The RST Spanish Treebank On-line Interface

Iria da Cunha
Instituto Universitario de Lingüística Aplicada (UPF)
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon, Universidad Nacional Autónoma de México, École Polytechnique de Montréal
juan-manuel.torres@univ-avignon.fr

Gerardo Sierra
Universidad Nacional Autónoma de México
gsierram@iingen.unam.mx

Luis-Adrián Cabrera-Diego
Universidad Nacional Autónoma de México
LCabreraD@iingen.unam.mx

Brenda-Gabriela Castro-Rolón
Universidad Nacional Autónoma de México
BCastroR@iingen.unam.mx

Juan-Miguel Rolland-Bartilotti
Universidad Nacional Autónoma de México
jrollandb@iingen.unam.mx

Abstract

In this article, we present the on-line interface that we have developed for the RST Spanish Treebank, the first corpus including Spanish texts annotated with rhetorical relations. This interface allows users to consult or download the texts and their corresponding annotations. In addition, it allows carrying out several tasks over a selected subcorpus: searching statistics in terms of words, rhetorical relations and Elementary Discourse Units (EDUs), and extracting information, in terms of texts passages marked with rhetorical relations (ex. Result, Cause or Background), which users may select.

1 Introduction

According to Hovy (2010), there are 7 core questions in corpus' design: selecting a corpus, instantiating the theory, designing the interface, selecting and training the annotators, designing and managing the annotation procedure, validating results, and delivering and maintaining the product. All these points are really relevant when compiling a corpus. However, we consider that usually one of them is underestimated: the interface design. When Hovy (2010) mentions this aspect, he mainly refers to the annotation interface. We think that the annotation interface is important but, as well, that, if there is an available annotation interface suitable for the purposes of a corpus project, it can be used. Nevertheless, we consider that an interface allowing users to consult or download the corpus' texts, and even carrying out searches (both statistics and linguistics) over a selected subcorpus, is really useful and necessary.

Compiling and annotating an adequate corpus is not a trivial task; it implies lots of people, resources, time and effort. Thus, we consider that it is important to develop a friendly and useful interface to be able to exploit the created corpus, and transform it into a most accessible resource. Therefore, in this article, we present the on-line interface that we have developed in order to include the RST Spanish Treebank (da Cunha et al., 2011). The RST Spanish Treebank is the first corpus including Spanish texts annotated with rhetorical relations of the Rhetorical Structure Theory (RST) by Mann and Thompson (1988). It contains texts from nine specialized domains (Astrophysics, Earthquake Engineering, Economy, Law, Linguistics, Mathematics, Medicine, Psychology and Sexuality). It includes 52,746 words, 267 texts, 2,256 sentences and 3,349 discourse segments. The segmentation criteria are similar to those employed by da Cunha et al. (2011). Each text was tagged by 1 person, from a team of 10 RST expert annotators. There is a 31% of the corpus double-annotated. The corpus is not annotated with syntactic structure, although we are conscious this would be interesting. The corpus will be useful for the

Proceedings of Recent Advances in Natural Language Processing, pages 698–703,
Hissar, Bulgaria, 12-14 September 2011.

development of a rhetorical parser for this language and several other applications related to computational linguistics (automatic translation, automatic summarization, information extraction, etc.). In addition, this corpus will be helpful for researchers and students interested on the analysis of rhetorical relations. Thus, before the search interface design, we wondered which kind of information they would need for their discourse studies. We considered that they would like to know the quantity of discourse segments included in a corpus (for example, to compare the discourse complexity among languages), the number of rhetorical relations of each type (for example, to try to characterize the discourse of a genre, a domain or a language, in the same line of Iruskieta and da Cunha, 2011), or to extract text passages corresponding to some rhetorical relations (for example, to determine how these relations are explicit in the text and if they are marked with discourse connectors). With these possible needs in mind, we have developed some search tools and we have included them in the interface. Thus, the interface allows users to consult or download the texts and their corresponding annotations and, in addition, it allows carrying out several statistical and linguistic searches over a selected subcorpus. The interface is available in: http://www.corpus.unam.mx/rst/.

In Section 2, we present some previous work. In Section 3, we explain the development of the interface: the website, the annotated texts selection and downloading interface, the search tools (statistical and linguistic), the annotated texts uploading interface and the administrator interface. In Section 4, we establish some conclusions and future work.

