recourse to non-finite clauses or prepositional complements.

3.2 Rhetorical Analysis Results

The results of quantitative method were presented in order to show the consistency of the qualitative method. To this end, first, we present below the results of the quantitative method; second, we present the results of the qualitative method, and after that we compare results from both methods.

[16] One-way ANOVA demonstrated significant differences across the three languages in the corpus ($p = 0.07$). We thought this was quite significant, therefore we performed a post-hoc Tukey's test and we observed that harmonization in Basque is the furthest from the other two.

3.2.1 Results of quantitative evaluation method

Results of quantitative evaluation are shown in Table 15.[17]

Language comparison		Evaluation		
1st Lang.	**2nd Lang.**	**Span**	**Nuclearity**	**Relation**
ENG	SPA	84.06%	67.43%	56.22%
ENG	BSQ	86.22%	68.24%	53.28%
SPA	BSQ	88.61%	71.02%	54.94%

Table 15 Quantitative evaluation results (F-measure)

Surprisingly, results for the quantitative evaluation are slightly better when Basque is involved in the comparison, which was not the case for the segmentation Span agreement results (Table 14). Agreement, however, is higher for the Nuclearity criterion when Basque is included (also the case for Span agreement results shown earlier). Finally, the Relation agreement drops when Basque is involved. We point out the source of this change and we discuss the results of the Relation comparison in Section 2.4.2, where we present the final results of both evaluation methods (Table 21).

3.2.2 Results of qualitative evaluation method

Table 16 and Table 17 include the final results for the entire corpus, which account for agreement and disagreement in a qualitative way. In Table 16 results from the agreement level obtained on the four types of measurements increases as the relaxation of the agreement increases too, being RCA the most demanding agreement, and R the more relaxed one.

Classification		**ENG-SPA**		**ENG-BSQ**		**SPA-BSQ**	
	%	**Gain**	**%**	**Gain**	**%**	**Gain**	
Agreement	RCA	44.67 %		40.33 %		42.33 %	
	RC	49.34 %	4.67	42.66 %	2.33	45.66 %	3.33
	RA	51.67 %	7	48.66 %	8.33	50.66 %	8.33
	R	59.67 %	3.33	54.66 %	3.67	56.99 %	3

Table 16 Qualitative evaluation results (F-measure): analysis of the sources of agreement

In Table 18 we show summarized results of the three sources: total agreement between annotators (Agreement), discrepancies because of annotation decisions (Annotation Discrepancies) and discrepancies because of linguistic differences (Translation Strategies).

As we observe in Table 18, the disagreement is higher when data of both A1 (English) and A2 (Spanish) are compared with A3 (Basque). That could be, as we

[17] EDUs are excluded because they are identical after harmonization.

Classification		ENG-SPA	ENG-BSQ	SPA-BSQ
Annotator-based Discrepancies	Nuclearity	4.00%	4.00%	3.33%
	N/N vs. N/S	5.33%	8.00%	6.00%
	Attachment span	2.00%	1.33%	0.67%
	Relation	6.67%	4.00%	2.67%
	Similar Relation	1.67%	4.33%	6.67%
	Mismatched Relation	6.00%	4.67%	5.67%
	Specificity	0.67%	4.33%	5.33%
	No Match	6.33%	6.67%	4.67%
Language-based Discrepancies	Marker Change	4.67%	3.33%	4.67%
	Clause Structure	1.67%	1.67%	1.33%
	Unit Shift	1.33%	2.67%	1.67%

Table 17 Qualitative evaluation results (F-measure): analysis of the sources of disagreement

Classification	ENG-SPA	ENG-BSQ	SPA-BSQ
Agreement	59.67%	54.66%	56.99%
Annotator-based Discrepancies	32.67%	37.33%	35.01%
Language-based Discrepancies	7.67%	7.67%	7.67%

Table 18 Qualitative evaluation results (F-measure): summary of results

discussed in Subsection 3.1.1, because English and Spanish are typologically closer to each other than Basque is to either English or Spanish. But this dispersion is not so large if we take into account the fact that there are more Similar Relations and Specificity when A3's data is compared with A1's and A2's.

Futhermore, the agreement attained across the three annotators was moderate with a Kappa [Fleiss, 1971] score of 0.484 (300 rhetorical relations, 15 texts). We show in Table 19 the agreement relation by relation across the three annotators.

	Kappa	z	p.value		Kappa	z	p.value
Antithesis	-0.008	-0.235	0.814	Justify	-0.009	-0.269	0.788
Background	0.420	12.589	0.000	List	0.554	16.629	0.000
Cause	0.352	10.552	0.000	Means	0.221	6.617	0.000
Circumstance	0.420	12.586	0.000	Motivation	0.136	4.084	0.000
Concession	0.705	21.155	0.000	Preparation	0.851	25.528	0.000
Condition	0.525	15.763	0.000	Purpose	0.335	10.057	0.000
Conjunction	0.172	5.151	0.000	Restatement	0.424	12.723	0.000
Contrast	0.376	11.272	0.000	Result	0.301	9.017	0.000
Disjunction	-0.001	-0.033	0.973	Sequence	0.499	14.966	0.000
Elaboration	0.531	15.933	0.000	Solutionhood	-0.011	-0.337	0.736
Evaluation	-0.003	-0.100	0.920	Summary	0.712	21.361	0.000
Evidence	-0.008	-0.235	0.814	Unless	-0.001	-0.033	0.973
Interpretatio n	0.080	2.390	0.017				

Table 19 Qualitative evaluation results (Fleiss' Kappa) for rhetorical relations

As we observe in Table 19, Fleiss' Kappa measure show different degrees of understanding rhetorical relations.

i) Almost perfect: PREPARATION.
ii) Substantial: SUMMARY and CONCESSION.
iii) Moderate agreement: LIST, ELABORATION, CONDITION, SEQUENCE, RESTATEMENT, BACKGROUND and CIRCUMSTANCE.
iv) Fair agreement: CONTRAST, CAUSE, PURPOSE, RESULT and MEANS.
v) Slight agreement: CONJUNCTION, MOVITATION and INTERPRETATION.
vi There is not enough data for: ANTITHESIS, DISJUNCTION, EVALUATION, EVIDENCE, JUSTIFY, SOLUTIONHOOD and UNLESS.

Translation Strategies. In carrying out the comparison of rhetorical structures, we observed some language differences. Some of them were produced when authors translated from one language into another (translation strategy),[18] and others were the result of comparing rhetorical structure in a pairwise manner, for instance in comparing English and Spanish with each other, when they are both translations of a Basque source. The latter cannot be regarded as translation strategies, so we will include only the first types under the umbrella term 'translation shift'. And the second type under the umbrella 'different language forms'.

On the one hand, we do not analyze translation strategies which do not lead the annotator to choose a different relation, as in Example (3); where in Basque the rhetorical relation was made explicit with the marker (*izan ere,* 'in fact'), but remains the same, a CAUSE relation is in the A1 analysis.[19]

(3) a. [In the recent past, a trend has been noted, and reported by many researchers in the area of Serbian scientific terminology, of importing borrowings of lexical and larger structural units from English into specific scientific registers, rather that to opt for translations, calques, etc.]$_{3N}$ [This corresponds closely to the fact that a consensus has been reached among Serbian scientists of various orientations regarding the status of English as the only language of scientific communication in the last several decades.]$_{4S-CAUSE}$

b. [Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unitate lexikalen maileguak eta unitate-egitura luzeagoen maileguak hartzen dira zientzia-erregistro zehatz baterako, itzulpenak edo kalkoak egin ordez.]$_{3N}$ [Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingelesari eman diote zientzia-komunikaziorako hizkuntza bakarraren estatusa.]$_{4S-CAUSE}$ TERM18_ENG

On the other hand, we do analyze all the directions (ENG>SPA, ENG>BSQ and so on) in Table 20 and three types of translation differences that influence rhetorical relation and reveal local translation strategies:

[18] Catford [1965, pg. 73] defines translation shift as "departures from formal correspondence in the process of going from the SL to the TL". Chesterman [1997] states that changes from original to translated text are due to a translation strategy.

[19] Note that here there is another translation strategy (CSC hierarchical upgrading in Basque with a coordination of two finite verbs *lortu dute* '[they] achieve [it]' and *eman diote* '[they] give [him]') which is not under consideration due to harmonization process.

1) Relation signalling has a different configuration (Marker Change). Within Marker Change, we found three subtypes:
   i) inclusion of a marker,
   ii) exclusion of a marker, and
   iii) changing a marker.
2) Differences because of the use of a distinct language configuration (Clause Structure Change):
   i) hierarchical downgrading, and
   ii) hierarchical upgrading.
3) Punctuation is used differently (Unit Shift):
   i) an independent sentence is integrated in another sentence, and
   ii) a clause is translated in an independent sentence. We detail some of them below.

