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.
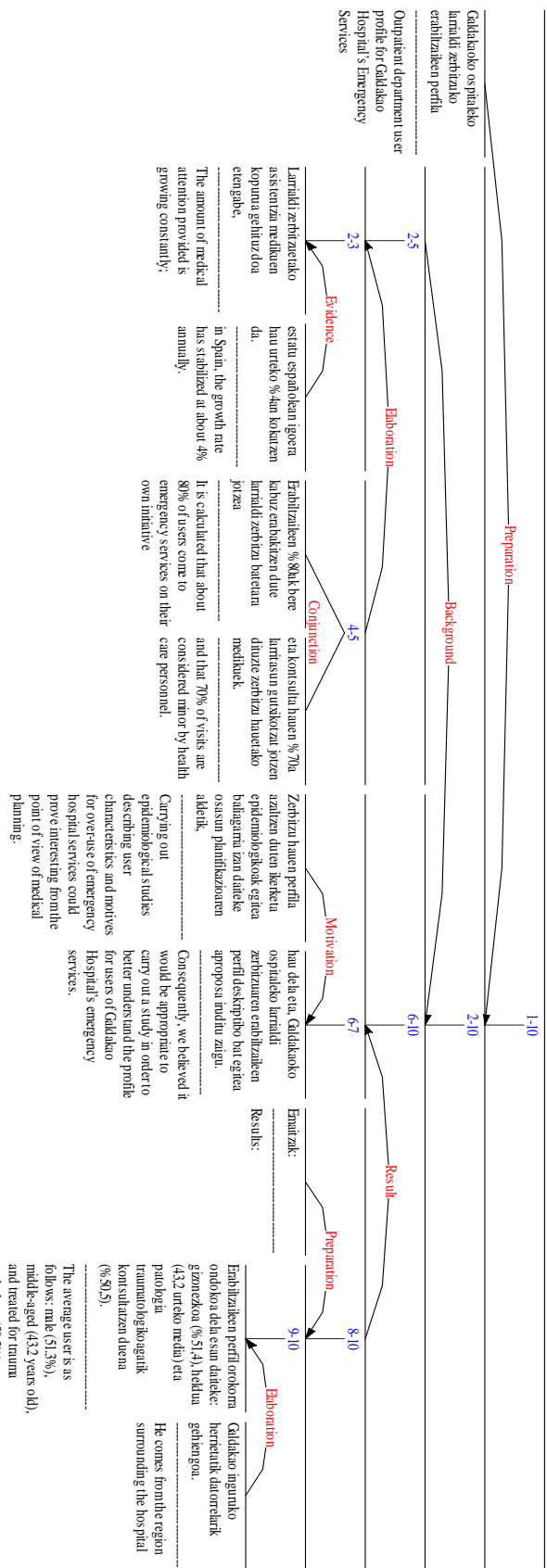
Galdakaoko ospitaleko larrialdi zerbitzuko erabiltzaileen perfila

Outpatient department user profile for Galdakao Hospital's Emergency Services

Larrialdi zerbitzuetako asistentzia medikuen kopurua gehituz doa etengabe,

The amount of medical attention provided is growing constantly;

estatu espaniolean igoera hau urteko %4an kokatzen da.

in Spain, the growth rate has stabilized at about 4% annually.

Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jo_tzea

It is calculated that about 80% of users come to emergency services on their own initiative

eta kontsulta hauen % 70a larriasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.

and that 70% of visits are considered minor by health care personnel.

Zerbitzu hauen perfila azaltzen duten ikerketa epidemiologikoak egitea baliagarria izan daiteke osasun planifikazioaren aldetik,

Carrying out epidemiological studies describing user characteristics and motives for over-use of emergency hospital services could prove interesting from the point of view of medical planning.

hau dela eta, Galdakaoko ospitaleko larrialdi zerbitzuaren erabiltzaileen perfil deskribtibo bat egitea aproposa iruditu zaigu.

Consequently, we believed it would be appropriate to carry out a study in order to better understand the profile for users of Galdakao Hospital's emergency services.

Emaitzak:

Results:

Erabiltzaileen perfil orokorra ondokoa dela esan daiteke: gizonezkoa (%51,4), heldua (43,2 urteko media) eta patologia traumatologikoagatik kontsultatzen duena (%50,5).

The average user is as follows: male (51,3%), middle-aged (43,2 years old), and treated for trauma pathology (50,5%).

Galdakao inguruko herrietatik datorrelarik gehiengoa.

He comes from the region surrounding the hospital

Relations: Preparation, Background, Elaboration, Evidence, Conjunction, Motivation, Result, Preparation, Elaboration

Node spans: 1-10, 2-10, 2-5, 2-3, 4-5, 6-10, 6-7, 8-10, 9-10

Figure 1: A rhetorical structure tree for text GMB0401

4

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	A2 No	Total
	Yes	243	0	243
A1	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.

