Lang Resources & Evaluation
DOI 10.1007/s10579-014-9271-6

ORIGINAL PAPER

A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora

Mikel Iruskieta · Iria da Cunha · Maite Taboada

Received: 26 June 2013 / Accepted: 8 May 2014

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. 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 (✉)
Department of Didactics of Language and Literature, University of the Basque Country, Sarriena auzoa z/g, 48940 Leioa, Spain
e-mail: mikel.iruskieta@ehu.es

I. da Cunha
University Institute for Applied Linguistics, Universitat Pompeu Fabra, C/ Roc Boronat 138, 08018 Barcelona, Spain
e-mail: iria.dacunha@upf.edu

M. Taboada
Department of Linguistics, Simon Fraser University, 8888 University Dr, Burnaby, BC V5A 1S6, Canada
e-mail: mtaboada@sfu.ca

Published online: 28 May 2014

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. 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. Translation literature deals with many aspects 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 done 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 separately. Mitocariu et al. (2013) have proposed an evaluation method (for RST and Veins Theory Cristea et al. 1998) 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 dispersion annotation, while at the same time allows the quantitative evaluation.

The second aim of this paper is to test this method. 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), 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

[1] Soricut and Marcu (2003, pg. 152) use the term "attachment point" or "dominance set".

constructed. Moreover, the study provides rhetorical analyzes of a less-commonly studied language,[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 that 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 speaker'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

[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.

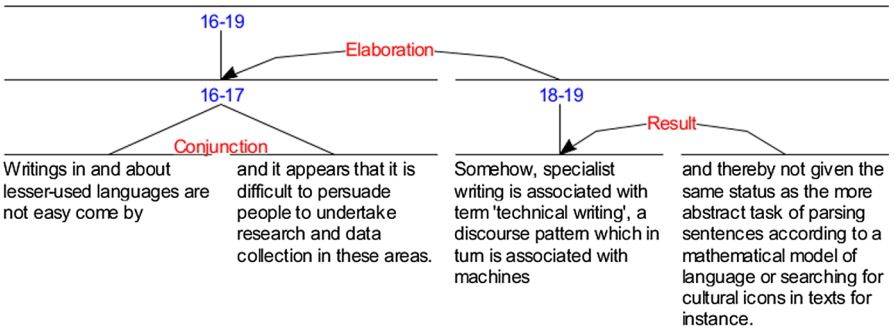

Fig. 1 Example of an RST tree, TERM30_ENG

relations (RESULT and ELABORATION). The annotator recognized that spans 16 and 17 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: 18 and 19 in Fig. 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.

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) of the text for the most important unit of an rhetorical structure tree (RS-tree) and "Central Subconstituent(s)" (CS) of a relation for the most important unit of the modifier span that is the most important unit of the satellite span. When there is a simple constituent (that is no more than one EDU), we formalized this simple constituent as the CS, and when there is a multinuclear relation, we describe it with all of its constituents.

Table 1 provides a representation of this example.

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).

Table 1 Formalization of Fig. 1, TERM30_ENG

Relation	Left span	Right span	CS	Nuclearity
Result	18	19	19	NS
Conjunction	16	17	16–17	NN
Elaboration	16–17	18–19	18	NS

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 search for similarities and differences in large corpora, 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. In sum, our corpus includes 45 texts. 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.

[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.

Table 2 Corpus statistics

Subcorpus	Annotators	Texts	Words	Sentences	EDUs
ENG	A1	15	5,706	201	318
SPA	A2	15	6,324	193	318
BSQ	A3	15	4,800	197	318

This qualitative method applies to any type of text,[5] since the principles on which it is based are 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 Sect. 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 with RST consists of segmenting the text into spans. Exactly what a span is, in the framework of 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 subcorpus (A1 segmented English texts, A2 segmented Spanish texts, and A3 segmented Basque texts).[6]

[5] It was used also to evaluate the RST Basque TreeBank (Iruskieta et al. 2013a), available at: http://ixa2.si.ehu.es/diskurtsoa/en/.

[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 if the corpus is small, the comparison work is extensive. 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. 2011a) and Basque RST TreeBank (Iruskieta et al. 2013a)].

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 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.

2.3.1 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" for a detailed explanation).

2.3.2 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.

2.3.3 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 not to elevate them to the status of independent segments.

2.3.4 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

[7] See the paragraph on Truncated EDUs in this section.

individual span if they modify a noun or adjective, but they do if they are independent units, with a finite verb.

