The RST Spanish-Chinese Treebank

Shuyuan Cao
Universitat Pompeu Fabra (UPF)
shuyuan.cao@hotmail.com

Iria da Cunha
Universidad Nacional de Educación a Distancia (UNED)
iriad@flog.uned.es

Mikel Iruskieta
University of Basque Country (UPV-EHU)
mikel.iruskieta@ehu.eus

Abstract

Discourse analysis is necessary for different tasks of Natural Language Processing (NLP). As two of the most spoken languages in the world, discourse analysis between Spanish and Chinese is important for NLP research. This paper aims to present the first open Spanish-Chinese parallel corpus annotated with discourse information, whose theoretical framework is based on the Rhetorical Structure Theory (RST). We have evaluated and harmonized each annotation part to obtain a high annotated-quality corpus. The corpus is already available to the public.

1 Introduction

Spanish and Chinese are two of the most spoken languages in the world; the language pair occupies an important position in the Natural Language Processing (NLP) research world. Recently, discourse analysis has called much attention as an unsolved problem and is crucial for many NLP tasks (Zhou et al., 2014). The great language distance causes a great number of discourse differences between Spanish and Chinese. Comparative or contrastive studies of discourse structures reveal information to identify properly equivalent discourse elements in a language pair (Cao and Gete, 2018). Here we give an example to show the discourse similarity and difference between the two languages.

Ex.1[1]:

1.1 Sp: Aunque aún no contamos con resultados, intuimos que el modelo será más amplio que el del sintagma nominal.

[Aunque aún no contamos con resultados,]Unit$_1$ [intuimos que el modelo será más amplio que el del sintagma nominal.]Unit$_2$

[DM[2] still no get results,] [we consider that the model will more extensive than the sentence group nominal.] [3]

1.2 Sp: Intuimos que el modelo será más amplio que el del sintagma nominal, aunque aún no contamos con resultados.

[Intuimos que el modelo será más amplio que el del sintagma nominal,]Unit$_1$ [aunque aún no contamos con resultados.]Unit$_2$

[We consider that the model will more extensive than the sentence group nominal,] [DM still no get results.]

1.3 Ch: 尽管还没有取得最终结果，但是我们认为该模型已囊括了语段模型涉及的内容。

[1] The examples have been extracted from the corpus.

[2] DM means discourse marker. In this work, we use the definition of DM by Eckle-Kohler, Kluge and Gurevych (2015). DMs are used to signal discourse relations in a text segment. Specially, the DMs in our work are traditional markers and markers including verbal structures, as da Cunha indicates (2013).

[3] In this work, we give an English literal translation for both Spanish and Chinese examples in order to make the readers understand the content better.

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 156–166
Santa Fe, New Mexico, USA, August 25-26, 2018.

[尽管还没有取得最终结果，]Unit$_1$ [但是我们认为该模型已囊括了语段模型涉及的内容。]Unit$_2$

[DM1 still no get results,] [DM2 we consider that the model contains the sentence group nominal.]

1.4 Eng: Although we haven’t got the results yet, we consider that the model will be more extensive than the nominal sentence group.

In Example 1, we can see that the Spanish passage has a similar discourse structure to the Chinese passage. Both passages start the text with a discourse marker in the first unit. However, the usage of discourse markers in both languages is different. To show same meaning, in Chinese, it is mandatory to include two discourse markers: one marker is “*jinguan*” (尽管), at the beginning of the first unit, and another marker is “*danshi*” (但是), at the beginning of the second unit. These two discourse markers are equivalent to the English discourse marker ‘although’. By contrast, in Spanish, just one discourse marker “*aunque*” is being used at the beginning of the first unit, and this discourse marker is also equivalent to the English discourse marker *although*. Moreover, the order of the discourse units in the Spanish passage can be changed and it makes sense syntactically, but the order cannot be changed in the Chinese passage, because neither syntactically nor grammatically makes sense.

