On the Development of the RST Spanish Treebank

Iria da Cunha
Institute for Applied Linguistics (UPF), Spain
Instituto de Ingeniería (UNAM), Mexico
Laboratoire Informatique d'Avignon (UAPV), France
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon (UAPV), France
Instituto de Ingeniería (UNAM), Mexico
École Polytechnique de Montréal, Canada
juan-manuel.torres@univ-avignon.fr

Gerardo Sierra
Instituto de Ingeniería (UNAM), Mexico
gsierram@iingen.unam.mx

Abstract

In this article we present the RST Spanish Treebank, the first corpus annotated with rhetorical relations for this language. We describe the characteristics of the corpus, the annotation criteria, the annotation procedure, the inter-annotator agreement, and other related aspects. Moreover, we show the interface that we have developed to carry out searches over the corpus' annotated texts.

1 Introduction

The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a language independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs) linked by means of nucleus-satellite or multinuclear rhetorical relations. In the first case, the satellite gives additional information about the other one, the nucleus, on which it depends (ex. Result, Condition, Elaboration or Concession). In the second case, several elements, all nuclei, are connected at the same level, that is, there are no elements dependent on others and they all have the same importance with regard to the intentions of the author of the text (ex. Contrast, List, Joint or Sequence). The rhetorical analysis of a text by means of RST includes 3 phases: segmentation, detection of relations and building of hierarchical rhetorical trees. For more information about RST we recommend the original article of Mann and Thompson (1988), the web site of RST[1] and the RST review by Taboada and Mann (2006a).

RST has been used to develop several applications, like automatic summarization, information extraction (IE), text generation, question-answering, automatic translation, etc. (Taboada and Mann, 2006b). Nevertheless, most of these works have been developed for English, German or Portuguese. This is due to the fact that at present corpora annotated with RST relations are available only for these languages (for English: Carlson et al., 2002, Taboada and Renkema, 2008; for German: Stede, 2004; for Portuguese: Pardo et al., 2008) and there are automatic RST parsers for two of them (for English: Marcu, 2000; for Portuguese: Pardo et al., 2008) or automatic RST segmenters (for English: Tofiloski et al., 2009). Scientific community working on RST applied to Spanish is very small. For example, Bouayad-Agha et al. (2006) apply RST to text generation in several languages, Spanish among them. Da Cunha et al. (2007) develop a summarization system for medical texts in Spanish based on RST. Da Cunha and Iruskieta (2010) perform a contrastive analysis of Spanish and Basque texts. Romera (2004) analyzes coherence relations by means of RST in spoken Spanish. Taboada (2004) applies RST to analyze the resources used by speakers to elaborate conversations in English and Spanish.

We consider that it is necessary to build a Spanish corpus annotated by means of RST. This corpus should be useful for the development of a rhetorical parser for this language and several other applications related to computational linguistics, like those developed for other languages

[1] http://www.sfu.ca/rst/index.html

Proceedings of the Fifth Law Workshop (LAW V), pages 1–10,
Portland, Oregon, 23-24 June 2011.

(automatic translation, automatic summarization, IE, etc.). And that is what we pretend to achieve with our work. We present the development of the RST Spanish Treebank, the first Spanish corpus annotated by means of RST.

In Section 2, we present the state of the art about RST annotated corpora. In Section 3, we explain the characteristics of the RST Spanish Treebank. In Section 4, we show the search interface we have developed. In Section 5, we establish some conclusions and future work.

2 State of the Art

The most known RST corpus is the RST Discourse Treebank, for English (Carlson et al., 2002a, 2002b). It includes 385 texts of the journalistic domain, extracted from the Penn Treebank (Marcus et al., 1993), such as cultural reviews, editorials, economy articles, etc. 347 texts are used as a learning corpus and 38 texts are used as a test corpus. It contains 176,389 words and 21,789 EDUs. 13.8% of the texts (that is, 53) were annotated by two people with a list of 78 relations. For annotation, the annotation tool RSTtool[2] (O'Donnell, 2000) was used, with some adaptations. The principal advantages of this corpus stand on the high number of annotated texts (for the moment it is the biggest RST corpus) and the clarity of the annotation method (specified in the annotation manual by Carlson and Marcu, 2001). However, some drawbacks remain. The corpus is not free, it is not on-line and it only includes texts of one domain (journalistic).