2 Previous Work

Nowadays, there are lots of corpora containing texts annotated at different levels (morphological, syntactic, semantic, etc.), for the majority of the most used languages. Despite this fact, there are not so many corpora annotated with rhetorical relations. The most used rhetorical framework for this task is the RST, an independent language theory departing from 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. Some examples are the relations of Antithesis, Background, Cause, Reformulation or Result. 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. They are the relations of Contrast, List, Joint or Sequence, among others.

Until now, there were RST corpora only for three languages: English (Carlson et al., 2002; Taboada and Renkema, 2008), German (Stede, 2004) and Portuguese (Pardo et al., 2008; Pardo and Seno, 2005). These RST corpora suppose an important step on the RST research and they have been very useful to develop several applications, like information extraction, text generation, automatic summarization, etc. Each one has some advantages and disadvantages, related to the number of included texts and words, the annotation systematicity, the texts' domain heterogeneity, the amount of double-annotated texts (to measure the agreement between annotators), etc. (see da Cunha et al., 2011) Nevertheless, we consider that there is one limitation shared by almost all these corpora: the lack of a free on-line corpus interface, to consult the corpus and to carry out searches over it. Most of these corpora offer a folder containing all the annotated texts individually into the format of the annotation interface RSTtool (O'Donnell, 2000). The only one offering a search tool (allowing to users to search at different linguistic levels) is the German Potsdam Commentary Corpus (Stede, 2004), although, to our knowledge, this tool is not available on-line.

3 Developing the Interface

In this section, we explain all the aspects regarding the developing of the interface.

3.1 The Website

The RST Spanish Treebank is free for research purposes and it can be consulted or downloaded by means of the on-line interface we have developed for it. Ide and Pustejovsky (2010) mention several different kinds of documentation which a corpus project must provide. Following these guideliness, the website including the RST Spanish Treebank contains a high level description of the resource for non-specialist public, annotation guidelines, information on the theoretical framework, project documentation (location, personnel, contact, etc.), corpus documentation, among other information.

The RST Spanish Treebank interface and all the related information are written in Spanish, although they will be also in English soon.

3.2 Annotated Texts Selection and Downloading Interface

The RST Spanish Treebank interface allows the visualization and downloading of all the original documents in plain text format (txt), with their corresponding annotated trees in RSTtool format (rs3), as well as in image format (png). Each text includes its title, its reference, its web link (if it is an on-line text) and its number of words.

The copyright of the texts included in a corpus is a polemical subject. Usually, written authorization to the authors of the texts must be requested in order to include the texts in a corpus. However, as Sierra (2008) explains, there are exceptions or limits in some cases. One of them is the case of non-profit research projects, where only passages of texts (not complete texts) are provided and their origin and corresponding bibliographic reference are stated. This is precisely the case of the RST Spanish Treebank, since it is a non-profit research project which provides the corpus through an interface that includes only passages of the original texts (for example, abstracts of scientific articles, sections of webpages, thesis introductions, etc.) and the bibliographic references (and links, in the cases of electronic publications) of all the documents.

The interface shows texts by areas and allows the user to select a subcorpus (including individual files and/or folders containing several files). The selection of the subcorpus can be saved on local disk (generating an xml file including the IDs of the selected texts) for future analyses.

As the RST Spanish Treebank is a growing corpus, our interface is dynamic too, in order to be able to do changes (for example, to include new domains categories) without modifying the interface code. To solve this challenge, we have developed an in-house program that recursively reads the entire corpus' directory and creates a general xml with the information of each document (as location, number of words, etc.). As well, at the same time, this program creates an individual xml for each file, which contains its bibliographic reference, origin, among other data.

Appendix A includes a screenshot of the texts selection and downloading interface.

3.3 Search Tools

Until now, we have developed four search tools, which are included in the RST Spanish Treebank interface. Three of them are statistical; the other one is linguistic. The four tools are developed in Perl and can be applied over the total corpus or over a subcorpus selected by the user.

3.3.1 Statistic Tools

Firstly, users can know the number of words of the selected subcorpus automatically and in real time. This tool is simple but it is important, because it allows the user to increase or decrease his subcorpus easily regarding his research aims.

Secondly, users may obtain the number of EDUs of the selected corpus, using the tool RST_stats_EDUs. This tool analyses automatically the rs3 archives of the selected subcorpus and it calculates the amount of EDUs present into these texts. This tool is useful to have an idea of the discourse "potential" of a corpus.