2.3.5 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 considered as a separated EDU. There are, in any case, no examples of reported speech in our corpus.

2.3.6 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 Sect. 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 Sect. 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), a graphical interface widely used for RST annotation. 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 a 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:

1. Two of the factors evaluated, nuclearity and relation, are not independent of each other: factor conflation.
2. 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.

1. 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 Fig. 2, one of the annotators has labelled the relation as ELABORATION, and the other as EVIDENCE.

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 Fig. 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 Fig. 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]

RSTeval does not take into account the language of the rhetorical structures; however, it eliminates the stopwords of each language from the text, which are not used to build the EDUs and Spans. In the first table of Fig. 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.

[8] This evaluation method has been automated by Maziero and Pardo (2009) and nowadays it can be used in four languages: English, Spanish, Portuguese and Basque. Available at http://www.nilc.icmc.usp.br/rsteval/.

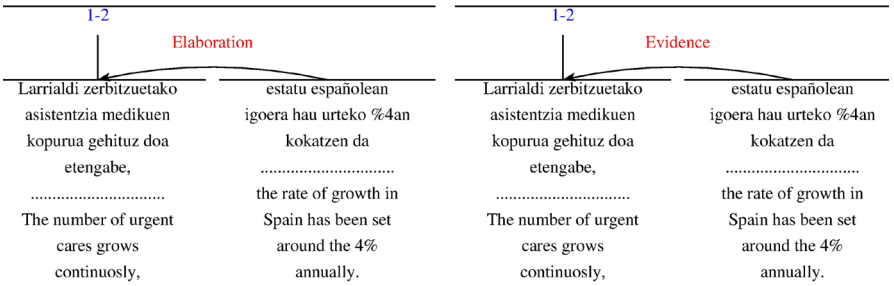

Fig. 2 Quantitative evaluation: factor conflation (Iruskieta et al. 2013a, GMB0401)

The second table of Fig. 3 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 at the beginning 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.

Figure 4 shows a real example extracted from Iruskieta et al. (2013a).

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 italicize.