1. **Marker Change**. In Example (4) a discourse maker (*de ahí*, 'hence') was not translated from Spanish into either English or Basque. In English the marker *por ejemplo* 'for example' was also elided and the punctuation changed (from semicolon into colon). This is why annotators in English and Basque labelled the relation Elaboration; whereas in Spanish, the marker *de ahí*, 'hence', resulted in an annotation with the evidence label.

(4) a. [Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados;]$_{9N}$ [de ahí, por ejemplo, los apartados que encontramos en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar.]$_{10S-EVIDENCE}$

b. [Furthermore, terms can be compiled, discussed and assessed anywhere:]$_{9N}$ [many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.]$_{10S-ELABORATION}$

c. [Are gehiago, edozein tokitatik biltzen dira terminoak, baita komentatu eta haztatu ere;]$_{9N}$ [adibidez, Interneti buruzko terminoen glosarioak zabaltzen dira Web askotan, eta izendegietarako proposamenak egin ere bai, eta erabiltzaileek botoa eman ahal izaten diete.]$_{10S-ELABORATION}$ TERM38_SPA

2. **Clause Structure Change**. In Example (5) the clauses under the relative used in the original Spanish text were avoided in the same way in English and in Basque (*que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos,* 'that, although [it] has enriched it, [it] has also called into question some of its basic concepts'), in favour of an adversative coordination using a finite verb in English (*but*), and a conjunction coordination (*eta,* 'and') and a finite verb in Basque (*jarri ditu*, '[it] places [them]'). That was the reason for A1 to annotate a CONTRAST relation, whereas A3 annotated a LIST relation. The relative form[20] analyzed here is a product of the harmonization and it was annotated by A2 as an ELABORATION relation.

[20] Again, this goes against the principles of our segmentation.

(5) a. [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,$]_{6N}$ {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos $(\ldots)]_{7-11S-ELABORATION}$[21]

b. [All these factors lead to an increase in the number of specialist terms which enrich terminology$]_{6N-CONTRAST}$ [but also call into question some of its basic concepts $(\ldots)]_{7N-CONTRAST}$

c. [Alderdi horiek guztiek, espezialitateko terminologiaren gehikuntza kuantitatiboa eragiteaz gain, terminologia lanen ikuspegia ere zabaldu egin dute;$]_{6N-LIST}$ [eta, egia bada ere ikuspegi berri horrek terminologia aberastu egin duela esatea, zalantzan jarri ditu terminologiaren oinarrizko zenbait kontzeptu $(\ldots)]_{7N-LIST}$ TERM19_SPA

3. **Unit Shift**. A different punctuation can lead the annotator to interpret a different relation. In the original text in Spanish in Example (6), the spans were linked with comma, whereas in the English text the punctuation was changed, using a period. The punctuation led A1 to consider a hypotactic relation between the first and the following two spans.

(6) a. [En esta comunicación, a partir de la experiencia en trabajos de normalización de terminología catalana, se planteará la necesidad social de la normalización terminológica,$]_{N12-LIST}$ [se comentarán algunas de las dificultades con que se enfrenta y se apuntarán ideas para su enfoque dentro de la sociedad actual.$]_{N13-14-LIST}$

b. [This paper looks, on the basis of experience in the standardisation of terminology in Catalan, at the social need for standardisation of terminology.$]_{N12}$ [Some of the difficulties faced will be discussed, and ideas will be given for approaching this field in present day society.$]_{S13-14-ELABORATION}$ TERM19_SPA

We present, in Table 20, the influence of translation strategies and different language forms more in depth.

	Translation Strategies						Different Language Forms		
	ENG>SPA	ENG>BSQ	SPA>ENG	SPA>BSQ	BSQ>ENG	BSQ>SPA	ENG-SPA	ENG-BSQ	SPA-BSQ
MC	1.45%	–	4.35%	7.25%	10.14%	11.59%	14.49%	4.35%	1.45%
CSC	1.45%	1.45%	2.90%	4.35%	4.35%	1.45%	2.90%	1.45%	–
US	2.90%	2.90%	2.90%	1.45%	4.35%	2.90%	0.00%	4.35%	2.90%
Total			**68.12%**					**31.88%**	

Table 20 Translation strategies and different language pairs

It is worth mentioning that when English is the SL there are not so many translation strategies (10.14%) as when other languages are SL (Spanish: 23.19% and

[21] Note here the human annotation error which does not follow the modular and incremental annotation that Pardo [2005] proposes.

Basque: 34.78%). Another interesting aspect is that the Marker Change translation strategy is the most prominent one (MC: 34.78% vs CSC: 15.94% and US: 17.39%), and changes in discourse markers have an influence on rhetorical annotation.[22] These results are merely describing tendencies, because the corpus is not big enough (although is comparable to other corpora in the literature Scott et al [1998]). The results are sensitive to segmentation granularity or harmonization decisions and to text characteristics (genre and domain). But what is relevant here is that the method presented here can describe and quantify translation strategies.

3.2.3 Comparing Quantitative and Qualitative Methodologies

To determine whether the proposed method is consistent, we compare the quantitative results of the relation factor from both methods in Table 21, where we present the final results from both evaluation methods, providing the F-measure of relation factor.

	Quantitative Evaluation	**Qualitative Evaluation**
ENG-SPA	56.22%	59.67%
ENG-BSQ	53.28%	54.66%
SPA-BSQ	54.94%	56.99%

Table 21 Comparison of relation factor in quantitative and qualitative evaluation methods (F-measure)

We can highlight two findings in this comparison:

1. The qualitative method finds slightly higher agreement than the quantitative method. The difference goes from almost 2% to 4% when we compare results in a pairwise manner.
2. Both methods show the same relative agreement rate per language pair. The pair with the highest agreement corresponds to English-Spanish, second comes the pair Spanish-Basque, and finally the pair English-Basque shows the lowest agreement.

In the rhetorical analysis, unlike those we have achieved in the harmonization (changes made in languages to carry out the alignment of discourse units), we see no significant difference (Translation Strategies in Table 20) between languages typologically more distant. It is worth noting, however, that for the closest languages, the English-Spanish pair, the agreement in relation is higher. Languages with more contact like the Spanish-Basque pair obtain better agreement than the English-Basque pair (Table 21).

We see clear advantages to the use of the qualitative evaluation method. First of all, with a qualitative evaluation, we measure inter-annotator agreement using only RST relations. Relations and nuclearity are phenomena of a different nature, and we

[22] This phenomenon (marker change is the first reason to mismatch relations) is repeated when we compare translated texts (TL) among them (MC 20.29%, CSC 4,35% and US 7.25%).

believe they ought not to be included in the same factor. Secondly, the qualitative evaluation clearly distinguishes the most relevant sources of disagreement; because of that, results are more reliable. The translation of discourse structure from one language to another is not one to one. As Marcu [2000a] has mentioned, sometimes a particular rhetorical structure has to be translated as a different structure. Moreover, translation strategies can affect the rhetorical structure and annotation, and the qualitative method presented here could be used to identify and measure these translation strategies.

4 Conclusions and Further Work

The methodology we have proposed has two main implications for RST theory and for annotation methodology. First of all, in terms of RST theory, we have shown that it is possible to conduct cross-linguistic studies using the same set of principles. In our study we have shown that, although RST structures may not be exactly the same across languages, they do show a large similarity. Secondly, we have provided a clear and detailed method to identify where structures differ. Thirdly, the annotated files are available to anyone who wishes to use them and on our website,[23] the tagged multilingual corpus can be consulted, as for example: i) the rhetorical structure of a text (in RS3 format) and it's image (in JPG format). ii) all instances of a selected rhetorical relation in three languages; iii) discourse units of a text in each language or aligned in three languages.

Ours is, to our knowledge, the first study that provides a rigorous qualitative methodology, which solves the deficiencies of quantitative evaluations and provides a qualitative description of agreement and disagreement. This method distinguishes and locates translation strategies when those strategies are the sources of annotator disagreement, as opposed to simple annotator discrepancies. The methodology helps determine whether the same passage in different languages has different RST structures because those structures correspond to different applications of the theory, or whether the discrepancy in RST structures is due to different linguistic realizations (due to translation strategies, broadly understood).

The study has some limitations with regard to the source of the translation differences that the analysis reveals. We believe that in order to detect these sources a translation theory "must include both a descriptive and an evaluative element", as Chesterman [1993] suggests, so that we can decide whether translation strategies may or may not be well motivated. We have presented some suggestions for the translation differences that the analysis evidenced, showing that the typological differences between the languages affected mostly the segmentation. More detail, informed by a rigorous translation theory, is necessary, but is beyond the scope of this paper.