Thirdly, the interface includes a statistical tool that allows obtaining statistics of rhetorical relations in a subcorpus selected by the user. It is called RST_stats_Rel. We consider that this is the most useful tool, because the user may carry out statistical researches about the rhetorical relations existing into the texts of the studied corpus, which usually are performed by hand. The RSTtool also offers this option but it can be only used for one text at time. We consider that it is more useful for the user to obtain statistics from various texts, so as 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. For example, Figure 1 shows a RST tree containing a multinuclear relation of Contrast. If we select the strategy a), the tool will count 1, and if we select the strategy b), the tool will count 2.

English translation: One patient was found in breathing acidosis, whereas 5 presented chronic breathing alkalosis.

Figure 1: Example of multinuclear Contrast relation

Table 1 contains the list of the nucleus-satellite relations of the RST Spanish Treebank,

with the number and percentage of rhetorical relations, calculated by RST_stats_Rel.

Relation	Quantity	
	Nº	%
Elaboration	765	24.56
Preparation	475	15.25
Background	204	6.55
Result	193	6.20
Means	175	5.62
Circumstance	140	4.49
Purpose	122	3.92
Interpretation	88	2.83
Antithesis	80	2.57
Cause	77	2.47
Evidence	59	1.89
Condition	53	1.70
Concession	50	1.61
Justification	39	1.25
Solution	32	1.03
Motivation	28	0.90
Reformulation	22	0.71
Otherwise	3	0.10
Evaluation	11	0.35
Summary	8	0.26
Enablement	5	0.16
Unless	2	0.06

Table 1: Amount of nucleus-satellite rhetorical relations in the RST Spanish Treebank

Table 2 includes the list of multinuclear relations of the corpus, using strategies a) and b). As it can be observed, using b), the amount of detected relations is higher than using a).

Relation	Quantity			
	Strategy a		Strategy b	
	Nº	%	Nº	%
List	172	5.52	864	19.09
Joint	160	5.14	537	11.86
Sequence	74	2.38	289	6.39
Contrast	58	1.86	153	3.38
Conjunction	11	0.35	28	0.62
Disjunction	9	0.29	24	0.53

Table 2: Amount of multinuclear rhetorical relations in the RST Spanish Treebank

3.3.2 Linguistic Tool

The RST_extract is a tool aimed to extract information from the annotated texts. This tool allows the user to select a subcorpus and to extract from it the EDUs corresponding to the rhetorical relation selected, like a multidocument specialized summarizer guided by user's interests. This tool might be useful, for example, to elaborate a compendium of results of diverse medical articles about a certain topic (selecting the relation of Result) or to compile a state of the art about one topic (selecting the relation of Background). At present some monodocument summarizers based on RST exist for some languages (Marcu 2000; Pardo and Rino, 2001; da Cunha et al., 2007, among others), but, at our knowledge, no multidocument specialized RST summarizers exist. We can mention here the works about multidocument summarization for Portuguese based on the Cross-document Structure Theory (CSS) (Radev, 2000), a theory derived from RST (Jorge and Pardo, 2010). Figure 2 contains a passage of the output of the RST_extract, applying it over the subcorpus of Sexuality and selecting the rhetorical relation of Result (the English tranlation is ours). We show 3 of the 20 extracted Result satellites.

```
se00028.rs3: La hipertrofia del epitelio produce acantosis y la aparición de papiloma de 3 meses a 2 años después del inicio de la infección.
    The epithelium hypertrophy causes acanthosis and the occurrence of papilloma from 3 months to 2 years after the beginning of the infection.

se00032.rs3: Las complicaciones más graves de la enfermedad inflamatoria pélvica son la esterilidad y embarazo ectópico secundario.
    The most severe complications of the pelvic inflammatory illness are the sterility and the secondary ectopic pregnancy.

se00032.rs3: La infección puede ascender y dar como resultado salpingitis, abscesos tubo-ováricos y enfermedad inflamatoria pélvica.
    The infection can rise and to give as result salpingitis, tube-ovarian abscesses and pelvic inflammatory illness.
```

Figure 2: Example of the output of RST_extract

RST_extract uses as input the rs3 files from RSTTool. Due to the complexity of this kind of format, for the moment, our tool only extracts satellites of nucleus-satellite relations, being simple EDUs (not SPANs).