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

Fig. 3 Quantitative evaluation of Fig. 2 with RSTeval

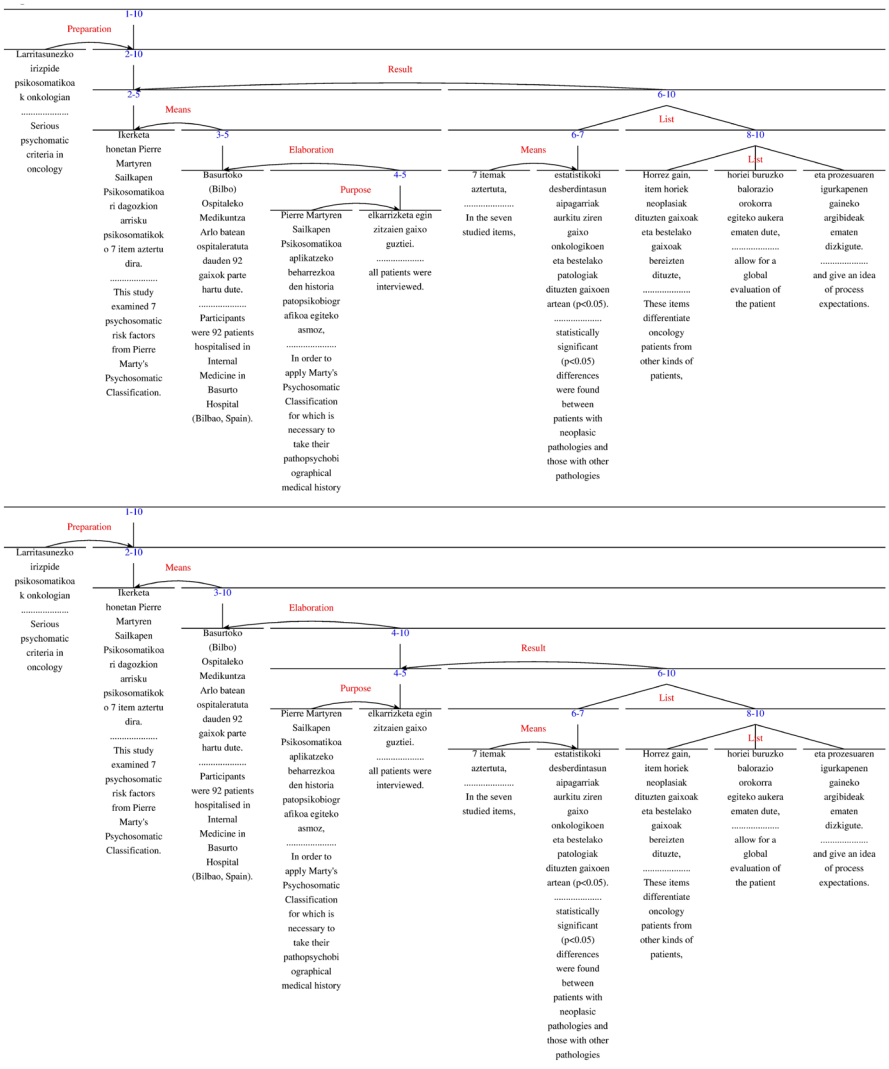

Fig. 4 Annotations of text GMB0701 (Iruskieta et al. 2013a)

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).[9]

[9] Note that, after harmonizing discourse segmentation, accuracy, precision, recall and F-measure obtain the same value. Therefore, although this results in a somewhat artificial level of agreement, we are conscious about this fact, we use the standard measure employed in the RST literature (Marcu 2000a; Maziero and Pardo 2009).

Table 3 Qualitative method for text GMB0701

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 4 Quantitative method: agreement level 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 5 Agreement level according to rhetorical relations in GMB0701

Relations		
Match	R	P
7 of 9	0.778	0.778

2. 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), the method proposed by Marcu (2000a) is not able to compare the relations where constituents has changed. Observe the following issues in Fig. 4:

- 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'.
- 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.

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 in Marcu's (2000b) method; only the central subconstituent (CS) has to be the same.[10] 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 not 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 not 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 formalizes the method to check the agreement types in rhetorical relations (see Fig. 5). As we mentioned before, to check agreement in rhetorical relation, the constituent of this relation must have the same central subconstituent (CS). If this condition is fulfilled, we check if 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 and 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)**.
2. Different choice in nuclearity entailed discrepancy in N/S relations **(N/S)**.

[10] If there is more than one CS (because there is a multinuclear relation) at least one of them has to be the same for N/S-N/N mix-up.

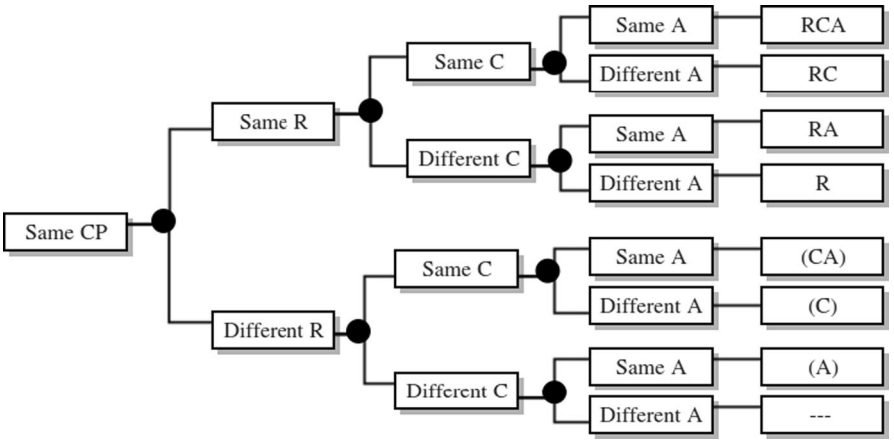

Fig. 5 Decision tree based on CS to establish the agreement types about R

3. A relation has the same constituent and attachment point, but not the same relation label (≠ **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)**.

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 signaled 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 which was segmented and harmonized in Spanish (A2) and in English (A1) (Fig. 7) to illustrate the qualitative method (Table 7).[11]

[11] Basque segments (A3) were also harmonized, but space constraints preclude us to align with Spanish and English. Anyway, the harmonization of TERM38_SPA segmentation in the three languages can be consulted at: http://ixa2.si.ehu.es/rst/segmentuak_multiling.php?bilatzekoa=TERM38%. The English RS-tree can be consulted at: http://ixa2.si.ehu.es/rst/diskurtsoa_jpg/TERM38_A1.jpg. The Spanish RS-tree can be consulted at: http://ixa2.si.ehu.es/rst/diskurtsoa_jpg/TERM38_A2.jpg.