Our results show that RST, in conjunction with our methodological proposal for the comparison of RST annotations, are valid tools for the study of translated corpora. The results of our corpus analysis provide some evidence that, in segmentation, the linguistic distance calculated by change in the harmonization process is very small

[23] http://ixa2.si.ehu.es/rst

between languages from the same family such as English-Spanish and it is large between languages from distinct families such as Spanish-Basque and English-Basque. Surprisingly, the dispersion in relation agreement caused by translation strategies was very small when comparing English-Basque and Spanish-Basque with English-Spanish. In the same line, the linguistic distance in rhetorical relations, calculated as the F-score result when comparing RST annotations, is not as large as the segmentation differences. It appears that there is more dispersion in segmentation than in rhetorical relations; this may be due to the fact that there is more distance at the level of clause linking than at the level of discourse relational structure. It is worth noting, however, that each language is affected by a particular translation strategy in this corpus.

Although the results obtained by both methods in the annotations for different languages show that there are different interpretations, this is not due to interlingual differences. The problem of annotation subjectivity arises also when three annotators analyze the same text in a language: this problem is even more important when the annotators do not have the same training (although in our experiment the three annotators started their annotation from the same departure criteria). As we said, the purpose of this paper is to present a methodology to compare RS-trees and no to describe the structure of text in the three languages. To see a description of those texts and a detailed work in these three languages, we recommended consulting the corpora developed by the authors in these three languages (English SFU corpus[24] [Taboada and Renkema, 2008], Spanish RST TreeBank[25] [da Cunha et al, 2011] and Basque RST TreeBank[26] [Iruskieta et al, 2013a]). We are aware that in this work we do not account for the problem of multiple relations in RST [Taboada and Mann, 2006b; Marcu, 2000b] or all the possibilities comparing RS-trees in parallel corpora.

The qualitative evaluation is in certain respects more complex than Marcu's quantitative evaluation, which has been automated by Maziero and Pardo [2009]. Despite its complexity, it solves some inherent problems of the quantitative evaluation and it has advantages when describing the sources of disagreement.

We plan to perform two tasks as future work. First of all, we will carry out a larger RST multilingual corpus analysis, but limited to a smaller number of rhetorical relations, with the objective of detecting translation strategies in order to improve machine translation discourse tasks. Second, we will carry out an automatic implementation of the qualitative rhetorical evaluation that we propose in our work, which will be valid for monolingual [Iruskieta et al, 2013a] and multilingual annotation, so that it can be used by all the scientific community working on RST.

[24] SFU corpus is available at http://www.sfu.ca/~mtaboada/download/downloadRST.html.

[25] RST Spanish TreeBank is available at http://corpus.iingen.unam.mx/rst/corpus_en.html.

[26] Basque RST TreeBank is available at http://ixa2.si.ehu.es/diskurtsoa/en/.

APPENDIX A

Discourse segmentation details

The first step in analyzing texts under RST consists of segmenting the text into spans. Exactly what a span is, under RST, and more generally in discourse, is a well-debated topic. RST Mann and Thompson [1988] proposes that spans, the minimal units of discourse —later called elementary discourse units (EDUs) [Marcu, 2000a] —are clauses, but that other definitions of units are possible:

> The first step in analyzing a text is dividing it into units. Unit size is arbitrary, but the division of the text into units should be based on some theory-neutral classification. That is, for interesting results, the units should have independent functional integrity. In our analyzes, units are essentially clauses, except that clausal subjects and complement and non-restrictive relative clauses are considered as part of their host clause units rather than as separate units.
>
> [Mann and Thompson, 1988, pg. 248]

This definition is the basis of our work. From our point of view, adjunct clauses stand in clear rhetorical relations (cause, condition, concession, etc.). Complement clauses, however, have a syntactic, but not discourse, relation to their host clause. Complement clauses include, as Mann and Thompson [1988] point out, subject and object clauses, and restrictive relative clauses, but also embedded report complements, which are, strictly speaking, also object clauses.

Other possibilities for segmentation exist; one of the better-known ones is the proposal by Carlson et al [2003] for segmentation of the RST Discourse Treebank [Carlson et al, 2002]. Carlson et al [2003] propose a much more fine-grained segmentation, where report complements, relative clauses and appositive elements constitute their own EDUs.

In our work three annotators segmented the EDUs of each corpus (A1 segmented English texts, A2 segmented Spanish texts, and A3 segmented Basque texts). These annotators are experts on RST, since they have been researching in this field since years ago, and they have participated in several projects related to the design and elaboration of RST corpora in the three languages of this work. Annotators performed this segmentation task separately and without contact among them. In our segmentation, we follow then the general guidelines proposed by Mann and Thompson [1988], which we have operationalized for this paper. We detail the principles below.

Every EDU Should Have a Verb In general, EDUs should contain a (finite) verb. The main exception to this rule is the case of titles, which are always EDUs, whether they contain a verb or not.

Non-finite verbs form their own EDUs only when introducing an adjunct clause (but not a modifier clause, as we will see below). In (7), the non-finite clause *Focussing on less widely...* is an independent EDU, because it is an adjunct clause. Note that in both Spanish and Basque the same proposition was translated as an independent sentence.

(7) a. [Focussing on less widely used and taught languages (LWUTLs) including Irish,] [the VOCALL partners are compiling multilingual glossaries of technical terms in the areas of computers, office skills and electronics] [and this involves the creation of a large number of new Irish terms in the above areas.]

b. [El proyecto está enfocado hacia lenguas minoritarias en cuanto al uso y enseñanza, incluido el irlandés.] [El proyecto VOCALL está en proceso de recopilación de un glosario plurilingüe de términos técnicos de las áreas de informática, secretariado y construcción,] [y esto supone la creación de una larga serie de nuevos términos en irlandés, en las áreas mencionadas.]

c. [Gutxi erabiltzen eta irakasten diren hizkuntzetan kontzentratzen da proiektua (LWUTL), irlandera barne.] [Informatika, bulego-lana eta eraikuntzako arloetako termino teknikoen glosario eleanizduna biltzen ari da VOCALL,] [eta horrek esan nahi du arlo horietako irlanderazko termino berri ugari sortzen ari dela.] TERM23_ENG

In some cases, a prepositional phrase (especially one containing a nominalised verb) in one language was realized as an independent clause in another. The final decision in such cases is typically to segment minimally, that is, to unify the segmentation across the three languages, so that the language with the fewer segments determines how the texts in the other languages have to be segmented. See also Subsection 3.1.1, on harmonization of the segmentation, for more examples of our final decisions across the three languages.

Coordination and Ellipsis. Coordinated clauses are separated into two segments, including cases where the subject is elliptical in the second clause. In Spanish and Basque, both pro-drop languages, this is in fact the default for both first and second clause, and therefore we see no reason why a clause with a pro-drop subject cannot be an independent unit. We follow the same principle for English. In (8), the first two EDUs in Spanish are coordinated with an elliptical subject in both cases, referring to the authors (*venimos traduciendo,* '[we] have been translating' and *queremos expresar,* '[we] wish to indicate'). They constitute separate EDUs. In the English and Basque versions, the two clauses are expressed as separate sentences.

(8) a. [To attain this goal we have been translating doctrinal texts in law at the University of Deusto since 1994.] [We wish to indicate the difficulties we have had over the years and also our achievements,] [if there can be said to be any.]

b. [Para poder alcanzar ese objetivo en la Universidad de Deusto venimos traduciendo textos doctrinales del campo del Derecho desde 1994] [y queremos expresar las dificultades que hemos tenido a lo largo de estos años y, así mismo, también los logros conseguidos,] [si es que realmente los ha habido.]

c. [Xede hori iristeko, 1994. urteaz geroztik, Deustuko Unibertsitatean Zuzenbidearen inguruko testu doktrinalak itzultzen dihardugu.] [Esperientzia horretan izandako zailtasunak eta,] [halakorik izanez gero,][27] [lorpenak ere azaldu nahi ditugu.] TERM25_BSQ

Coordinated verb phrases (VPs) or verbs do not constitute their own EDUs. We differentiate coordinated clauses from coordinated VPs because the former can be independent clauses with the repetition of a subject; the latter, in the second part of the coordination, typically contain elliptical verbal forms, most frequently a finite verb or modal auxiliary.