3.4 Annotated Texts Uploading Interface

The RST Spanish Treebank interface also includes a screen that permits the users to send comments, suggestions, and also to send their

own annotated texts. Our aim is for the RST Spanish Treebank to become a dynamic corpus, in constant evolution, increasing 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.

3.5 Administrator Interface

The sustainability of a language resource is a crucial aspect. As Ide and Pustejovsky (2010) assess, "means for resource preservation and maintenance should be established prior to publication to ensure continued availability [...]. In the case where resources are distributed via the web [...], ensured sustainability is the reponsibility of the resource developer". Having this requirement in mind, we have a data manager who is the responsible for the administration of the RST Spanish Treebank and its interface. This manager is the person in charge of the new texts and information that will be included in the corpus (both texts from users and texts selected by our research team). Data manager work is important because, although a part of the task is automatic (texts uploading), the texts data (ID, title, bibliographic reference and link) are included semi-automatically.

The administrator interface is divided in two parts. The first one is a program that connects to the server through Secure Shell Protocol; using this application, the data manager can upload all the files to be added at the corpus and set their location. The second part of the administrator interface is an on-line template that includes four fixed fields (text ID, title, reference and link). Once the documents are uploaded and their templates are filled, the data manager can press an update button. This button brings up to date the general xml of the corpus and the individual xml of each file, and executes the first statistical tool to count the number of words of each new file at the server.

4 Conclusions

In this paper, we have presented the RST Spanish Treebank interface that we have developed in order to include the RST Spanish Treebank, the first corpus containing Spanish texts annotated with RST relations. As we have shown, this interface allows users to consult or download the texts and their corresponding annotations freely and on-line. Moreover, it allows carrying out several statistical and linguistic searches over a selected subcorpus. We consider this interface is necessary and useful to exploit all the data contained in a corpus, which in this case will be in continuous growth.

We think that this work means an important step for the RST research in Spanish. Additionally, the RST Spanish Treebank and its interface will be useful to carry out diverse researches about RST in this language. These researches can be developed both from a descriptive point of view (contrastive analysis among specialized texts from different domains, analysis of genres, analysis of discourse markers, etc.) and an applied point of view (development of discourse parsers, development of natural language processing applications, like automatic summarization, automatic translation, information extraction, etc.). In addition, we consider that this interface would be useful to contain and analyze automatically RST corpora for other languages, because the interface architecture would allow it without too much adaptation effort.

As future work, we would like to insert a sentence segmentator (in order to count sentences automatically) and to optimize the RST_extract tool (in order to extract satellites and nuclei being SPANs, not only EDUs).

Acknowledgments

This research was supported by the research project CONACyT (Mexico), n. 82050; and the Spanish projects RICOTERM (FFI2010-21365-C03-01) and APLE (FFI2009-12188-C05-01).

References

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

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the Development of the RST Spanish Treebank. In Proc. of the 5th Linguistic Annotation Workshop. 49th Annual Meeting of the ACL. 1-10.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2011. DiSeg 1.0: The First System for Spanish Discourse Segmentation. Expert Systems with Applications.

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

Mikel Iruskieta and Iria da Cunha. 2010. El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera-español. Calidoscópio, 8(3):181-202.

Maria Lucía del Rosario Castro Jorge and Thiago Alexandre Salgueiro Pardo. 2010. Experiments with CST-based Multidocument Summarization. In Proc. of the ACL Work. TextGraphs-5: Graph-based Methods for NLP. 74-82.

Eduard Hovy. 2010. Annotation. A Tutorial. Presented at the 48th Annual Meeting of ACL.

Nancy Ide and James Pustejovsky. 2010. What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In Proc. of the 2º Int. Conf. on Global Interoperability for Language Resources.

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.

Michael O'Donnell. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory. In Proc. of Int. Natural Lang. Generation Conf.. 253-256.

Dragomir R. Radev. 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In Proc. of the 1st ACL SIGDIAL Work. on Discourse and Dialogue.

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 Proc. of the IX Encuentro Inter. de Lingüística en el Noroeste 2. 445-462.

Manfred Stede. 2004. The Potsdam commentary corpus. In Proc. of the Workshop on Discourse Annotation. 42nd Meeting of ACL.

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

Appendix A. Screenshot of the interface with an annotated text