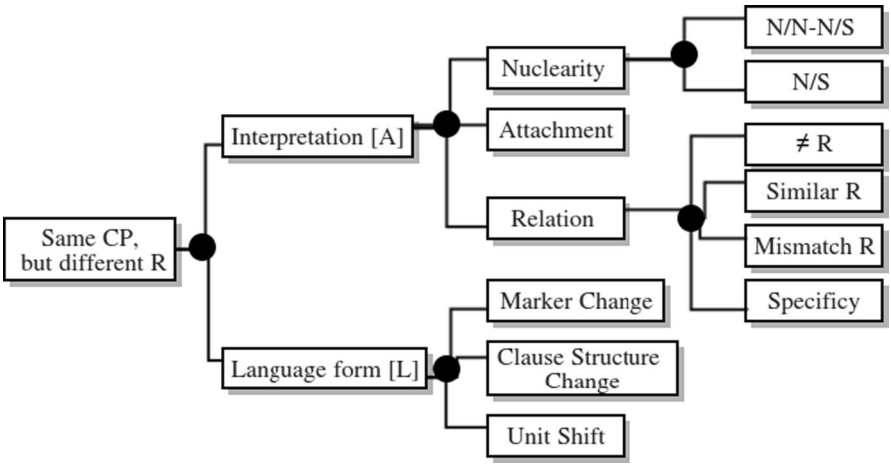

Fig. 6 Decision tree to establish the sources of agreement and disagreement about R

Table 7 includes the analyzed factors for Fig. 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 CSs is not trivial, because multinuclear relations have more than one CS. 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 annotate as "no-match"[12] and (b) compare first non-ambiguous CSs and leave problematic comparisons (lines 22 and 23) for the end. Following the last choice there is not any ambiguous CS in Table 7, because the other CS candidate (CS 12 in line 10) was used in other structure. Because of that, when we have to compare relations with more than one CS with another that has only one CS, at least one of the CSs 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 Marcu (2000b), since the most important unit can be formed by more than one EDU.[13]

In Table 8 we present the results of our evaluation method for the example in Fig. 7.

[12] If we follow this decision, we could not compare structures that contain a N/N–N/S mix-up inside the relation.

[13] 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 not 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).

Table 6 TERM38_SPA segmented and harmonized in Spanish and English

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

Table 6 continued

Tables		Languages	
7	9	Spanish	English
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?