Relative, Modifying and Appositive Clauses. We do not consider that relative clauses (restrictive or non-restrictive), clauses modifying a noun or adjective, or appositive clauses constitute their own EDUs. We include them as part of the same segment together with the element that they are modifying. This departs from RST practice, where (restrictive) relative clauses are often independent spans, as seen in many of the examples in the original literature and the analyzes on the RST web site (Mann and Thompson, 1988; Mann and Taboada, 2010). We found that relative clauses and other modifiers often lead to truncated EDUs, resulting in repeated use of the Same-unit relation (see Truncated EDUs in 4 subsection), and thus decided that it was best to not elevate them to the status of independent segments.

An example is presented in (9), where the relative clause is in parentheses in the Spanish original. Note, however, that the coordinated clauses (with an elliptical subject in all cases) are independent segments, as explained above. In Basque, on the other hand, the relative clause is translated as an independent clause with a finite verb (*mugatzen da,* '[it] is limited to'). We have not segmented it in Basque, to agree with the other two languages.

(9) a. [. . .] [Internet terminology extends beyond the bounds of its specialist field (which by definition is part of the lexicon of science and technology)] [and breaks into general language.]

b. [. . .] [la terminología de Internet traspasa los límites del área de especialidad (a la que se circunscribe por definición el léxico científico y técnico)] [e irrumpe en la lengua de uso general,] [. . .]

c. [...] [espezialitateko eremuaren mugak gainditzen dituela Interneteko terminologiak (espezialitatera mugatzen da, definizioz, lexiko zientifiko eta teknikoa),] [eta erabilera orokorreko hizkeran sartzen dela indartsu;] [. . .] TERM38_SPA

Parentheticals. The same principle applies to parentheticals and other units typographically marked as separate from the main text (with parentheses or dashes). They do not form an individual span if they modify a noun or adjective as in Example 10, but they do if they are independent units, with a finite verb. Such is the case in (11), with a full sentence in the parenthetical unit (in English, composed of three finite clauses: *can. . . be represented*, *is* and *are*).

[27] Truncated EDU. English translation: 'if there can be said to be any' (See Subsection 4).

(10) a. The analysis of the data at hand —international terms most of which have not yet been standardized in Serbian— indicate that a hierarchy of criteria for evaluating the terms, (. . .). TERM18_ENG

(11) a. [The design and management of terminological databases pose theoretical and methodological problems] [(how can a term be represented?] [Is there a minimum representation?] [How are terms to be classified?),] (. . .)

b. [Efectivamente, el diseño y la gestión de las bases de datos terminológicos plantean problemas diversos tanto de índole teórica y metodológica] [(¿cómo se representa un término?,] [¿existe una representación mínima?,] [¿cómo se clasifican los términos?)] (. . .)

c. [Hala da, terminologiako datu-baseak diseinatzeak eta kudeatzeak hainbat arazo dakar bai teoria eta metodologiaren aldetik] [(nola adierazi terminoa?] [Ba al da gutxieneko adierazpenik?] [Nola sailkatu terminoak?),] (. . .) TERM29_SPA

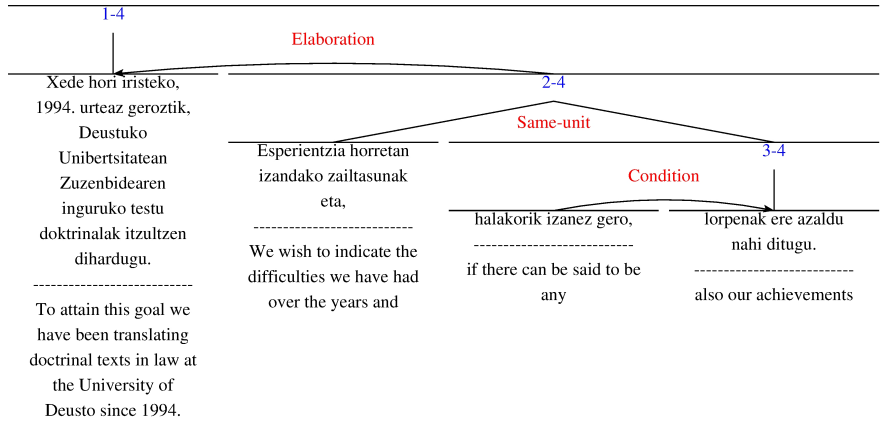
Reported Speech. We believe that reported and quoted speech do not stand in rhetorical relations to the reporting units that introduce them, and thus should not constitute separate EDUs, also following clear arguments presented elsewhere (da Cunha and Iruskieta, 2010; Stede, 2008a). This is in contrast to the approach in the RST Discourse Treebank [Carlson et al, 2003], where reported speech (there named ATTRIBUTION) is a separated EDU. There are, in any case, no examples of reported speech in our corpus.

Truncated EDUs. In some cases, a unit contains a parenthetical or inserted unit, breaking it into two separate parts, which do not have any particular rhetorical relation between each other. In those cases, we make use of a non-relation label, Same-unit, proposed for the RST Discourse Treebank [Carlson et al, 2003].

We see one such example in (11) above. The element that corresponds to the third unit in English is, in fact, inserted in the middle of the second unit in Basque. In order to align or harmonize segmentation and to preserve the integrity of that unit, we use the Same-unit (non) relation, as shown in Figure 8, which follows the Basque word order.

Once our segmentation criteria were established and the three annotators carried out the segmentation, the three segmentations were compared in terms of precision and recall. In this way, we quantified agreement and disagreement across segmentations. Moreover, we analyzed the main causes of the disagreements. Results are shown in Subsection 3. After the segmentation agreement evaluation, we harmonized the segmentation, ensuring that units were comparable across the languages. At this point, we also calculated linguistic distance between the pairs of languages, We understand linguistic distance as "the extent to which languages differ from each other" [Chiswick and Miller, 2005, pg. 1]. Although this concept is well known among linguists, there is not a single measure to evaluate this distance Chiswick and Miller [2005]. In our work, in order to measure this distance we calculated which language required the most changes in the harmonization process. This harmonization process was necessary to start out the analysis with similar units, and to avoid confusing

Fig. 8 Example of a Same-unit (non) relation



analysis disagreement and segmentation agreement. Marcu et al [2000] and Ghorbel et al [2001] also align (which we termed harmonize) their texts, decreasing the granularity of their segmentation to avoid complexity. With this decision, we lose some rhetorical information at the most detailed level of the tree. This does not, however, affect higher levels of tree structure. The results of this harmonization are shown in Subsection 3.1.

Acknowledgements This work has been partially financed by the Spanish projects RICOTERM 4 (FFI-2010-21365-C03-01) and APLE 2 (FFI2012-37260), and a Juan de la Cierva grant (JCI-2011-09665) to Iria da Cunha. Maite Taboada was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (261104-2008). Mikel Iruskieta was supported by the following projects: OPENMT-2 (TIN2009-14675-C03-01) [Spanish Ministry], Ber2Tek (IE12-333) [Basque Government] and IXA group (GIU09/19) [University of the Basque Country].

We would like to thank to reviewers and to Nynke van der Vliet for their comments on the evaluation method, to Esther Miranda for designing the website and to Oier Lopez de Lacalle for helping with scripts to calculate statistics.

References

Abelen E, Redeker G, Thompson SA (1993) The rhetorical structure of US-American and Dutch fund-raising letters. Text 13(3):323–350

Baker M (2004) A corpus-based view of similarity and difference in translation. International Journal of Corpus Linguistics 9(2):167–193

Bateman JA, Rondhuis KJ (1997) Coherence relations: Towards a general specification. Discourse Processes 24(1):3–49

Carlson L, Okurowski ME, Marcu D (2002) RST Discourse Treebank, LDC2002T07 [Corpus]. PA: Linguistic Data Consortium, Philadelphia

Carlson L, Marcu D, Okurowski ME (2003) Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory, Springer, Berlin, pp 85–112. Current and new directions in discourse and dialogue

Catford JC (1965) A linguistic theory of translation: An essay in applied linguistics, vol 8. Oxford University Press, New York

Cenoz J (2003) The role of typology in the organization of the multilingual lexicon, Springer, New York, pp 103–116. The multilingual lexicon

Chesterman A (1993) From 'is' to 'ought': Laws, norms and strategies in translation studies. Target 5(1):1–20

Chesterman A (1997) Memes of translation: The spread of ideas in translation theory. 22, Benjamins, Amsterdam and Philadelphia

Chiswick BR, Miller PW (2005) Linguistic distance: A quantitative measure of the distance between english and other languages. Journal of Multilingual and Multicultural Development 26(1):1–11