Additionally, as a large electronic library, a corpus can provide a large amount of linguistic information (Wu, 2014). Therefore, this paper aims to present the first open Spanish-Chinese parallel corpus with annotated discourse information under the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988).

In the second section, we present the theoretical framework of this study. In the third section, we talk about some related works. In the fourth section, we give detailed information about the research corpus. In the fifth section, we discuss how we carry out the study by introducing different annotation steps. In the sixth section, we evaluate the annotation results and give the qualitative analysis about the annotation quality. In the last section, we conclude our work and look ahead at our future work.

2 Theoretical framework

The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a theory that was created especially for discourse analysis. RST addresses both hierarchical and relational aspects of text structures for discourse analysis. Elementary Discourse Units (EDUs) (Marcu, 2000) and coherence relations are established in RST. Relations are recursive in RST and are held between EDUs, which can be Nuclei or Satellites, denoted by N and S. Satellites offer additional information about nuclei. EDUs can be linked among them holding a nucleus-satellite (e.g. CAUSE, JUSTIFY, EVIDENCE) function or a multinuclear (e.g. CONJUNCTION, LIST, SEQUENCE) function. As relations are recursive, all the discourse units of the text have a function in a treelike structure, if and only if the text is coherent.

Comparing to other discourse theory, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), RST focuses on the hierarchical structure of a whole text, where discourse relations can be annotated within a sentence (intra-sentence style) and between sentences (inter-sentence style). The intra-sentence annotation and inter-sentence annotation styles help to inform how discourse elements are being expressed in a language.

3 State of the art

3.1 Comparative discourse study

Some previous research using RST for comparative discourse has compared Chinese and English. Cui (1986) presents some aspects regarding discourse relations between Chinese and English; Kong (1998) compares Chinese and English business letters; Guy (2000, 2001) compares Chinese and English journalistic news texts. The only work that compares Spanish and Chinese using RST is by Cao, da Cunha and Bel (2016). They explore sentences that contain the Spanish discourse marker *aunque* (‘although’) and their Chinese parallel sentences in the UN subcorpus.

3.2 RST Treebanks for different languages

With the development of discourse analysis, annotated corpora with relational discourse structure under the RST exist for several languages: (i) for English, the RST Discourse Treebank (Carlson, Marcu and Okurowski, 2001)[4] and the Discourse Relations Reference Corpus (Taboada and Renkema, 2008)[5]; (ii) for German, the Potsdam Commentary Corpus (Stede and Neumann, 2014)[6]; (iii) for Spanish, the RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011)[7]; (iv) for Basque, the RST Basque Treebank (Iruskieta et al., 2013[8]; (v) for Portuguese, the CorpusTCC (Pardo, Nunes and Rino, 2008) and *Rhetalho* (Pardo and Seno, 2005)[9]; (vi) for Russian, the Russian RST Treebank (Toldova et al., 2017)[10].

Bilingual and multilingual RST Treebanks are not common; Iruskieta, da Cunha and Taboada, (2015) create one of the few with the Multilingual RST Treebank[11] for Spanish, Basque and English. For Basque and Spanish, Imaz and Iruskieta (2017) establish the RST Basque-Spanish DELIB Treebank[12]. To our knowledge, our corpus is the first bilingual corpus serves for the discourse analysis between Spanish and Chinese under the RST.

4 Research corpus

There are currently few parallel Spanish-Chinese corpora. the already existing parallel corpora are: (i) The Holy Bible (Resnik, Olsen & Diab, 1999), (ii) The United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009) and (iii) Sina Weibo Parallel Corpus (Wang et al., 2013). Cao, da Cunha and Iruskieta (2017) indicate the three corpora contain their own limitations for Spanish-Chinese comparative discourse analysis. To carry out our work, we develop a new Spanish-Chinese parallel corpus.

Complexity of discourse structure and heterogeneity are the main characteristics taken into account for corpus development. The specific considerations are the following: (a) texts with different sizes (between 100 and 2,000 words), (b) specialized texts and non-specialized texts, (c) texts from different domains, (d) texts from different genres, (e) texts from different original publications, and (f) texts from different authors.