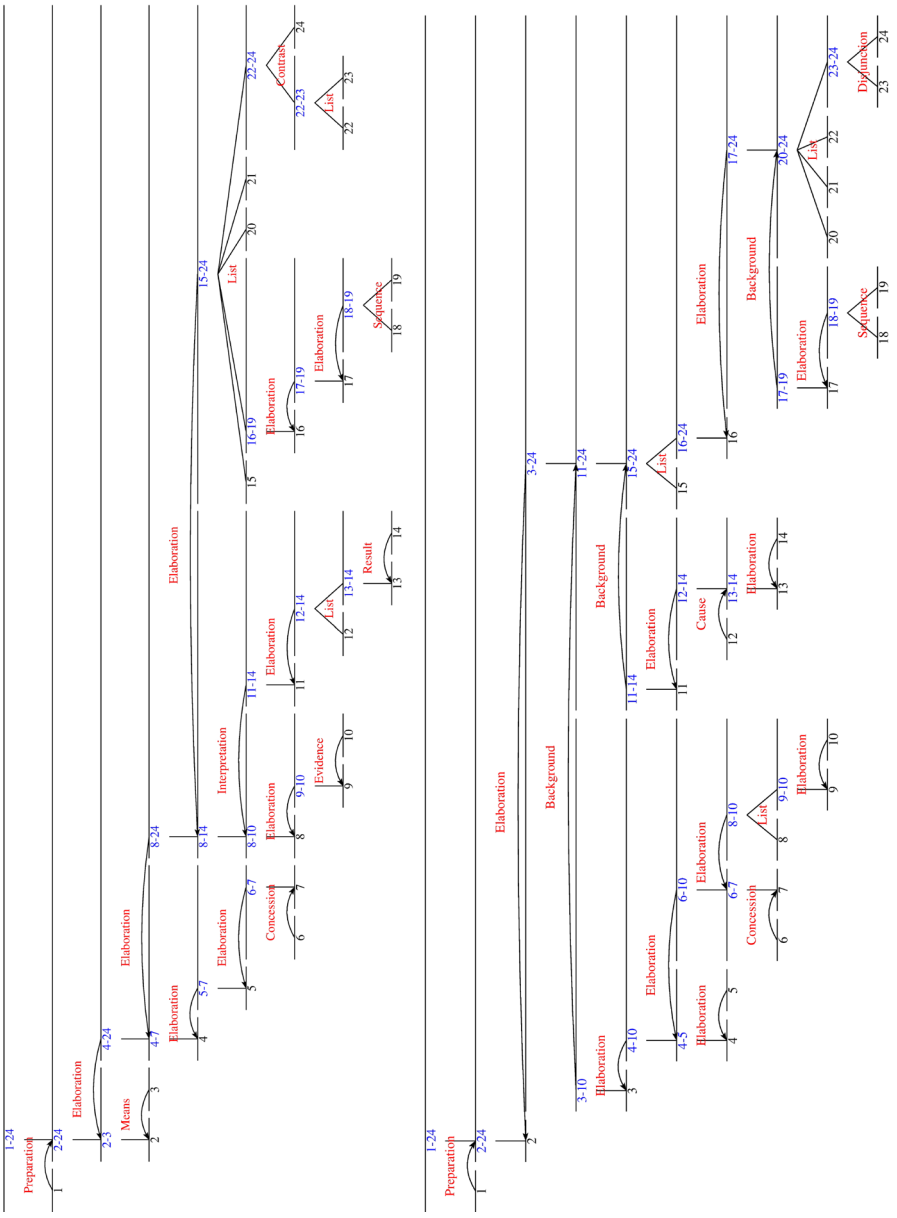

Fig. 7 Rhetorical tree elaborated by A2 (Spanish) and A1 (English), TERM38_SPA

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.

Table 7 Qualitative evaluation matrix TERM38_SPA

L	ENG				SPA				Qualitative evaluation					
	CS(s)	R	C	A	CS(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)	$N/NversusN/S[A]$
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)	$N/NversusN/S[A]$
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 8 Qualitative evaluation results for the example in Fig. 7, TERM38_SPA

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

Both methods measure the similar factors: (1) EDUs and spans (constituent and attachment), (2) 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.

In Table 11 both methods describe total agreement in segmentation. This is 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, 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 italicize the most important differences.

As we can see in Table 12, an important part of the agreement in quantitative evaluation method is captured 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 relations. The quantitative method captures an agreement of 4.35 %, while 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 of them 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 of advantages are shown in contingency Table 7:

Table 9 Contingency table for text TERM38_SPA with quantitative method, using *RSTeval*

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	

Table 9 continued

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
!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 10 Quantitative method results for text TERM38_SPA

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 11 Comparison using both methods, TERM38_SPA

	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 12 Comparison of agreement using both methods for text TERM38

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

1. 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 a higher level in the tree structure) and not in relation as quantitative method does. Moreover, the qualitative method accounts for the source of disagreement (Disagree).

2. Only rhetorical relations are compared. The description allows for a full coincidence in structure (RCA), or a partial match (RA, RC or R).
3. Reasons for annotator disagreement are captured: *a*) because of differences in the linguistic expression [L] or *b*) because of interpretation [A].
4. 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 CS (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 in this case in CS (5), relation (ELABORATION) and attachment point (4N). By contrast, slight disagreement in constituent (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 CS (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, 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 7, one of our segmentation principles is that every EDU should contain a

Table 13 Segmentation agreement

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

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.[14] 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.[15]

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.

[14] This harmonization work can be found at http://ixa2.si.ehu.es/rst/segmentuak_multiling.php.

[15] For Kappa segment candidates were calculated automatically by counting verbs.

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.[16]

(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.]

(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.]

[16] In the example, the original segmentation is marked with square brackets and the segmentation after harmonization with curly brackets.

(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.

We can see from the table that the language with more changes is Basque.[17] 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 suggest that 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 using non-finite clauses or prepositional complements.

3.2 Rhetorical analysis results

Results of quantitative method were presented in order to show the consistency of this 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.

3.2.1 Results of the quantitative evaluation method

Results of the quantitative evaluation are shown in Table 15.[18]

[17] 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.

[18] EDUs are excluded because they are identical after harmonization.

Table 14 Segmentation changes

	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 15 Quantitative evaluation results (F-measure)

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

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 Sect. 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.

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 have interpreted from the results of Table 14, because English and Spanish are typologically closer to each other than Basque is to either English or Spanish (Cenoz

Table 16 Qualitative evaluation results (F-measure): analysis of the sources of agreement

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 17 Qualitative evaluation results (F-measure): analysis of the sources of disagreement

Classification		ENG-SPA (%)	ENG-BSQ (%)	SPA-BSQ (%)
Annotator-based discrepancies	Nuclearity	4.00	4.00	3.33
	N/N versus 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 18 Qualitative evaluation results (F-measure): summary of results

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

2003). 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.