For English there is also the Discourse Relations Reference Corpus (Taboada and Renkema, 2008). This corpus includes 65 texts (each one tagged by one annotator) of several types and from several sources: 21 articles from the Wall Street Journal extracted from the RST Discourse Treebank, 30 movies and books' reviews extracted from the epinions.com website, and 14 diverse texts, including letters, webs, magazine articles, newspaper editorials, etc. The tool used for annotation was also the RSTtool. The advantages of this corpus are that it is free and on-line, and it includes texts of several types and domains. The disadvantages are that the amount of texts is not very high, the annotation methodology is not specified and it does not include texts annotated by several people.

Another well-known corpus is the Potsdam Commentary Corpus, for German (Stede, 2004; Reitter and Stede, 2003). This corpus includes 173 texts on politics from the on-line newspaper Märkische Allgemeine Zeitung. It contains 32,962 words and 2,195 sentences. It is annotated with several data: morphology, syntax, rhetorical structure, connectors, correference and informative structure. Nevertheless, only a part of this corpus (10 texts), which the authors name "core corpus", is annotated with all this information. The texts were annotated with the RSTtool. This corpus has several advantages: it is annotated at different levels (the annotation of connectors is especially interesting); all the texts were annotated by two people (with a previous RST training phase); it is free for research purposes, and there is a tool for searching over the corpus (although it is not available on-line). The disadvantages are: the genre and domain of all the texts are the same, the methodology of annotation was quite intuitive (without a manual or specific criteria) and the inter-annotator agreement is not given.

For Portuguese, there are 2 corpora, built in order to develop a rhetorical parser (Pardo et al., 2008). The first one, the CorpusTCC (Pardo et al., 2008), was used as learning corpus for detection of linguistic patterns indicating rhetorical relations. It contains 100 introduction sections of computer science theses (53,000 words and 1,350 sentences). To annotate the corpus a list of 32 rhetorical relations was used. The annotation manual by Carlson and Marcu (2001) was adapted to Portuguese. The annotation tool was the ISI RST Annotation Tool[3], an extension of the RSTtool. The advantages of this corpus are: it is free, it contains an acceptable number of texts and words and it follows a specific annotation methodology. The disadvantage is: it only includes texts of one genre and domain, only annotated by one person.

The second one, Rhetalho (Pardo and Seno, 2005), was used as reference corpus for the parser evaluation. It contains 50 texts: 20 introduction sections and 10 conclusion sections from computer science scientific articles, and 20 texts from the on-line newspaper Folha de São Paulo (7 from the Daily section, 7 from the World section and 6 from

[2] http://www.wagsoft.com/RSTTool/

[3] http://www.isi.edu/~marcu/discourse/

the Science section). It includes approximately 5,000 words. The relations and the annotation tool are the same as those used in the CorpusTCC. The advantages of this corpus are that it is free, it was annotated by 2 people (they both were RST experts and followed an annotation manual) and it contains texts of several genres and domains. The main disadvantage is the scarce amount of texts.

The Penn Discourse Treebank (Rashmi et al., 2008)f for English includes texts annotated with information related to discourse structure and semantics (without a specific theoretical approach). Its advantages are: its big size (it contains 40,600 annotated discourse relations) allows to apply machine learning, and the discourse annotations are aligned with the syntactic constituency annotations of the Penn Treebank. Its limitations are: dependencies across relations are not marked, it only includes texts of the journalistic domain, and it is not free. Although there are several corpora annotated with discourse relations, there is not a corpus of this type for Spanish.

3 The RST Spanish Treebank

As Sierra (2008) states, a corpus consists of a compilation of a set of written and/or spoken texts sharing some characteristics, created for certain investigation purposes. According to Hovy (2010), we use 7 core questions in corpus design, detailed in the next subsections.