Cui S (1986) A comparison of English and Chinese expository rhetorical structures. PhD thesis, UCLA

da Cunha I, Iruskieta M (2010) Comparing rhetorical structures in different languages: The influence of translation strategies. Discourse Studies 12(5):563–598

da Cunha I, Torres-Moreno JM, Sierra G, Cabrera-Diego LA, Castro-Rolón BG (2011) The RST Spanish Treebank On-line Interface. In: International Conference Recent Advances in NLP, Bulgaria

Delin J, Hartley AF, Paris C, Scott DR, Linden KV (1994) Expressing procedural relationships in multilingual instructions. In: Seventh International Workshop on Natural Language Generation, Association for Computational Linguistics, pp 61–70

Delin J, Hartley AF, Scott DR (1996) Towards a contrastive pragmatics: Syntactic choice in English and French instructions. Language Sciences 18(3-4):897–931

Egg M, Redeker G (2010) How complex is discourse structure? In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, p 1619–1623

Fetzer A, Johansson M (2010) Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse. International Journal of Corpus Linguistics 15(2):240–266

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378–382

Flowerdew J (2010) Use of signalling nouns across l1 and l2 writer corpora. International Journal of Corpus Linguistics 15(1):36–55

Fung P (1995) Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In: 3rd Workshop on Very Large Corpora, Boston, Massachusetts, vol 78, pp 173–183

Ghorbel H, Ballim A, Coray G (2001) ROSETTA: Rhetorical and semantic environment for text alignment. In: Corpus Linguistics, Lancaster University (UK), pp 224–233

Gomez X, Simoes A (2009) Parallel corpus-based bilingual terminology extraction. In: 8th International Conference on Terminology and Artificial Intelligence, Toulouse

Granger S (2003) The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies, Rodopi, Amsterdam/New York, pp 17–29. Corpus-based approaches to contrastive linguistics and translation studies

House J (2004) Explicitness in discourse across languages, AKS, Bochum, pp 185–208. Neue Perspektiven in der Übersetzungs-und Dolmetschwissenschaft

Iruskieta M, Aranzabe MJ, Díaz de Ilarraza A, Gonzalez I, Lersundi M, Lopez de la Calle O (2013a) The RST Basque TreeBank: an online search interface to check rhetorical relations. In: 4th Workshop "RST and Discourse Studies", Brasil

Iruskieta M, Díaz de Ilarraza A, Lersundi M (2013b) Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. Corpus Linguistics and Linguistic Theory 0:1–32

Kanté I (2010) Mood and modality in finite noun complement clauses A French-English contrastive study. International Journal of Corpus Linguistics 15(2):267–290

Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: MT summit, Phuket, Thailand

Kong KCC (1998) Are simple business request letters really simple? A comparison of Chinese and English business request letters. Text 18(1):103–141

Mann WC, Taboada M (2010) RST web-site. URL http://www.sfu.ca/rst/, Accessed: 2012-09-30

Mann WC, Thompson SA (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse 8(3):243–281

Marcu D (2000a) The rhetorical parsing of unrestricted texts: A surface-based approach. Computational Linguistics 26(3):395–448

Marcu D (2000b) The theory and practice of discourse parsing and summarization. The MIT press, Cambridge

Marcu D, Carlson L, Watanabe M (2000) The automatic translation of discourse structures. In: 1st North American chapter of the Association for Computational Linguistics conference, Morgan Kaufmann Publishers Inc., Seattle (USA), pp 9–17

Maxwell M (2010) Limitations of corpora. International Journal of Corpus Linguistics 15(3):379–383

Maziero EG, Pardo TAS (2009) Automatização de um método de avaliação de estruturas retóricas. In: RST Brazilian Meeting, São Paulo, Brazil

Mitocariu E, Anechitei DA, Cristea D (2013) Comparing Discourse Tree Structures, Springer, pp 513–522. Computational Linguistics and Intelligent Text Processing

Mohamed AH, Omer MR (1999) Syntax as a marker of rhetorical organization in written texts: Arabic and English. International Review of Applied Linguistics in Language Teaching (IRAL) 37(4):291–305

Morin E, Daille B, Takeuchi K, Kageura K (2007) Bilingual terminology mining-using brain, not brawn comparable corpora. In: Annual meetings ACL, Prague, vol 45, pp 664–671

Mortier L, Degand L (2009) Adversative discourse markers in contrast: the need for a combined corpus approach. International Journal of Corpus Linguistics 14(3):338–366

O'Donnell M (2000) RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In: First International Conference on Natural Language Generation INLG '00, ACL, Mitzpe Ramon, vol 14, pp 253–256

Pardo TAS (2005) Métodos para análise discursiva automática. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP

Ramsay G (2000) Linearity in rhetorical organisation: A comparative cross-cultural analysis of newstext from the People's Republic of China and Australia. International Journal of Applied Linguistics 10(2):241–258

Ramsay G (2001) Rhetorical styles and newstexts: A contrastive analysis of rhetorical relations in Chinese and Australian news-journal text. ASAA E-Journal of Asian Linguistics and Language-teaching 1(1):1–22

Salkie R, Oates SL (1999) Contrast and concession in French and English. Languages in Contrast 2(1):27–56

Sarjala M (1994) Signalling of reason and cause relations in academic discourse. Anglicana Turkuensia 13:89–98

Scott DR, Delin J, Hartley AF (1998) Identifying congruent pragmatic relations in procedural texts. Languages in Contrast 1(1):45–82

Soricut R, Marcu D (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, Association for Computational Linguistics, vol 1, pp 149–156

Stede M (2008a) Disambiguating rhetorical structure. Research on Language and Computation 6(3):311–332

Stede M (2008b) RST revisited: Disentangling nuclearity, John Benjamins, Amsterdam and Philadelphia, pp 33–57. 'Subordination' versus 'coordination' in sentence and text

Taboada M (2004a) Building coherence and cohesion: Task-oriented dialogue in English and Spanish. John Benjamins, Amsterdam and Philadelphia

Taboada M (2004b) Rhetorical relations in dialogue: A contrastive study, John Benjamins, Amsterdam and Philadelphia, pp 75–97. Discourse across Languages and Cultures

Taboada M, Mann WC (2006a) Applications of Rhetorical Structure Theory. Discourse Studies 8(4):567–588

Taboada M, Mann WC (2006b) Rhetorical Structure Theory: looking back and moving ahead. Discourse Studies 8(3):423–459

Taboada M, Renkema J (2008) Discourse relations reference corpus. Simon Fraser University and Tilburg University, URL http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html, Accessed: 2012-09-30

Trask RL (1997) The history of Basque. Routledge, London

Usoniene A, Soliene A (2010) Choice of strategies in realizations of epistemic possibility in English and Lithuanian A corpus-based study. International Journal of Corpus Linguistics 15(2):291–316

UZEI, HAEE-IVAP (1997) International Congress on Terminology. UZEI; HAEE-IVAP, Donostia and Gasteiz

van der Vliet N (2010) Inter annotator agreement in discourse analysis. URL http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/

Wu D, Xia X (1994) Learning an English-Chinese lexicon from a parallel corpus. In: First Conference of the AMTA, Citeseer, Columbia, pp 206–213

Xiao R (2010) How different is translated Chinese from native Chinese? A corpus-based study of translation universals. International Journal of Corpus Linguistics 15(1):5–35

10

The RST Basque TreeBank: an online search interface to check rhetorical relations

The RST Basque TreeBank: an online search interface to check rhetorical relations

Mikel Iruskieta[1], María Jesús Aranzabe[2], Arantza Diaz de Ilarraza[3], Itziar Gonzalez-Dios[3], Mikel Lersundi[2], Oier Lopez de Lacalle[3]

[1]Department of Didactics of Language and Literature
University of the Basque Country (UPV/EHU)
Postcode 48940 – 0034.94601.7569 – Leioa – Basque Country

mikel.iruskieta@ehu.es

[2]Department of Basque Language and Communication (UPV/EHU)

[3]Department of Computer Science (UPV/EHU)

Abstract. *This paper introduces the first Basque discourse TreeBank annotated with rhetorical relations following Rhetorical Structure Theory. We report the main features of the corpus, such as the annotation criteria, inter-annotator agreement and harmonization procedure. We describe an online search system to check the annotation of discourse relations.*

1. Introduction

In computational linguistics discourse analysis covers a wide range of structural phenomena, such as identification of referential and relational structures. The main task when studying referential structures is correference resolution [Mitkov 2002, Recasens et al. 2010] while relational structures are related to coherence relation assignment [Asher and Lascarides 2003, Mann and Thompson 1988].