After aligning the contingency tables of the qualitative evaluation from all the RS-structure in English, Spanish and Basque, we measured the agreement of rhetorical relations with Fleiss Kappa (see Table 19) for assessing the reliability of agreement between more than two annotators. 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 between the three annotators.

As we observe in Table 19, Fleiss' Kappa measures show different degrees of understanding rhetorical relations.

1. Almost perfect: PREPARATION.
2. Substantial: SUMMARY and CONCESSION.

Table 19 Qualitative evaluation results (Fleiss' Kappa) for rhetorical relations

Relation	Kappa	z	*p* value
Preparation	0.851	25.528	0.000
Summary	0.712	21.361	0.000
Concession	0.705	21.155	0.000
List	0.554	16.629	0.000
Elaboration	0.531	15.933	0.000
Condition	0.525	15.763	0.000
Sequence	0.499	14.966	0.000
Restatement	0.424	12.723	0.000
Background	0.420	12.589	0.000
Circumstance	0.420	12.586	0.000
Contrast	0.376	11.272	0.000
Cause	0.352	10.552	0.000
Purpose	0.335	10.057	0.000
Result	0.301	9.017	0.000
Means	0.221	6.617	0.000
Conjunction	0.172	5.151	0.000
Motivation	0.136	4.084	0.000
Interpretation	0.080	2.390	0.017
Solutionhood	−0.011	−0.337	0.736
Justify	−0.009	−0.269	0.788
Antithesis	−0.008	−0.235	0.814
Evidence	−0.008	−0.235	0.814
Evaluation	−0.003	−0.100	0.920
Disjunction	−0.001	−0.033	0.973
Unless	−0.001	−0.033	0.973

3. Moderate agreement: LIST, ELABORATION, CONDITION, SEQUENCE, RESTATEMENT, BACKGROUND and CIRCUMSTANCE.
4. Fair agreement: CONTRAST, CAUSE, PURPOSE, RESULT and MEANS.
5. Slight agreement: CONJUNCTION, MOTIVATION and INTERPRETATION.
6. No observed agreement for: ANTITHESIS, DISJUNCTION, EVALUATION, EVIDENCE, JUSTIFY, SOLUTIONHOOD and UNLESS.[19]

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),[20] 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 relation, a CAUSE relation is in the A1 analysis.[21]

(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

[19] "Values of agreement between −A_e/1−A_e (no observed agreement) and 1 (observed agreement = 1), with the value 0 signifying chance agreement (observed agreement = expected agreement)." (Artstein and Poesio 2008, p. 559).

[20] Catford (1965, pg. 73) defines translation shifts as "departures from formal correspondence in the process of going from the SL to the TL" (from the Source Language to the Target Language). Chesterman (1997) states that changes from original to translated text are due to a translation strategy.

[21] 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.

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 relations and reveal local translation strategies:

1. Relation signaling has a different configuration (Marker Change). Within Marker Change, we found three subtypes:
   (a) inclusion of a marker,
   (b) exclusion of a marker, and
   (c) changing a marker.
2. Differences because of the use of a distinct language configuration (Clause Structure Change):
   (a) hierarchical downgrading, and
   (b) hierarchical upgrading.
3. Punctuation is used differently (Unit Shift):
   (a) an independent sentence is integrated in another sentence, and
   (b) 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[22] analyzed here is a product of the harmonization and it was annotated by A2 as an ELABORATION relation.

(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 $(\ldots)]_{7-11S-ELABORATION}$[23]

(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

[22] Again, this goes against the principles of our segmentation.

[23] Note here the human annotation error which does not follow the modular and incremental annotation that Pardo (2005) proposes.

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.

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 Basque: 34.78 %). Another interesting aspect is that the Marker Change translation strategy is the most prominent one (MC: 34.78 % versus CSC: 15.94 % and US: 17.39 %), and changes in discourse markers have an influence on rhetorical annotation.[24] These results are merely describing tendencies, because the corpus is not big enough (although is comparable to other corpora in the literature, such as Scott et al. (1998)). The results are sensitive to segmentation granularity or harmonization decisions and to text characteristics (genre and domain). However what is relevant 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. In this table, we present the final results from both evaluation methods, providing the F-measure of relation factor.

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 believe they ought not to be included in the same factor. Secondly, the qualitative evaluation clearly distinguishes the most relevant sources of disagreement; because

[24] 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 %).