3.1 Selecting a Corpus

For the RST Spanish Treebank, we wanted to include short texts (finally, the average is 197 words by text; the longest containing 1,051 words and the shortest, 25) in order to get a best on-line visualization of the RST trees. Moreover, in the first stage of the project, we preferred to select specialized texts of very different areas, although in the future we plan to include also non-specialized texts (ex. blogs, news, websites) in order to guarantee the representativity of the corpus. We did not find a pre-existing Spanish corpus with these characteristics, so we decided to build our own corpus. Following Cabré (1999), we consider that a text is specialized if it is written by a professional in a given domain. According to this work, specialized texts can be divided in three levels: high (both the author and the potential reader of the text are specialists), average (the author of the text is a specialist, and the potential reader of that text is a student or someone interested in or possessing some prior knowledge about the subject) and low (the author of the text is a specialist, and the potential reader is the general public). The RST Spanish Treebank includes specialized texts of the three mentioned levels: high (scientific articles, conference proceedings, doctoral theses, etc.), average (textbooks) and low (articles and reports from popular magazines, associations' websites, etc.). The texts have been divided in 9 domains (some of them including subdivisions): Astrophysics, Earthquake Engineering, Economy, Law, Linguistics (Applied Linguistics, Language Acquisition, PLN, Terminology), Mathematics (Primary Education, Secondary Education, Scientific Articles), Medicine (Administration of Health Services, Oncology, Orthopedy), Psychology and Sexuality (Clinical Perspective, Psychological Perspective).

The size of a corpus is also a polemic question. If the corpus is developed for machine learning, its size will be enough when the application we want to develop obtains acceptable percentages of precision and recall (in the context of that application). Nevertheless, if the corpus is built with descriptive purposes, it is difficult to determine the corpus size. In the case of a corpus annotated with rhetorical relations, it is even more difficult, because there are various factors involved: EDUs, SPANs (that is, a group of related EDUs), nuclearity and relations. In addition, relations are multiple (we use 28). As Hovy (2010: 13) mentions, one of the most difficult phenomena to annotate is the discourse structure. Our corpus contains 52,746 words and 267 texts. Table 1 includes RST Spanish Treebank statistics in terms of texts, words, sentences and EDUs.

	Texts	Words	Sentences	EDUs
Learning corpus	183	41,555	1,759	2,655
Test corpus	84	11,191	497	694
Total corpus	267	52,746	2,256	3,349

Table 1: RST Spanish Treebank statistics

To increase the linear performance of a statistical method, it is necessary that the training corpus size grows exponentially (Zhao et al., 2010). However, the RST Spanish Treebank is not designed only to use statistical methods; we think it will be useful to employ symbolic or hybrid algorithms (combining symbolic and statistical methods). Moreover, this corpus will be dynamic, so we expect to have a bigger corpus in the future, useful to apply machine learning methods.

If we measure the corpus size in terms of words or texts, we can take as a reference the other RST corpora. Nevertheless, as Sierra states (2008), it is "absurd" to try to build an exhaustive corpus covering all the aspects of a language. On the contrary, the linguist looks for the representativeness of the texts, that is, tries to create a sample of the studied language, selecting examples which represent the linguistic reality, in order to analyze them in a pertinent way. In this sense and in the frame of this work, we consider that the size will be adequate if the rhetorical trees of the corpus include a representative number of examples of rhetorical relations, at least 20 examples of each one (taking into account that the corpus contains 3115 relations, we consider that this quantity is acceptable; however, we expect to have even more examples when the corpus grows). Table 2 shows the number of examples of each relation currently included into the RST Spanish Treebank (N-S: nucleus-satellite relation; N-N: multinuclear relation). As it can be observed, it contains more than 20 examples of most of the relations. The exceptions are the nucleus-satellite relations of Enablement, Evaluation, Summary, Otherwise and Unless, and the multinuclear relations of Conjunction and Disjunction, because it is not so usual to find these rhetorical relations in the language, in comparison with others. Hovy (2010: 128) states that, given the lack of examples in the corpus, there are 2 possible strategies: a) to leave the corpus as it is, with few or no examples of some cases (but the problem will be the lack of training examples for machine learning systems), or b) to add low-frequency examples artificially to "enrich" the corpus (but the problem will be the distortion of the native frequency distribution and perhaps the confusion of machine learning systems). In the current state of our project, we have chosen the first option. We think that, including specialized texts in a second stage, we will get more examples of these less common relations. If we carry out a more granulated segmentation maybe we could obtain more examples; however, we wanted to employ the segmentation criteria used to develop the Spanish RST discourse segmenter (da Cunha et al., 2011).