Based on the mentioned aspects, finally, we selected 100 texts to form our research corpus[13]. The genres of the texts are the following: (a) abstracts of research papers, (b) news, (c) advertisements, and (d) announcements. The longest text of the corpus contains 1,774 words and the shortest one contains 111 words.

The sources of these texts are: (a) International Conference about Terminology (1997), (b) Shanghai Miguel Cervantes Library, (c) Chamber of Commerce and Investment of China in Spain, (d) Spain Embassy in Beijing, (e) Spain-China Council Foundation, (f) Confucius Institute Foundation in Barcelona, (g) Beijing Cervantes Institute and (h) Granada Confucius Institute.

The corpus includes texts related to seven domains: (a) terminology (30 texts), (b) culture (12 texts), (c) language (16 texts), (d) economy (14 texts), (e) education (8 texts), (f) art (10 texts), and (g) international affairs (10 texts).

The corpus was enriched automatically with POS information by using the Stanford parser (Levy and Manning, 2003) for Chinese.

Finally, we created an online interface to access the research corpus: `http://ixa2.si.ehu.es/rst/zh/`. Users can search POS information, discourse segments, key information and discourse structure of each text in the research corpus. Moreover, users can also download the texts of the corpus.

[4] https://catalog.ldc.upenn.edu/LDC2002T07 [Last consulted: 06 of July of 2017]

[5] http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html [Last consulted: 06 of July of 2017]

[6] http://angcl.ling.uni-potsdam.de/resources/pcc.html [Last consulted: 06 of July of 2017]

[7] http://corpus.iingen.unam.mx/rst/citar.html [Last consulted: 06 of July of 2017]

[8] http://ixa2.si.ehu.es/diskurtsoa/en/ [Last consulted: 06 of July of 2016]

[9] http://www.icmc.usp.br/~taspardo/projects.htm [Last consulted: 06 of July of 2017]

[10] https://github.com/nasedkinav/rst_corpus_rus [Last consulted: 06 of July of 2017]

[11] http://ixa2.si.ehu.es/rst/ [Last consulted: 06 of July of 2017]

[12] http://ixa2.si.ehu.es/diskurtsoa/rstfilo/ [Last consulted: 06 of July of 2017]

[13] Due to the limited resources that guarantee the complexity of discourse structure and heterogeneity, finally, we choose 50 Spanish texts and their translated Chinese texts (50 texts) to form the corpus.

5 Methodology

In this work, firstly, we elaborate some criteria to segment the corpus. Secondly, we annotate the Central Unit (CU) for each text following Iruskieta et al. (2014). Lastly, we annotate the discourse structure for each text in the corpus following Pardo (2005).

5.1 Segmentation annotation

Segmentation affects the discourse annotation quality; this makes it a crucial step for RST study. Two notable works for Spanish segmentation from the discourse level are mentioned previously: the RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011) and the Multilingual RST Treebank (Iruskieta, da Cunha and Taboada, 2015). Few works focus on the Chinese segmentation from the discourse level under RST. There are three works that use form-based criteria that use punctuation marks to elaborate segmentation rules for Chinese (Yue, 2006; Qiu, 2010; Li, Feng and Zhou, 2013).

In our work, we elaborate the discourse segmentation criteria proposal for both Spanish and Chinese based on linguistic function (the function of the syntactic components) and linguistic form (punctuation category and verbs). We have not considered the meaning (of any coherence relation between propositions) to segment EDUs to avoid circularity in the annotation process. For the function and form perspective, we adopt the segmentation criteria from Iruskieta, da Cunha and Taboada (2015). A Spanish-Chinese bilingual linguists and two Spanish linguists are in charge of the segmentation for the Spanish subcorpus while the bilingual linguists and a Chinese linguists carry out the segmentation task for the Chinese subcorpus[14]. The segmentation tool is the RSTTool (O'Donnell, 2000). Table 1 shows the segmentation criteria[15].