Annotated corpus are necessary in order to build advanced applications such as automatic text generation systems [Bouayad-Agha 2000], automatic summarizers [Marcu 2000b] or machine translation systems [Marcu et al. 2000]. These systems rely on different linguistic information, including the discourse level. Consequently, it is important to have a corpus which is annotated at different linguistic levels. Aforementioned systems could take advantage of the available *automatic discourse analyzers* [Marcu 2000b, Pardo et al. 2004], in order to improve their output.

There are a few works that deal with the annotation of referential structures for corpus written in languages such as English [Carlson et al. 2002, Taboada and Renkema 2011], German [Stede 2004], Dutch [van der Vliet et al. 2011], Portuguese [Pardo and Seno 2005] and Spanish [da Cunha et al. 2011a].

In the case of corpus annotation for Basque, we can find studies on referential structure [Goenaga et al. 2012, Ceberio et al. 2009] and relational structure [Iruskieta et al. 2013, Iruskieta et al. 2011]. From the linguistic point of view it is interesting to study languages with a different typology as Basque and to offer annotated corpus to the scientific community.

This work is the first RST corpus for Basque created to serve as a reference for several NLP applications for this language. The annotations follow the RST theory introduced by [Mann and Thompson 1988]. From our point of view: *i*) RST facilitates the

Anais do IV Workshop "A RST e os Estudos do Texto", páginas 40–49, Fortaleza, CE, Brasil, Outubro 21–23, 2013.

representation of coherence in real texts, establishing relations among all the units in a tree-like structure; *ii*) RST has been applied to different languages and used for advanced applications and, *iii*) there are tools which facilitate working with RST annotated corpora: RSTTool [O'Donnell 2000] and Rhetorical DataBase [Pardo 2005]. We present the annotated corpus and we describe an online search interface to check the annotated discourse structure.

The remainder of this paper is structured as follows. Section 2 lays out the theoretical framework and Section 3 the methodology utilized to annotate the corpus. Section 4 sets out the results of the annotation and presents the online search interface. Finally, Section 5 presents the discussion and establishes directions for future work.

2. Annotation in Rhetorical Structured Theory

Rhetorical Structured Theory is a language-independent theory describing coherence between text fragments. It combines the idea of nuclearity, i.e. the importance of an individual fragment from within the discourse, with the presence of rhetorical relations (R) (hypotactic and paratactic relations) between these fragments. Hypotactic and paratactic relations connect discourse units, either a single unit (EDU) or groups of units (span). According to the theory, these relations can be paratactic (N-N) —when they establish relations between fragments that are equally important to the author (LIST, CONTRAST, DISJUNCTION, etc.)— or hypotactic (N-S) —when they connect a less-important unit with a unit the author views to be more important (ELABORATION, MEANS, PREPARATION, CONCESSION, CAUSE, RESULT, etc.). Relations are defined in light of the restrictions established between the nucleus and satellite and by describing the effect they have on the reader. A more detailed explanation of RST can be found in [Mann and Thompson 1988] and in [Mann and Taboada 2010].

Refering to the annotation process, it is well known that agreement is higher when there is training among coders. Works in which annotators did not have a training phase present a similar agreement [van der Vliet et al. 2011]. This fact is reported in the work carried out on the English language [Carlson et al. 2003]; a total of six professional annotators tagged the corpus measuring inter-annotator agreement in different texts (53 to be precise) in a pairwise manner (and in a few cases three-wise manner). There are methods for improving inter-annotator agreement: in [Carlson et al. 2003], for example, it is reported that at the beginning of the project the highest level of agreement attained between the three annotators in a small sample was a Kappa score of 0.602, while at the end of the project, after training, it was 0.755. In this project, in addition to the professional annotators, the authors also measured the agreement between two non-profesional annotators, with very different results: Kappa scores of between 0.597 and 0.792 (1918 EDUs, 30 texts).

The size of the corpus is another aspect to take into acount. We can say that, while the size of our corpus is smaller than that of the corpora found in the bibliography, the fragment tagged in a pairs was comparable as regards both size and number of annotators.

Although the delivery phase is important in annotation [Hovy 2010], it is usually forgotten. This is not the case in the RST Spanish Treebank [da Cunha et al. 2011b]. Relation extraction from a corpus is very helpful for a better understanding of the relation itself or for the study of patterns (this information will be useful to be on the design

of automatic rules or as features in machine learning algorithms). In the RST Basque TreeBank the delivery phase is of great importance as we will see in the Section 4.

3. Methodological principles

Our corpus is composed by abstracts, short but well structured texts, written in Basque.[1]

Regarding coherence relations, abstracts function as independent discourse and summarize the main idea of the paper. The percentages of each relation —which are available on the web— are similar to the ones of [Pardo and Nunes 2004].

As regards relational structure, agreement between annotators was measured manually, using the evaluation system based on rhetorical relations presented in [da Cunha and Iruskieta 2010]. We decided not to use the evaluation system that assesses the tree structure [Marcu 2000a], mainly in order to avoid the shortfalls described in [Iruskieta et al. 2013]. According to these authors, span and nuclearity factors are not independent phenomena in the tree structure evaluation proposed in [Marcu 2000a], since they influence the evaluated factor of rhetorical relations. In contrast, [da Cunha and Iruskieta 2010] propose an evaluation method based on rhetorical relations where three factors are assessed: satellite unit or composition span (C), nuclear unit or attachment span (A),[2] and rhetorical relations (R).

3.1. Annotated corpus

The corpus utilized in this study is composed of abstracts from three specialized domains: medicine, terminology and science. Medical texts include the abstracts of all medical articles written in Basque in the Medical Journal of Bilbao (GMB) between 2000 and 2008. Texts related to terminology were extracted from the proceedings of the International Conference on Terminology (TERM) organized in 1997 by UZEI, while scientific articles are papers from the University of the Basque Country's Faculty of Science and Technology (ZTF) Research Conference, which took place in 2008. We have collected 60 documents that contain 15566 words (803 sentences). The created gold standard contains 1355 EDUs and 1292 Rs.

3.2. Annotators

The corpus was annotated by two linguists. The two annotators had previously annotated other linguistic levels (morphosyntax, syntax and semantics), and were familiar with RST and its annotation interface, RSTTool, but no training was provided.

3.3. Annotation phases

The process of tagging the rhetorical structure was divided into four phases. Each phase was evaluated and harmonized by a judge, in order to ensure that all annotators started each new phase from the same basic criteria. The four phases were as follows:

i) **Segmentation:** annotators were asked to divide the text into EDUs; in general, each EDU is either a subordinate clause containing a verb or an independent clause (more details in [da Cunha and Iruskieta 2010]).

[1] In the same sense as [Swales 1990] mentions that abstracts follows an IMRaD (*Introduction*, *Method*, *Results* and *Discussion*) structure.

[2] In multinuclear relations any of the nucleus can be considered as composition or attachment span.

ii) **Identifying the macrostructure:** before identifying the rhetorical relations, annotators were asked to identify most important part of the text or central unit (CU).

iii) **Representing the relational structure:** bearing in mind the CU, rhetorical structure was annotated in a modular and incremental way as proposed in the work by [Pardo 2005] and with the extended classification of rhetorical relations [Mann and Taboada 2010].

iv) **Annotating the signals of relations:** one annotator has tagged the signals of rhetorical relations, as proposed in [Taboada and Das Forthcoming]. The cause subset (CAUSE, RESULT and PURPOSE) was annotated by two annotators and evaluated.

The method mainly used in RST to increase annotator agreement on rhetorical relations is to establish a training phase. From our point of view this could carry a circular process between relations and their signals [Spenader and Lobanova 2009]. To provide a more reliable annotated corpus and do not fall in this circular problem, we analyzed the problems arising amongst annotators, and, in order to achieve our aim (a reference corpus annotated with relational structure), we established the criteria for annotation and we designed a manual for a judge to decide the cases of disagreement.

3.4. Results

We carried out an evaluation to assess each of the annotation steps by means of different agreement measures. This way, we calculated the agreements of segmentation (EDU), the agreement on CU identification, the agreement on rhetorical structure and the agreement on signals of the cause subset. At the rhetorical structure level we provide an analysis of the source of the disagreement, categorizing them in different types.

Segmentation (EDU). Inter-annotator agreement between annotators is 81.35%.

CUs identification. The overall mean agreement between annotators is 81.67%.[3]

Relational structure level. Based on the factors we defined —composition span (C), attachment span (A) and rhetorical relations (R)— the following types of agreements: *i*) **CAR**: agreement in composition span, attachment span and relation, *ii*) **CR**: agreement in composition span and relation, *iii*) **AR**: agreement in attachment span and relation and *iv*) **R**: agreement only in relation. Table 1 shows the agreement level obtained on the four types of measurements.