Relation	**Type**	**Quantity**	
		Nº	**%**
Elaboration	N-S	765	24.56
Preparation	N-S	475	15.25
Background	N-S	204	6.55
Result	N-S	193	6.20
Means	N-S	175	5.62
List	N-N	172	5.52
Joint	N-N	160	5.14
Circumstance	N-S	140	4.49
Purpose	N-S	122	3.92
Interpretation	N-S	88	2.83
Antithesis	N-S	80	2.57
Cause	N-S	77	2.47
Sequency	N-N	74	2.38
Evidence	N-S	59	1.89
Contrast	N-N	58	1.86
Condition	N-S	53	1.70
Concession	N-S	50	1.61
Justification	N-S	39	1.25
Solution	N-S	32	1.03
Motivation	N-S	28	0.90
Reformulation	N-S	22	0.71
Otherwise	N-S	3	0.10
Conjunction	N-N	11	0.35
Evaluation	N-S	11	0.35
Disjunction	N-N	9	0.29
Summary	N-S	8	0.26
Enablement	N-S	5	0.16
Unless	N-S	2	0.06

Table 2: Rhetorical relations in RST Spanish Treebank

3.2 Instantiating the Theory

Our segmentation and annotation criteria are very similar to the original ones used by Mann and Thompson (1988) for English, and by da Cunha and Iruskieta (2010) for Spanish. We also explore the annotation manual for English by Carlon and Marcu (2001). Though we use some of their postulates, we think that their analysis is too meticulous in some aspects. Because of this, we consider that it is not adjusted to our interest, which is the finding of the simplest and most objective annotation method, orientated to the future development of a rhetorical parser for Spanish. To sum up, our segmentation criteria are:

a) All the sentences of the text are segmented as EDUs (we consider that a sentence is a textual passage between a period and another period, a semicolon, a question mark or an exclamation point; texts' titles are also segmented). Exs.[4]

[Éstas son las razones fundamentales que motivaron este trabajo.]

[These are the fundamental reasons which motivated this work.]

[Estudio de caso único sobre violencia conyugal]

[Study of a case on conjugal violence]

b) Intra-sentence EDUs are segmented, using the following criteria:

b1) An intra-sentence EDU has to include a finite verb, an infinitive or a gerund. Ex.

[Siendo una variante de la eliminación Gaussiana,] [posee características didácticas ventajosas.]

[Being a variant of Gaussian elimination,] [it possesses didactic profitable characteristics.]

b2) Subject/object subordinate clauses or substantive sentences are not segmented. Ex.

[Se muestra que el modelo discreto en diferencias finitas es convergente y que su realización se reduce a resolver una sucesión de sistemas lineales tridiagonales.]

[It appears that the discreet model in finite differences is convergent and that its accomplishment is to solve a succession of tridiagonal linear systems.]

b3) Subordinate relative clauses are not segmented. Ex.

[Durante el proceso, que utiliza solo aritmética entera, se obtiene el determinante de la matriz de coeficientes del sistema, sin necesidad de cálculos adicionales.]

[During the process, which only uses entire arithmetic, the determinant of the system coefficient matrix is obtained, without additional calculations.]

b4) Elements in parentheses are only segmented if they follow the criterion b1. Ex.

[Este año se cumple el bicentenario del nacimiento de Niels (Nicolás, en nuestro idioma) Henrik Abel.]

[This year is the bicentenary of Niels's birth (Nicolás, in our language) Henrik Abel.]

b5) Embedded units are segmented by means of the non-relation Same-Unit proposed by Carlon and Marcu (2001). Figure 1 shows this structure.

[En décadas precedentes se ha puesto de manifiesto,] [y así lo han atestiguado muchos investigadores de la terminología científica serbia,] [una tendencia a importar préstamos del inglés.]

[In previous decades it has been shown,] [and it has been testified by many researchers of the scientific Serbian terminology,] [a trend to import loanwords from English.]

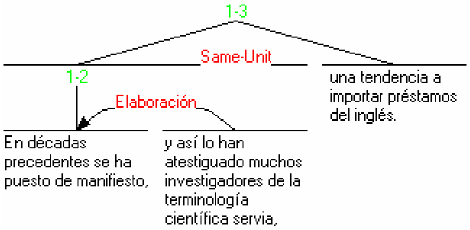
1-3
Same-Unit
1-2
Elaboración
En décadas precedentes se ha puesto de manifiesto,
y así lo han atestiguado muchos investigadores de la terminología científica servia,
una tendencia a importar préstamos del inglés.