6

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).
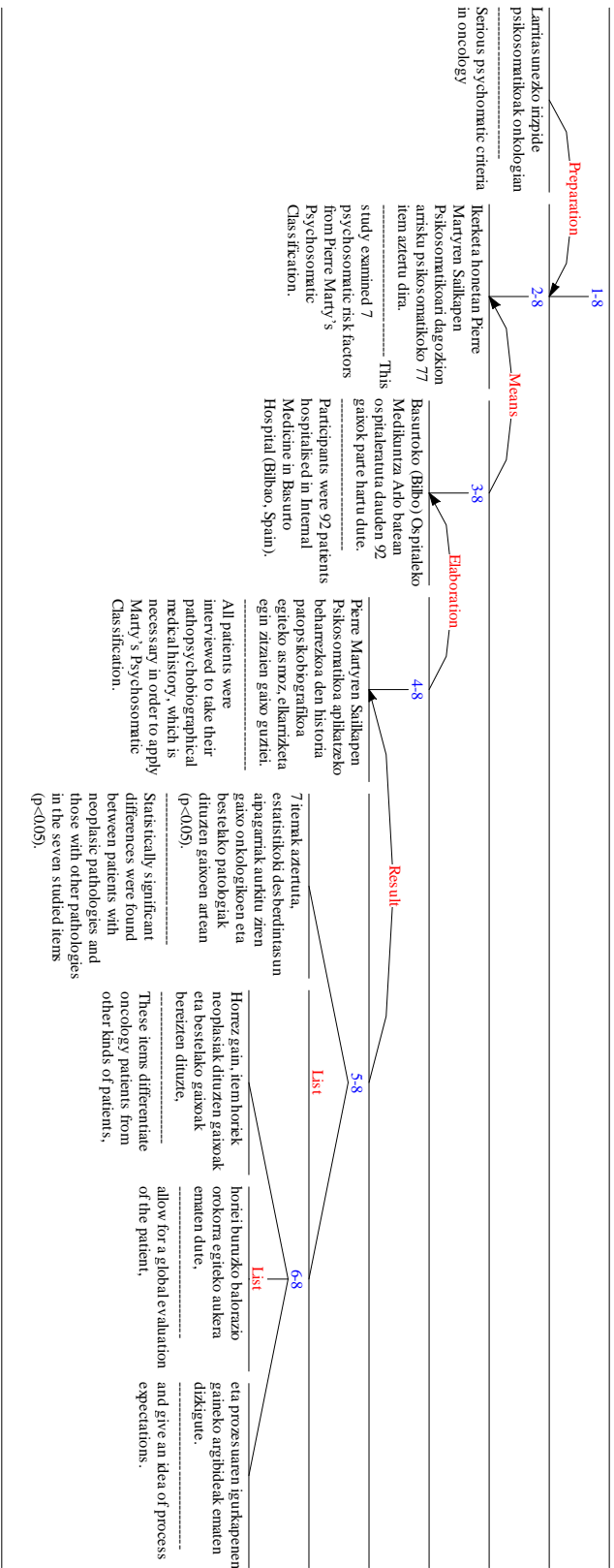
4. Evaluation of the rhetorical annotation

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).

A1

Preparation — 1-8 — 2-8 — Means — 2-4 — 3-4 — Elaboration — Result — List — 5-8 — 6-8

Ikerketa honetan Pierre Martyren Sailkapen Psikosomatikoari dagozkion arrisku psikosomatikoko 7 item aztertu dira.
———
This study examined 7 psychosomatic risk factors from Pierre Marty's Psychosomatic Classification.

Basuroko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute.
———
Participants were 92 patients hospitalised in Internal Medicine in Basurto Hospital (Bilbao, Spain).

Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, elkarrizketa egin zizaien gaixo guztiei.
———
All patients were interviewed to take their pathopsychobiographical medical history, which is necessary in order to apply Marty's Psychosomatic Classification.

7 itemak aztertuta, estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05).
———
Statistically significant differences were found between patients with neoplasic pathologies and those with other pathologies in the seven studied items (p<0.05).

Horrez gain, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte,
———
These items differentiate oncology patients from other kinds of patients,

horiei buruzko balorazio orokorra egiteko aukera ematen dute, eta prozesuaren igurikapenen gaineko argibideak ematen dizkigute.
———
allow for a global evaluation of the patient, and give an idea of process expectations.

A2

Preparation — 1-8 — 2-8 — Means — 3-8 — Elaboration — 4-8 — Result — 5-8 — List — 6-8 — List

Ikerketa honetan Pierre Martyren Sailkapen Psikosomatikoari dagozkion arrisku psikosomatikoko 7 item aztertu dira.
———
This study examined 7 psychosomatic risk factors from Pierre Marty's Psychosomatic Classification.

Basuroko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute.
———
Participants were 92 patients hospitalised in Internal Medicine in Basurto Hospital (Bilbao, Spain).

Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, elkarrizketa egin zizaien gaixo guztiei.
———
All patients were interviewed to take their pathopsychobiographical medical history, which is necessary in order to apply Marty's Psychosomatic Classification.

7 itemak aztertuta, estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean (p<0.05).
———
Statistically significant differences were found between patients with neoplasic pathologies and other kinds of patients.

Horrez gain, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte,
———
These items differentiate oncology patients from other kinds of patients,

horiei buruzko balorazio orokorra egiteko aukera ematen dute, eta prozesuaren igurikapenen gaineko argibideak ematen dizkigute.
———
allow for a global evaluation of the patient, and give an idea of process expectations.
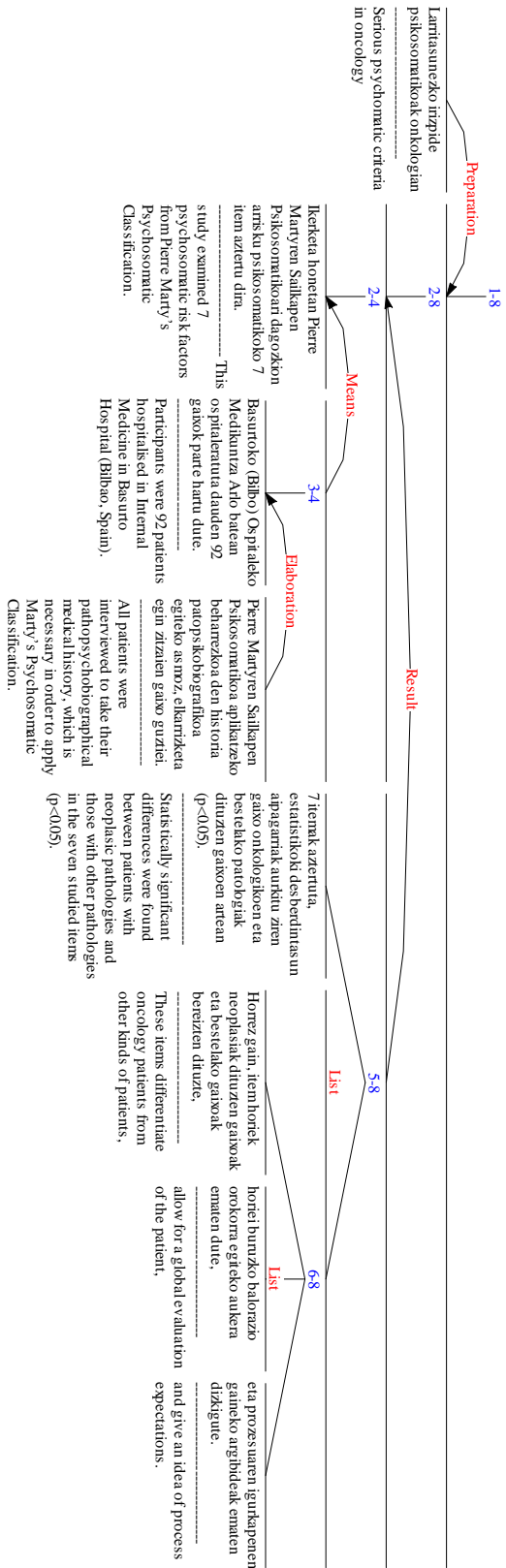
Figure 2: Text GMB0701

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

Node	EDU A1	EDU A2	Span A1	Span A2	N/S A1	N/S A2	RR A1	RR 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.

14

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.
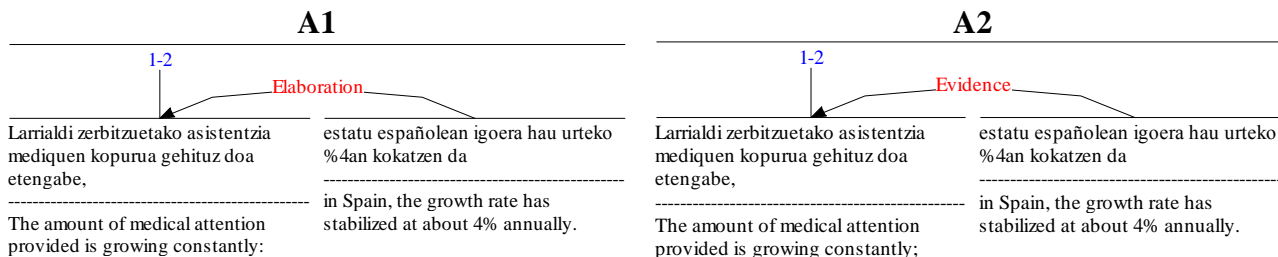

Figure 3: Disagreement regarding relation, GMB0401

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

Node	EDU A1	A2	Span A1	A2	N/S A1	A2	RR 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 structure*s]. 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.