Criteria to form an EDU	Non EDU criteria
Every EDU should have an adjunct verb clause	Relative, modifying and appositive clauses
Paragraphs with line breaks (titles)	Reported speech
Period and question exclamation marks	Truncated EDUs (same-unit)
Comma + adjunct verb clause	
Semicolon + adjunct verb clause	
Colon + adjunct verb clause	
Parenthetical & dash + adjunct verb clause	
Coordination with two subordinate verb clauses	

Table 1: The segmentation criteria

5.2 Central Unit (CU) annotation

Under RST, for each segmented text, among the EDUs, there is an EDU called Central Unit (CU) that contains the key information of the text (Iruskieta, Labaka and Desiderato, 2016).

According to van Dijk (1980), language users are able to summarize discourses, expressing the main topics of the summarized discourse. For our work, for all the segmented texts, the annotators decide which EDUs represent the main idea of the text. Table 2 shows the statistical information of the segmented texts and the annotated CUs in the corpus.

[14] The bilingual expert annotates all the 100 texts, each of the two Spanish experts annotate 25 Spanish texts. The Chinese expert annotate all the 50 Chinese texts.

[15] For the examples of the segmentation criteria, consult Cao et al (2017).

Corpus part	EDUs	CUs
Spanish	840	76
Chinese	953	81

Table 2: Statistical information of EDUs and CUs in the corpus

A Spanish-Chinese bilingual linguist and a Spanish linguist annotate the CUs for all the Spanish texts. The bilingual linguist and a Chinese linguist selected the CUs for all the Chinese texts.

5.3 Discourse structure annotation

Discourse structure annotation is one of the most difficult challenges for annotation works (Hovy and Lavid, 2010). Our study adopts the intra-sentence annotation and inter-sentence annotation styles. Figure 1 shows an annotated parallel Spanish-Chinese text[16].

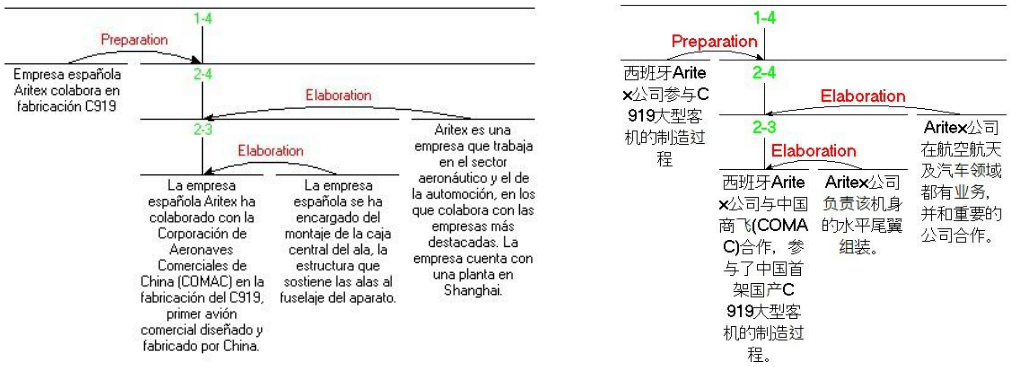
1-4
Preparation
Empresa española Aritex colabora en fabricación C919
2-4
Elaboration
2-3
Elaboration
La empresa española Aritex ha colaborado con la Corporación de Aeronaves Comerciales de China (COMAC) en la fabricación del C919, primer avión comercial diseñado y fabricado por China.
La empresa española se ha encargado del montaje de la caja central del ala, la estructura que sostiene las alas al fuselaje del aparato.
Aritex es una empresa que trabaja en el sector aeronáutico y el de la automoción, en los que colabora con las empresas más destacadas. La empresa cuenta con una planta en Shanghai.
1-4
Preparation
西班牙Aritex公司参与C919大型客机的制造过程
2-4
Elaboration
2-3
Elaboration
西班牙Aritex公司与中国商飞(COMAC)合作，参与了中国首架国产C919大型客机的制造过程。
Aritex公司负责该机身的水平尾翼组装。
Aritex公司在航空航天及汽车领域都有业务，并和重要的公司合作。

Figure 1: Example of the annotated parallel text (EEP1)

The selected discourse relations are presented in the Table 3. The annotation tool we use is the RSTTool (O'Donnell, 2000). All the annotation results are saved by the rstWeb (Zeldes, 2016).