Figure 1: Example of the non-relation Same-Unit

3.3 Designing the Interface

The annotation tool used in this work is the RSTtool, since it is free and easy to use. Therefore, we preferred to use it instead of designing a new one. Nevertheless, we have designed an on-line interface to include the corpus and to carry out searches over it (see Section 4).

3.4 Selecting and Training the Annotators

With regard to the corpus annotators, we have a team of 10 people (last year Bachelor's degree students, Master's degree students and PhDs)[5]. Before the annotation, they took a RST course of 6 months (100 hours), where the segmentation and annotation methodology used for the development of the RST Spanish Treebank was explained.[6] We called this period "training phase". The course had a theoretical and a practical part. In the theoretical part, some criteria with regard to the 3 phases of rhetorical analysis (segmentation, detection of relations, and rhetorical trees building) were given to annotators. In the practical part, firstly, it was explained how to use the RSTtool. Secondly, annotators extracted several texts from the web, following their personal interests, as for example, music, video games, cookery or art webs. They segmented those texts, using the established segmentation criteria. Once segmented, all the doubts and problematic examples were discussed, and they tried to get an agreement on the most complicated cases. Thirdly, the relations were

[4] Spanish examples were extracted from the corpus. English translations are ours.

[5] We thank annotators (Adriana Valerio, Brenda Castro, Daniel Rodríguez, Ita Cruz, Jessica Méndez, Josué Careaga, Luis Cabrera, Marina Fomicheva and Paulina De La Vega) and interface developers (Luis Cabrera and Juan Rolland).

[6] This course was given in the framework of a last-year subject in the Spanish Linguistics Degree at UNAM (Mexico City).

analyzed (using a given relations list) and, once again, annotators discussed the difficult cases. After the discussion, texts were re-annotated to verify if the difficulties were solved. This process was doubly interesting, since it helped to create common criteria for the annotation of the final corpus and to define the annotation criteria more clearly and consensually, in order to include them in the RST Spanish Treebank annotation manual. Once annotators agreed on the most difficult cases, we consider that the training phase finished.

3.5 Designing and Managing the Annotation Procedure

We start from the following annotation definition:

> Annotation ('tagging') is the process of adding new information into source material by humans (annotators) or suitably trained machines. [...]. The addition process usually requires some sort of mental decision that depends both on the source material and on some theory or knowledge that the annotator has internalized earlier. (Hovy, 2010: 6)

Exactly, after our annotators internalized the theory and annotation criteria during the training phase, the "annotation phase" of the final texts included in the RST Spanish Treebank started. In this phase, the annotation tasks were assigned to annotators (the number of texts assigned to each annotator was different, depending on their availability). They were asked to carry out the annotation individually and without questions among them. We calculated that the average time to carry out the annotation of one text was between 15 minutes and 1 hour. This time difference is due to the fact that the corpus includes both short and long texts. The annotation process is the following: once a text is segmented, rhetorical relations between EDUs are annotated. First, EDUs inside the same sentence are annotated in a binary way. Second, sentences inside the same paragraph are linked. Finally, paragraphs are linked.

Hovy (2010) states that it is difficult to determine if, for the same money (we add "for the same time"), it is better to double-annotate less, or to single-annotate more. As he explains, Dligach et al. (2010) made an experiment with OntoNotes (Pradhan et al., 2007) verb sense annotation. The result was that, assuming the annotation is stable (that is, inter-annotator agreement is high), it is better to annotate more, even with only one annotator. The problem with RST annotation is that there are so many categories to annotate, that is very difficult to obtain a stable annotation. Therefore, we consider it is necessary to have at least some texts double-annotated (or even triple-annotated), in order to have an adequate discourse corpus. This is the reason why, following the RST Discourse Treebank methodology, we use some texts as learning corpus and some others (from the Mathematics, Psychology and Sexuality domains) as test corpus: 69% (183 texts) and 31% (84 texts), respectively. The texts of the learning corpus were annotated by 1 person, whereas the texts of the test corpus were annotated by 2 people.