Agree	K. α	%	Gain
CAR	0.394	47.76%	-
CR	0.458	54.03%	6.27%
AR	0.431	51.17%	3.41%
R	0.561	61.47%	13.71%

Table 1. Types of agreement

Disagree	%	Disagree	%
No-Match	0.23%	Different R	13.62%
Nuclearity	6.73%	Similar R	5.88%
N/N-N/S	8.90%	MissMatch R	2.01%
Attachment	0.08%	Specificy	0.93%
Composition	0.15%	Segmentation	0.15%

Table 2. Types of disagreement

The results show how the agreement increases as the relaxation of the agreement increases too, being CAR the most demanding agreement, and R the more relaxed one.

[3] Agreement related to CU has been different in the three domains. The agreement is related to the number of candidates (text size) and to the enough explicit linguistic evidence which highlights the CU.

The inter-annotator agreement level [Krippendorff 2012] is moderate for relations. It must be noted that we are in the initial phase of the annotation project. Nevertheless, the results obtained are comparable to those achieved in the initial phases of the main work of rhetorical relation annotation carried out for English [Carlson et al. 2003].

On the other hand, we defined different types of disagreement, taking into account the following phenomena: *i*) **No-match**: The composition of the tree results in relations that cannot be compared. *ii*) **Nuclearity**: Different choices in nuclearity entailed discrepancy in hypotactic relations. *iii*) **N/N vs N/S**: Different choices in nuclearity entailed a paratactic/hypotactic mix-up. *iv*) **Attachment span**: Different choices in attachment span entailed a different relation. *v*) **Different R**: A relation has the same composition and attachment span, but not the same relation. *vi*) **Similar R**: Relations chosen are similar in nature. *vii*) **Mismatch R**: Relations with mismatched RST trees. *viii*) **Specificity**: The relation chosen is more specific in one annotation than in the other. *ix*) **Segmentation**: Segmentation does not match.

As shown in Table 2, although the Different R label is the main source of disagreement (13.62% of the times), one of the main disagreement comes from the choice of nuclearity: in total, 15.63% of the annotation disagree on Nuclearity or the N/N-N/S factors. The other types of disagreement (the 8.82% of the annotations) can easily be resolved explaining how the annotator understand the relations involved in Similar R, Mismatch R and Specificity labels.

Signals for rhetorical relations. Finally, a judge resolved the disagreements between annotators, establishing the relational structure model and specifying the signals for rhetorical relations. The average agreement between annotators of the cause subset —which is often signalled— was 78.11% (PURPOSE 90%, CAUSE 76.79% and RESULT 59.7%).

4. The RST Basque TreeBank

When entering in the website,[4] you can find information of the general characteristics of the RST Basque TreeBank and facilities to consult the contents of the tagged corpus, as for example: *i*) discourse units, the central unit and relations linked to the central unit (4.1 subsection); *ii*) all instances of a selected rhetorical relation in the corpus (4.2 subsection); *iii*) the rhetorical structure of a desired text (4.3 subsection); *iv*) all the signals of relations (4.4 subsection) and, *v*) searching facilities for further studies about typical patterns about combination of word-forms, lemma and POS present in the corpus (4.5 subsection).

4.1. Consulting EDUs and CU of a tree

The application offers the possibility to check the linear segmentation (EDUs) of a document as well as its CU. Table 3 shows the segmentation for the GMB0301 document. The text has seven EDUs[5] and the last one, EDU_7, has a button called *See* in the CU column. If you click on this button, you will see all the relations linked to the CU of this text.

4.2. Dealing with rhetorical relations

The web application allows you to look up all the occurrences of a specific relation, or restrict your search to a particular sub-corpus (GMB, TERM or ZTF). If the segments are

[4] http://ixa2.si.ehu.es/diskurtsoa/en/

[5] Translations thereof are found underneath these.

GMB0301-GS.rs3 (7)			
EDU	**Segment**	**Annotator**	**CU**
1	Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.	GS	
	Recurrent aphthous stomatitis (I): epidemiologic, etiologic and clinical features.		
2	"Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da.	GS	
	"Recurrent aphthous stomatitis" is one of the most frequent oral pathologies.		
3	tamainu, kokapena eta iraunkortasuna aldakorra izanik.	GS	
	having a variable size, location and duration.		
4	Honen etiologia eztabaidagarria da.	GS	
	It has a controversial etiology.		
5	Ultzera mingarri batzu bezela agertzen da,	GS	
	It is characterized by the apparition of painful ulcers,		
6	Hauek periodiki beragertzen dira.	GS	
	These ulcers appear recurrently.		
7	Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.	GS	See
	In this paper we analyze the most important epidemiological, etiological, pathological and clinical features of this common oral pathology.		

Table 3. Example of the EDUs section, GMB0301

very long and you are only interested in the beginning of each, you can also limit the size.

Table 4 shows a fragment of a search conducted in the relation database. Since the search was limited to the TERM corpus, there are only 27 CAUSE relations, rather than the 56 shown in corpus. The first 3 columns of Table 4 describe the order and direction of the discourse units. Since the segments —left span and right span— follow the order in where they appear in the text, the second column specifies the nuclearity of the relations: if the relation is NS (nucleus on the left and satellite on the right), then the arrow points left (<–), towards the nucleus. If it is SN, then the arrow points right (–>). The fourth column specifies the relation and relation type: in this case, a single nucleus relation (N/S) CAUSE; when there are multiple nuclei, this is indicated by the letters (N/N). Finally, the source of the example (Ref.) and annotator (Annot.) is specified.[6]

Relation: Cause (27)					
Left span	**NS**	**Right span**	**Relation**	**Ref.**	**Annot.**
Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta horren berri eman dute: ingeleseko unita[...]	<–	Izan ere, iritzi ezberdinetako zientzialari serbiarrek adostasuna lortu dute eta aurreko hamarkadetan ingelesari eman diote [...]	Cause	TERM18	GS
In recent decades, many Serbian researchers working in different scientific fields have noticed a tendency and this is outlined here: the English unit [...]		Indeed, Serbian scientists from different schools of thought have reached a consensus and have given English [...]			
Terminologiak berak ere, uztartu egin behar ditu joera orokor horiek, eransten zaizkien beste batzuekin batera, hala nola: teknologien [...]	<–	gizartearekin lotuta dagoen jarduera denez,	Cause	TERM19	GS
Terminology itself must seek to unite these general trends, along with others related to them, for example: technology [...]		since it is an activity linked to society,			

Table 4. Example of a CAUSE relation search

[6]Note: due to space limitations we only mention here the most important information contained in the database. The signals for rhetorical relations are underlined in Table 4.

4.3. Checking all relations of a RST tree

You can also consult the database file by file: viewing the rhetorical relations of the chosen file or its image in JPG format. The rhetorical structure can be consulted in different formats (XML and RS3). Other information can be consulted here: text file in TXT format, morphosyntactic information annotated automatically in KAF format [Bosma et al. 2009], and the signals for relations annotated in RHETDB format.

4.4. Signals of rhetorical relations

You can check if a signal is in more that one relation. We show as an example a query based on the adversative conjunction *baina* 'but' in Table 5, which signals two similar relations (CONTRAST and CONCESSION).[7]

Signal: *baina* 'but'			
Gainerakoan, prokasu adierazle egokiak daude,	Kontzesioa	baina altan dagoen gaixoaren ahalmen funtzionalaren erregistro urria antzematen da,	GMB0504
With respect to the other aspects, the indicators of process are good	Concession	but there is poor recording of the patient's functional capacity on discharge,	
Bestalde, Euskaltzaindiak hitz elkartuen bidea (1995eko urtarrilaren 27an onartutako araua) proposatzen du adjektibo erreferentzialak itzultzeko,	Kontrastea	baina arauan bertan esaten denez, "...ahal den guztian...",	TERM22
Euskaltzaindia proposed a mechanism of compound words (in a standard approved on January 27th 1995) for the translation of referential adjectives.	Contrast	However the academy also confirmed, ..."whenever possible",	

Table 5. Example of the SIGNALS section, the discourse marker *baina* 'but'

4.5. Word form, lemma and POS search interface

Searches combining word-form, lemma and POS features can be done in the application due to the fact that all the words in the texts have associated morphological and syntactical information in KAF format.