N-S		N-N
Antithesis	Background	Conjunction
Circumstance	Cause	Contrast
Condition	Concession	Disjunction
Enablement	Elaboration	List
Evaluation	Evidence	Sequence
Justify	Interpretation	
Motivation	Means	
Purpose	Otherwise	
Restatement	Preparation	
Solutionhood	Result	
Summary		

Table 3: Selected relations for the discourse annotation[17]

[16] English translation of the text EEP1: [Spanish company Aritex collaborates in manufacturing C919]
[The Spanish company Aritex has collaborated with the Commercial Aircraft Corporation of China (COMAC) in the manufacture of the C919, the first commercial aircraft designed and manufactured by China.] [The Spanish company has been responsible for the assembly of the central wing box, the structure that holds the wings to the fuselage of the aircraft.] [Aritex is a company that works in the aeronautical and automotive sectors, in which it collaborates with the most outstanding companies. The company has a plant in Shanghai.]

[17] The selected relations are extracted from the RST webpage: http://www.sfu.ca/rst/02analyses/index.html [Last consulted: 06 of July of 2017]. The selected relations are the common used ones for RST studies.

6 Evaluation result

6.1 Segmentation annotation evaluation

In this annotation level, we use Cohen Kappa to measure inter-annotator agreement of the segmented discourse units[18]. Kappa calculates the agreement between annotators as:

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

where (A) represents the current observed agreement, and P(E) represents chance agreement. Kappa was calculated by considering titles, parentheses, and verbs, as EDUs candidates. Table 4 includes the agreement between the annotators for the Spanish subcorpus and the Chinese subcorpus.

Corpus Source	**Kappa Agreement**	
	Spanish	Chinese
ICT	0.895	0.815
SMCL	0.945	0.719
CCICS	0.855	0.744
SEB	0.786	0.711
SCCF	0.828	0.711
CIFB	0.716	0.616
BCI	0.863	0.759
GCI	0.873	0.705
Total	0.87	0.76

Table 4: Segmentation annotation agreement of the entire corpus

From Table 4, we can see that, for the Spanish subcorpus, the highest agreement between the annotators is 0.945, and the lowest agreement is 0.716. The agreement for the whole Spanish subcorpus is 0.87. The highest agreement result for the Chinese subcorpus is 0.815, and the lowest agreement result is 0.616. The agreement for the entire Chinese subcorpus is 0.76. The annotation results prove the segmentation criteria are reliable for the language pair. Based on the results, we analyze the segmentation errors to improve the segmentation annotation quality.

6.2 Central Unit (CU) annotation evaluation

Same as the segmentation evaluation, we also use Kappa to measure the CU annotation agreement for exact match. Table 5 shows the evaluation results of the Spanish subcorpus and Table 6 reflects the agreement of the Chinese subcorpus.

From Table 5, we can see that for the Spanish subcorpus, the agreement is 0.961 and the agreement is 0.977 for the Chinese subcorpus (see Table 6). The results show that the CU annotation for the whole corpus is almost perfect.

A1	**A2**		**Total**	**Kappa**
	Yes	No		
Yes	61	16	77	0.961
No	13	750	763	
Total	74	766	840	

Table 5: CU annotation evaluation result of the Spanish subcorpus

[18] For all the annotation steps, the Spanish-Chinese bilingual annotator is assigned as A1, the two Spanish annotators are assigned as A2 (considering as one annotator) and the Chinese annotator is assigned as A3. The agreement are measured between A1 and A2, A1 and A3.