3.6 Validating Results

Da Cunha and Iruskieta (2010) measure inter-annotator agreement by using the RST trees comparison methodology by Marcu (2000). This methodology evaluates the agreement on 4 elements (EDUs, SPANs, Nuclearity and Relations), by means of precision and recall measures (an annotation with regard to the other one). Following this methodology, we have measured inter-annotator agreement over the test corpus. We employ an on-line automatic tool for RST trees comparison, RSTeval (Mazeiro and Pardo, 2009), where Marcu's methodology has been implemented (for 4 languages: English, Portuguese, Spanish and Basque). We know that there are some other ways to measure agreement, such as Cohen's kappa (Cohen, 1960) or Fleiss's kappa (Fleiss, 1971), for example. Nevertheless, we consider that Marcu's methodology (2000) is suitable to compare adequately 2 annotations of the same original text, because it has been designed specifically for this task.

For each trees pair from the test corpus, precision and recall were measured separately. Afterwards, all those individual results were put together to obtain general results. Table 3 shows global results for the 4 categories. The category with more agreement was EDUs (recall: 91.04% / precision: 87.20%), that is, segmentation. This result was expected, since the segmentation criteria given to the annotators were quite precise and the possibility of mistake was low. The lowest agreement was obtained for the category Relations (recall: 78.48% / precision: 76.81%). This result is lower than the other, but we think it is acceptable. In the RST Discourse Treebank the trend was similar to the one detected in our corpus: the highest agreement is obtained at the segmentation level and the lowest at the relations level.

Category	Precision	Recall
EDUs	87.20%	91.04%
SPANs	86%	87.31%
Nuclearity	82.46%	84.66%
Relations	76.81%	78.48%

Table 3: Inter-annotator agreement

Precision and recall have not been calculated with respect to a gold standard because it does not exist for Spanish. Our future aim is to reach a consensus on the annotation of the test corpus (using an external "judge"), in order to establish a set of texts considered as a preliminary gold standard for this language. We consider that the annotations have quality at present, because inter-annotator agreement is quite high; however, this consensus could solve the typical annotation mistakes we have detected or some ambiguities.

We have analyzed the main discrepancy reasons between annotators. With regard to the segmentation, the main one was human mistake; ex. segmenting EDUs without a verb (one annotator segmented the following passage into 2 EDUs because she detected a Means relation, but the second EDU does not include any verb):

[Además estudiamos el desarrollo de criterios para determinar si un semigrupo dado tiene dicha propiedad] [mediante el estudio de desigualdades de curvatura-dimensión.]

[We also study the development of tests in order to determine if a given semi group has this property] [by means of curvature-dimension inequalities.]

The second reason was that in the manual some aspects were not explained in detail. For example, if a substantive sentence or a direct/object clause (which must not be segmented, according to the point b2) includes two coordinated clauses, these must not be segmented either. Thus, we found some erroneous segmentations. For example:

[Los hombres adultos tienen miedo de fracasar] [y no cumplir con el rol masculino de ser proveedores del hogar y de proteger a su familia.]

[Adult men are scared to fail] [and not to fulfill the masculine role of being the suppliers of the home and to protect their family.]

This kind of mistakes allowed us to refine our segmentation manual *a posteriori*. In the future, we will ask the test corpus annotators to make a new annotation of the texts, using the refined manual, in order to check if the agreement increases, in the same way as the RST Discourse Treebank.

With regard to rhetorical annotations, we detected 2 main reasons of inter-annotator disagreement. The first one was the ambiguity of some relations and their corresponding connectors; for example, Justification-Reason, Antithesis-Concession or Circumstance-Means relations, like in the following passage (in Spanish, “al” may indicate time or manner):

[Los niños aprenden matemáticas] [al resolver problemas.]

[Children learn mathematics] [when solving problems.]

The second one is due to differences between annotators when determining nuclearity. For example, in the following passage, one annotator marked Background and the other one Elaboration:

[Quedó un hueco en la pared de 60 x 1.20cm.]S_Background [Norma y Andrés quieren colocar en el hueco una pecera.]N_Background

[Quedó un hueco en la pared de 60 x 1.20cm.]N_Elaboration [Norma y Andrés quieren colocar en el hueco una pecera.]S_Elaboration

[A hole of 60 x 1.20 cm remained in the wall.] [Norma and Andrés want to place a fish tank in the hole.]

It is easier to solve segmentation disagreement than relations disagreement, since in this case annotator subjectivity is more evident; we must consider how to refine our manual in this sense.