	Doc.	Sent Id	Word	CU	Sentence
1	TERM50	sent2	taldeek / helburua	BAI	[...] Hitzaldi honek azken hiru urteotan lau unibertsitate hauen *talde*ek egindako ikerkuntzaren ondorioetako batzuk azaltzeko *helburua* izango luke.
			groups / aim	YES	"[...] The aim of this talk is to present some of the results of the research carried out by groups from these four universities over the last three years."
2	ZTF13	sent1	taldearen / helburu	BAI	[...] Gure *ikerkuntza talde*aren *helburu* nagusia, [...]
			group's / aim	YES	[...] Our research group's principal aim, [...]
3	ZTF13	sent17	taldearen / helburu	EZ	Alor honetan, gure *ikerkuntza talde*aren *helburu* nagusiak bi dira.
			group's / aim	NO	In this field, our research group has two main aims.
1	ZTF15	sent7	helburu / talde	EZ	[...] bestelako galdera zailagoei ere erantzutea dute *helburu*, hala nola, espezieen biogeografia, *talde*aren filogenia, eta abar.
			aim / group	NO	[...] the aim is to answer other such difficult questions, such as species biogeography, group phylogeny, etc.

Table 6. Example of the SEARCH section

These searches provide the option of searching patterns. For example, in a two-word search, you can specify to show the sentences which contain words starting with the forms *talde* 'group' or 'team' and *helburu* 'goal' or 'aim'. You can also define whether or not other words can be located between the target terms. Table 6 shows a search for the

[7] More information about ambiguity in this corpus can be read in [Iruskieta and da Cunha 2010] and in [Iruskieta et al. 2009].

terms *talde* 'group' and *helburu* 'aim' results in two YES responses for CU, but another search with the terms the other way round (aim and group) would only give one NO response for CU.

5. Discussion and Future Work

This paper presents the first RST Basque TreeBank, where the gold standard files that have been used to compile the database are at the disposal of anyone who wishes to use them. Moreover, the study also served to design the harmonization processes for the different annotation phases (segmentation, identification of central units, rhetorical relations and its signals), as well as giving the judge the opportunity of consulting both their annotations and those of the annotators, seeing at a single glance the frequency of each relation and its signals. This in turn enabled the detection of errors and incoherence during the establishment of the gold standards.

The work carried out is useful for certain language processing tasks. Indeed, during the course of the project we established a segmented gold standard for 60 texts, on the road towards automatic segmentation. As regards rhetorical relations, after establishing a gold standard for 60 texts, we marked the signals of those relations, being the size of the work similar to that of others in the literature [Taboada and Das Forthcoming]. In the future, this work will help us define rhetorical relation patterns, and this in turn will help us achieve automatic detection of those most commonly signaled relations.

The authors are currently striving to achieve the following aims: in the short medium term, their goal is to annotate texts from another genre: newspaper articles, texts from the EPEC corpus and to study deeply the signals of relations in the RST Basque TreeBank. With the data provided by the RST Basque TreeBank, they are implementing an automatic discourse segmentation program. Besides, and considering how time consuming the tagging and evaluation processes are, the authors are working on the implementation of a new interface to facilitate the editing of rhetorical relations and programs for automatic evaluation program based on rhetorical relations.

Acknowledgments

This study was carried out within the framework of the following projects: Ber2Tek (IE12-333); Hibrido Sint (TIN2010-20218); NewsReader project (FP7- ICT-2011-8-316404); IXA group, Research Group of type A (IT344-10). We would like to thank Esther Miranda and Kike Fernandez for their help in designing the web page.

References

[Asher and Lascarides 2003] Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge Univ Pr, Cambridge.

[Bosma et al. 2009] Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: a generic semantic annotation format. In *GL2009 Workshop on Semantic Annotation*, Italy.

[Bouayad-Agha 2000] Bouayad-Agha, N. (2000). Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts. In *Annual Meeting-ACL*, volume 38, pages 16–22.

[Carlson et al. 2003] Carlson, L., Marcu, D., and Okurowski, M. E. (2003). *Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory*, pages 85–112. Current and new directions in discourse and dialogue. Springer, Berlin.

[Carlson et al. 2002] Carlson, L., Okurowski, M. E., and Marcu, D. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus]*. PA: Linguistic Data Consortium, Philadelphia.

[Ceberio et al. 2009] Ceberio, K., Aduriz, I., Díaz de Ilarraza, A., and Garcıa, I. (2009). Empirical study of the relevance of semantic information for anaphora resolution: the case of adverbial anaphora. In *7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC09)*, pages 56–63, Goa, India.

[da Cunha and Iruskieta 2010] da Cunha, I. and Iruskieta, M. (2010). Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5):563–598.

[da Cunha et al. 2011a] da Cunha, I., Torres-Moreno, J. M., and Sierra, G. (2011a). On the Development of the RST Spanish Treebank. In *5th Linguistic Annotation Workshop (LAW V '11)*, pages 1–10, Portland, USA.

[da Cunha et al. 2011b] da Cunha, I., Torres-Moreno, J. M., Sierra, G., Cabrera-Diego, L.-A., and Castro-Rolón, B.-G. (2011b). The RST Spanish Treebank On-line Interface. In *International Conference Recent Advances in NLP*, Bulgaria.

[Goenaga et al. 2012] Goenaga, I., Arregi, O., Ceberio, K., de Ilarraza, A. D., and Jimeno, A. (2012). Automatic coreference annotation in basque. In *Eleventh International Workshop on Treebanks and Linguistic Theories*, Portugal.

[Hovy 2010] Hovy, E. (2010). Annotation: A Tutorial. In *48th Annual Meeting of the ACL*, Uppsala, Sweden.

[Iruskieta and da Cunha 2010] Iruskieta, M. and da Cunha, I. (2010). Marcadores y relaciones discursivas en el ámbito médico: un estudio en español y euskera. In *XXVIII Congreso Internacional AESLA: Analizar datos > Describir variación*, pages 13–159, Vigo.

[Iruskieta et al. 2009] Iruskieta, M., de Ilarraza, A. D., and Lersundi, M. (2009). Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso. In *Proceedings of 27th AESLA International Conference*, pages 963–971, Ciudad Real, Spain.

[Iruskieta et al. 2011] Iruskieta, M., Díaz de Ilarraza, A., and Lersundi, M. (2011). 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. 2013] Iruskieta, M., Díaz de Ilarraza, A., and Lersundi, M. (2013). A critical analysis of rhetorical annotation: fundamental principles of discourse segmentation in basque. *Corpus Linguistics and Linguistic Theory*, 0(0):1–32.

[Krippendorff 2012] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. SAGE, London.

[Mann and Taboada 2010] Mann, W. C. and Taboada, M. (2010). RST web-site. *http://www.sfu.ca/rst/*.

[Mann and Thompson 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[Marcu 2000a] Marcu, D. (2000a). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

[Marcu 2000b] Marcu, D. (2000b). *The theory and practice of discourse parsing and summarization*. The MIT press, Cambridge.

[Marcu et al. 2000] Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *1st North American chapter of the Association for Computational Linguistics conference*, pages 9–17, Seattle (USA).

[Mitkov 2002] Mitkov, R. (2002). *Anaphora resolution*, volume 134. Longman London.

[O'Donnell 2000] O'Donnell, M. (2000). Rsttool 2.4: a markup tool for rhetorical structure theory. In *6th European Workshop on Natural Language Generation*, Germany.

[Pardo 2005] Pardo, T. A. S. (2005). Métodos para análise discursiva automática. PhD Thesis. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

[Pardo and Nunes 2004] Pardo, T. A. S. and Nunes, M. G. V. (2004). Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil. Technical Report NILC-TR-04-03.

[Pardo et al. 2004] Pardo, T. A. S., Nunes, M. G. V., and Rino, L. H. M. (2004). DiZer: An automatic discourse analyzer for Brazilian Portuguese. *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234.

[Pardo and Seno 2005] Pardo, T. A. S. and Seno, E. R. M. (2005). Rhetalho: um corpus de referência anotado retoricamente. *Anais do V Encontro de Corpora*, pages 24–25.

[Recasens et al. 2010] Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *5th International Workshop on Semantic Evaluation*, pages 1–8, Sweden. Association for Computational Linguistics.

[Spenader and Lobanova 2009] Spenader, J. and Lobanova, A. (2009). Reliable discourse markers for contrast relations. In *Proceedings of the 8th International Conference on Computational Semantics*, Tilburg, The Netherlands.

[Stede 2004] Stede, M. (2004). The Potsdam Commentary Corpus. In *2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.

[Swales 1990] Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge Univ Pr, Cambridge, UK.

[Taboada and Das Forthcoming] Taboada, M. and Das, D. (Forthcoming). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*.

[Taboada and Renkema 2011] Taboada, M. and Renkema, J. (2011). Discourse Relations Reference Corpus. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

[van der Vliet et al. 2011] van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., and Redeker, G. (2011). Building a discourse-annotated dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, 3:157–171.

The abbreviated translation
of the thesis
was finished
in February 02, 2014