A1	A2		Total	Kappa
	Yes	No		
Yes	55	13	68	0.977
No	7	878	885	
Total	62	881	953	

Table 6: CU annotation evaluation result of the Chinese subcorpus

Based on the annotation results, the annotators discuss the disagreements to confirm the correct CUs for each text in the corpus.

6.3 Discourse structure annotation evaluation

For the discourse structure annotation evaluation, we follow a newly created qualitative method by Iruskieta, da Cunha and Taboada (2015). Under this qualitative method, four elements are being examined by using F-measure: Nuclearity (N), Relation (R), Composition (C) and Attachment (A). In addition, to use this method for the discourse evaluation between two or more languages, the comparison parts must be aligned and must contain the same number of EDUs, to avoid confusing analysis disagreement and segmentation disagreement. The following example explains how we follow this comparison rule by using our corpus:

Ex.2: Text name: CCICE3_ESP & CCICE3_CHN

Sp: [El jueves, el Tesoro volverá a los mercados con una subbasta de bonos y obligaciones en la que intentará colocar entre 3.000 y 4.000 millones.]

[On Thursday, the Treasury will return to the markets with a subbase of bonds and obligations in which it will try place between 3,000 and 4,000 million.]

Ch: [另外，财政部将在本周四再次回到市场拍卖中长期国债，] [欲拍卖30亿至40亿欧元。]

[In addition, the Ministry of Finance will on Thursday again return to the markets to auction of medium-term and long-term treasury bonds,] [to auction 3 billion to 4 billion euros.]

Eng: On Thursday, the Treasury will return to the markets with a sub-base of bonds and obligations in which it will try to place between 3,000 and 4,000 million.

From the above example, we can see the Spanish message is an interdependent EDU and its parallel Chinese message contains three EDUs. For the qualitative comparison, Iruskieta, da Cunha and Taboada (2015) suggest a simple rule, which is to erase the segmentation differences and get the same number of EDUs for the parallel content. Therefore, we combine three Chinese EDUs as a discourse unit (DU) or text span. Although the harmonization process erases some rhetorical relations, the higher level of RS-Tree structure is not affected.

Table 7 shows the evaluation results of the original Spanish subcorpus and the original Chinese subcorpus, meanwhile Table 8 shows the qualitative evaluation of the harmonized corpus.

From Table 7, we can conclude that in the Spanish subcorpus, the agreement of the Nuclearity is from 0.761 to 1, the agreement of the Relation is from 0.641 to 1, the agreement of the Composition is from 0.761 to 0.947, and the agreement of the Attachment is from 0.731 to 0.933. The annotation evaluation results of the Chinese subcorpus shows the agreement of the Nuclearity is from 0.864 to 0.978, the agreement of the Relation is from 0.727 to 0.844, the agreement of the Composition is from 0.864 to 0.978, and the agreement of the Attachment is from 0.84 to 0.978. The evaluation results prove that the annotation of the Spanish subcorpus and the annotation of the Chinese subcorpus are reliable. Two aspects explain why we have the good annotation results: (i) the annotation guideline has been discussed many times and (ii) some texts in the corpus are general publications and the discourse structure of these texts are more simple than others.

Table 8 informs that in the harmonized corpus, the agreement of the Nuclearity is from 0.855 to 1, the agreement of the Relation is from 0.794 to 0.923, the agreement of the Composition is from 0.855 to 1, and the agreement of the Attachment is from 0.855 to 1.The evaluation results of the harmonized corpus are better than the original corpus because of the removal of the annotation disagreements during the harmonized process for both Spanish subcorpus and Chinese subcorpus.

The qualitative analysis and quantitative evaluation results of the harmonized corpus demonstrate the reliability of the annotation quality. The reason that we get the good results is because of: (i)

Before carrying out the annotation work, we elaborate the annotation guideline, which requires the same inter-sentence annotation process and intra-sentence annotation process, and (ii) comparing to other annotation campaigns and texts (news, argumentation texts, scientific texts and abstracts), some texts have a simpler discourse structure.