Table 20 Translation strategies and different language pairs

	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 21 Comparison of relation factor in quantitative and qualitative evaluation methods (F-measure)

	Quantitative evaluation (%)	Qualitative evaluation (%)
ENG-SPA	56.22	59.67
ENG-BSQ	53.28	54.66
SPA-BSQ	54.94	56.99

of that, results are more reliable. The translation of discourse structure from one language to another does not result in a one-to-one mapping of relations. 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[25] the tagged multilingual corpus can be consulted, as for example: (1) the rhetorical structure of a text (in RS3 format) and its image (in JPG format); (2) all instances of a selected rhetorical relation in three languages; (3) 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 for comparison of rhetorical structures, 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

[25] http://ixa2.si.ehu.es/rst.

translation differences that the analysis evidenced, showing that typological differences between the languages affected mostly 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 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 not 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[26] (Taboada and Renkema 2008), Spanish RST TreeBank[27] (da Cunha et al. 2011b) and Basque RST TreeBank[28] (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

[26] SFU corpus is available at http://www.sfu.ca/~mtaboada/download/downloadRST.html.

[27] RST Spanish TreeBank is available at http://corpus.iingen.unam.mx/rst/corpus_en.html.

[28] Basque RST TreeBank is available at http://ixa2.si.ehu.es/diskurtsoa/en/.

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.

Acknowledgments This work has been partially financed by the Spanish projects RICOTERM 4 (FFI2010-21365-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 the anonymous reviewers for their comments and suggestions, Nynke van der Vliet for her feedback on the evaluation method, Esther Miranda for designing the website, and Oier Lopez de Lacalle for helping with the scripts to calculate the statistics.

Appendix: 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, p. 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 estáen 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 nominalized 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 Sect. 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, asímismo, 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,][29] [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 5 section), 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

[29] Truncated EDU. English translation: 'if there can be said to be any' (see Sect. 5).

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*).

(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 Fig. 8, which follows the Basque word order.

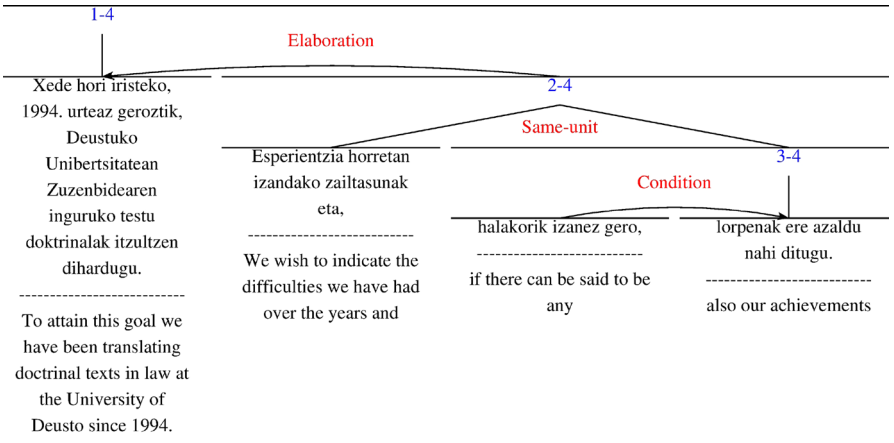

Fig. 8 Example of a Same-unit (non) relation

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 Sect. 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 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 Sect. 3.1.

References

Abelen, E., Redeker, G., & Thompson, S. A. (1993). The rhetorical structure of US-American and Dutch fund-raising letters. *Text*, *13*(3), 323–350.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, *9*(2), 167–193.

Bateman, J. A., & Rondhuis, K. J. (1997). Coherence relations: Towards a general specification. *Discourse Processes*, *24*(1), 3–49.

Carlson, L., Okurowski, M. E., & Marcu, D. (2002). *RST Discourse Treebank, LDC2002T07 [Corpus]*. Philadelphia, PA: Linguistic Data Consortium.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In van Kuppevelt, C. J. Jan & R. W. Smith (Eds.), *Current and new directions in discourse and dialogue* (pp. 85–112). Berlin: Springer.

Catford, J. C. (1965). *A linguistic theory of translation: An essay in applied linguistics* (Vol. 8). New York: Oxford University Press.

Cenoz, J. (2003). The role of typology in the organization of the multilingual lexicon. In J. Cenoz, B. Hufeisen & U. Jessner (Eds.), *The multilingual lexicon* (pp. 103–116), New York: Springer.

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* (Vol. 22). Amsterdam and Philadelphia: Benjamins.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, *26*(1), 1–11.