3.7 Delivering and Maintaining the Product

Hovy (2010) mentions some technical issues regarding these points: licensing, distribution, maintenance and updates. With regard to licensing and distribution, the RST Spanish Treebank will be free for research purposes. We have a data manager responsible for maintenance and updates.

The description of the annotated corpus is also a very important issue (Ide and Pustejovsky, 2010). It is important to provide a high level description of the corpus, including the theoretical framework, the methodology (annotators, annotation manual and tool, agreement, etc.), the means for resource maintenance, the technical aspects, the project leader, the contact, the team, etc. The RST Spanish Treebank includes all this detailed information.

XML (with a DTD) has been used, in order the corpus can be reused for several aplications. In the future, we plan to use the standard XCES.

To know more about resources development, linguistic annotation or inter-annotator agreement, we recommend: Palmer et al. (on-line), Palmer and Xue (2010), and Artstein and Poesio (2008).

4 The Search Interface of the RST Spanish Treebank

The RST Spanish Treebank interface is freely available on-line[7]. It allows the visualization and downloading of all the texts in txt format, with their corresponding annotated trees in RSTtool format (rs3), as well as in image format (jpg). Each text includes its title, its reference, its web link (if it is an on-line text) and its number of words. The interface shows texts by areas and allows the user to select a subcorpus (including individual files or folders containing several files). The selected subcorpus can be saved on local disk (generating a xml file) for future analyses.

The interface includes a statistical tool which allows obtaining statistics of rhetorical relations in a subcorpus selected by the user. The RSTtool also offers this option but it can be only used for one text. We consider that it is more useful for the user to obtain statistics from various texts, in order to get significant statistical results. As the RSTtool, our tool allows to count the multinuclear relations in two ways: a) one unit for each detected multinuclear relation, and b) one unit for each detected nucleus. If we use b), the statistics of the multinuclear relations of Table 2 are higher: List (864), Joint (537), Sequence (289), Contrast (153), Conjunction (28) and Disjunction (24).

We are developing another tool, aimed to extract information from the annotated texts, which we will soon include into the interface. This tool will allow to the user to select a subcorpus and to extract from it the EDUs corresponding to the rhetorical relations selected, like a multidocument specialized summarizer guided by user's interests.

The RST Spanish Treebank interface also includes a screen which permits the users to send their own annotated texts. Our aim is for the RST Spanish Treebank to become a dynamic corpus, in constant evolution, being increased with texts annotated by users. This has a double advantage since, on the one hand, the corpus will grow and, on the other hand, users will profit from the interface's applications, using their own subcorpora. The only requirement is to use the relations and the segmentation and annotation criteria of our project. Once the texts are sent, the RST Spanish Treebank data manager will verify if the annotation corresponds to these criteria.

5 Conclusions and Future Work

We think that this work means an important step for the RST research in Spanish, and that the RST Spanish Treebank will be useful to carry out diverse researches about RST in this language, from a descriptive point of view (ex. analysis of texts from different domains or genres) and an applied point of view (development of discourse parsers and NLP applications, like automatic summarization, automatic translation, IE, etc.).

For the moment the corpus' size is acceptable and, though the percentage of double-annotated texts is not very high, we think that having 10 annotators (using the same annotation manual) avoids the bias of only one annotator. In addition, the corpus includes texts of diverse domains and genres, which provides us with a heterogeneous Spanish corpus. Moreover, the corpus interface that we have designed allows the user to select a subcorpus and to analyze it statistically. In addition, we think that it is essential to release a free corpus, on-line and dynamic, that is, in continuous growth. Nevertheless, we are conscious that our work still has certain limitations, which we will try to solve in the future. In the short term, we have 5 aims:

a) To add one more annotator for the test corpus and to measure inter-annotator agreement.
b) To use more agreement measures, like kappa.
c) To reach a consensus on the annotation of the test corpus, in order to establish a set of texts considered as a preliminary gold standard.
d) To finish and to evaluate the IE tool.
e) To analyze the corpus to extract linguistic patterns for the automatic relations detection.

In the long term, we consider other aims:

f) To increase the corpus, by adding non-specialized texts, and new domains and genres.
g) To annotate all the texts by 3 people, to get a representative gold-standard for Spanish (this aim will depend on the funding of the project).