Cristea, D., Ide, N., & Romary, L. (1998). Veins theory: A model of global discourse cohesion and coherence. In C. Boitet & P. Whitelock (Eds.), *17th international conference on Computational linguistics* (Vol. 1 pp. 281–285). Montreal, Canada: Association for Computational Linguistics.

Cui, S. (1986). A comparison of English and Chinese expository rhetorical structures. Ph.D. 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, J. M., & Sierra, G. (2011a). On the Development of the RST Spanish Treebank. In *5th Linguistic annotation workshop. 49th annual meeting of the association for computational linguistics, ACL* (pp. 1–10). Portland, Oregon, USA.

da Cunha, I., Torres-Moreno, J. M., Sierra, G., Cabrera-Diego, L. A., Castro-Rolón, B. G., & Rolland-Bartilotti, J. M. (2011b). The RST Spanish Treebank On-line Interface. In *International conference recent advances in NLP* (pp. 698–703), Bulgaria.

Delin, J., Hartley, A. F., Paris, C., Scott, D. R., & Linden, K. V. (1994). Expressing procedural relationships in multilingual instructions. In *Seventh International Workshop on Natural Language Generation* (pp. 61–70), Association for Computational Linguistics.

Delin, J., Hartley, A. F., & Scott, D. R. (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)* (pp. 1619–1623), Valletta, Malta.

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, J. L. (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*, (Vol. 78, pp. 173–183). Boston, MA.

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* (pp. 17–29). Rodopi, Corpus-based approaches to contrastive linguistics and translation studies. Amsterdam/New York.

House, J. (2004). *Explicitness in discourse across languages. Neue Perspektiven in der Übersetzungs-und Dolmetschwissenschaft* (pp. 185–208), Bochum: AKS.

Iruskieta, M., Aranzabe, M. J., 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*, 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, K. C. C. (1998). Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text & Talk*, *18*(1), 103–141.
Mann, W. C., & Taboada, M. (2010). RST web-site. http://www.sfu.ca/rst/. Accessed 30 September 2012.
Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281.
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*. Cambridge: MIT press.
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* (pp. 9–17), Seattle (USA): Morgan Kaufmann Publishers.
Maxwell, M. (2010). Limitations of corpora. *International Journal of Corpus Linguistics*, *15*(3), 379–383.
Maziero, E. G., & Pardo, T. A. S. (2009). Automatização de um método de avaliação de estruturas retóricas. In: *RST Brazilian meeting*, São Paulo, Brazil.
Mitocariu, E., Anechitei, D. A., & Cristea, D. (2013). *Comparing discourse tree structures* (pp. 513–522). Berlin: Springer. Computational Linguistics and Intelligent Text Processing.
Mohamed, A. H., & Omer, M. R. (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* (Vol. 45, pp. 664–671). Prague.
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* (Vol. 14, pp. 253–256). Mitzpe Ramon: ACL.
Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. Ph.D. thesis, Instituto de Ciências Matemáticas e de Computação, São Carlos-SP: Universidade de São Paulo.
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, S. L. (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, D. R., Delin, J., & Hartley, A. F. (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* (Vol. 1, pp. 149–156). Association for Computational Linguistics.
Stede, M. (2008a). Disambiguating rhetorical structure. *Research on Language and Computation*, *6*(3), 311–332.
Stede, M. (2008b). *RST revisited: Disentangling nuclearity* (pp. 33–57). Amsterdam and Philadelphia: John Benjamins. 'Subordination' versus 'coordination' in sentence and text.
Taboada, M. (2004a). *Building coherence and cohesion: Task-oriented dialogue in English and Spanish*. Amsterdam and Philadelphia: John Benjamins.
Taboada, M. (2004b). *Rhetorical relations in dialogue: A contrastive study* (pp. 75–97), Amsterdam and Philadelphia: John Benjamins. Discourse across Languages and Cultures.

Taboada, M., & Mann, W. C. (2006a). Applications of rhetorical structure theory. *Discourse Studies*, *8*(4), 567–588.

Taboada, M., & Mann, W. C. (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. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html. Accessed 30 September 2012

Trask, R. L. (1997). *The history of Basque*. London: Routledge.

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 and HAEE-IVAP. (1997). *International congress on terminology*. Donostia and Gasteiz: UZEI; HAEE-IVAP.

van der Vliet, N. (2010). Inter annotator agreement in discourse analysis. 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* (pp. 206–213). Citeseer, Columbia.

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.