[7] http://www.corpus.unam.mx/rst/

References

Ron Artstein, and Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4):555-596.

Nadjet Bouayad-Agha, Leo Wanner, and Daniel Nicklass. 2006. Discourse structuring of dynamic content. Procesamiento del lenguaje natural, 37:207-213.

M. Teresa Cabré (1999). La terminología: representación y comunicación. Barcelona: IULA-UPF.

Lynn Carlson and Daniel Marcu. 2001. Discourse Tagging Reference Manual. ISI Technical Report ISITR-545. Los Ángeles: University of Southern California.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002a. RST Discourse Treebank. Pennsylvania: Linguistic Data Consortium.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002b. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37-46

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2010. Discourse Segmentation for Spanish based on Shallow Parsing. Lecture Notes in Computer Science, 6437:13-23.

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

Iria da Cunha, Leo Wanner, and M. Teresa Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. Terminology, 13(2):249-286.

Dmitriy Dligach, Rodney D. Nielsen, and Martha Palmer. 2010. To Annotate More Accurately or to Annotate More. In Proceedings of the 4th Linguistic Annotation Workshop (LAW-IV). 48th Annual Meeting of the Association for Computational Linguistics.

Joseph L. Fleis. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378-382.

Eduard Hovy. 2010. Annotation. A Tutorial. Presented at the 48th Annual Meeting of the Association for Computational Linguistics.

Nancy Ide and Pustejovsky, J. (2010). What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010).

William C. Mann, and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3):243-281.

Daniel Marcu. 2000. The Theory and Practice of Discourse Parsing Summarization. Massachusetts: Institute of Technology.

Mitchell P. Marcus, Beatrice Santorini, Mary A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treenbank. Computational Linguistics, 19(2):313-330.

Michael O'Donnell. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory. In Proceedings of the International Natural Language Generation Conference. 253-256.

Martha Palmer, and Nianwen Xue. 2010. Linguistic Annotation. Handbook of Computational Linguistics and Natural Language Processing.

Martha Palmer, Randee Tangi, Stephanie Strassel, Christiane Fellbaum, and Eduard Hovy (on-line). Historical Development and Future Directions in Data Resource Development. MINDS report. http://www-nlpir.nist.gov/MINDS/FINAL/data.web.pdf

Sameer Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In Proceedings of the First IEEE International Conference on Semantic Computing (ICSC-07).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

David Reitter, and Mandred Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In Proceedings of the 4th International

Workshop on Linguistically Interpreted Corpora (LINC-03).

Magdalena Romera. 2004. Discourse Functional Units: The Expression of Coherence Relations in Spoken Spanish. Munich: LINCOM.

Thiago Alexandre Salgueiro Pardo, and Lucia Helena Machado Rino. 2001. A summary planner based on a three-level discourse model. In Proceedings of Natural Language Processing Pacific Rim Symposium. 533-538.

Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes, and Lucia Helena Machado Rino. 2008. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. Lecture Notes in Artificial Intelligence, 3171:224-234.

Thiago Alexandre Salgueiro Pardo, and Eloize Rossi Marques Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. In Anais do V Encontro de Corpora. São Carlos-SP, Brasil.

Gerardo Sierra. 2008. Diseño de corpus textuales para fines lingüísticos. In Proceedings of the IX Encuentro Internacional de Lingüística en el Noroeste 2. 445-462.

Manfred Stede. 2004. The Potsdam commentary corpus. In Proceedings of the Workshop on Discourse Annotation, 42nd Meeting of the Association for Computational Linguistics.

Maite Taboada. 2004. Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish. Amsterdam/Philadelphia: John Benjamins.

Maite Taboada, and Jan Renkema. 2008. Discourse Relations Reference Corpus [Corpus]. Simon Fraser University and Tilburg University. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Maite Taboada, and William C. Mann. 2006a. Rhetorical Structure Theory: Looking Back and Moving Ahead. Discourse Studies, 8(3):423-459.

Maite Taboada, and William C. Mann. 2006b. Applications of Rhetorical Structure Theory. Discourse Studies, 8(4):567-588.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A Syntactic and Lexical-Based Discourse Segmenter. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics.

Hai Zhao, Yan Song, and Chunyu Kit. 2010. How Large a Corpus Do We Need: Statistical Method Versus Rule-based Method. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).