Source	**Corpus**	**Nuclearity**		**Relation**		**Composition**		**Attachment**	
		Match	F	Match	F	Match	F	Match	F
ICT	Spanish	290/315	0.921	268/315	0.851	290/315	0.921	288/315	0.914
	Chinese	313/357	0.877	278/357	0.779	313/357	0.877	312/357	0.874
SMCL	Spanish	51/67	0.761	43/67	0.641	51/67	0.761	49/67	0.731
	Chinese	66/72	0.917	58/72	0.806	66/72	0.917	66/72	0.917
CCICS	Spanish	37/41	0.902	30/41	0.732	36/41	0.878	37/41	0.902
	Chinese	44/45	0.978	38 /45	0.844	44/45	0.978	44/45	0.978
SEB	Spanish	54/57	0.947	50/57	0.877	54/57	0.947	53/57	0.930
	Chinese	60/64	0.938	54/64	0.844	60/64	0.938	60/64	0.938
SCCF	Spanish	46/50	0.92	37/50	0.74	46/50	0.92	45/50	0.90
	Chinese	62/65	0.954	51/65	0.785	62/65	0.954	62/65	0.954
CIFB	Spanish	39/44	0.886	34/44	0.773	39/44	0.886	38/44	0.864
	Chinese	44/50	0.88	41/50	0.82	44/50	0.88	42/50	0.84
BCI	Spanish	96/108	0.889	83/108	0.769	96/108	0.889	96/108	0.889
	Chinese	122/134	0.910	110/134	0.821	122/134	0.910	122 /134	0.910
GCI	Spanish	15/15	1	15/15	1	14/15	0.933	14/15	0.933
	Chinese	19/22	0.864	16/22	0.727	19/22	0.864	19/22	0.864

Table 7: Qualitative evaluation of the Spanish annotation and Chinese annotation

Source	**Nuclearity**		**Relation**		**Composition**		**Attachment**	
	Match	F	Match	F	Match	F	Match	F
ICT	275/285	0.965	242/285	0.846	274/285	0.961	274/285	0.961
SMCL	59/69	0.855	55/69	0.797	59/69	0.855	59/69	0.855
CCICS	34/34	1	27 /34	0.794	31/34	0.912	31/34	0.912
SEB	46/48	0.958	41/48	0.854	45/48	0.938	45/48	0.938
SCCF	40/42	0.952	35/42	0.833	40/42	0.952	40/42	0.952
CIFB	29/31	0.935	28/31	0.82	29/31	0.935	29/31	0.935
BCI	99/103	0.961	95/103	0.922	97/103	0.942	97/103	0.942
GCI	13/13	1	12/13	0.923	13/13	1	13/13	1

Table 8: Qualitative evaluation of the harmonized corpus between Spanish and Chinese

7 Conclusion

In this work, we present the first RST Spanish-Chinese Treebank with open access. We annotate the discourse information for all the 100 texts by using the RSTTool. We use Kappa to evaluate the annotation quality. The evaluation results for each annotation step show that we get an annotated corpus with high quality. Our corpus fills an important gap for Spanish-Chinese discourse analysis. Moreover, the corpus texts can be downloaded online. The POS information, discourse segments, CU information and the annotations of discourse structure can also be found online.

The corpus can be used for different NLP tasks, for instance, Spanish-Chinese language learning, evaluation of the machine translation (MT) between the two languages from the discourse level, information retrieval, etc. In the future, we will select and annotate more Spanish-Chinese parallel texts and will develop a protocol to help the MT for the language pair.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions on this work. Iria da Cunha is funded by a Ramón y Cajal contract (RYC-2014-16935) associated with the National Distance Education University (UNED) in Spain. Mikel Iruskieta is funded by the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE).

References

Cao Shuyuan, Xue Nianwen, da Cunha Iria, Iruskieta Mikel, and Wang Chuan. 2017. Discourse Segmentation for Building a RST Chinese Treebank. In *Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms*, 73-81.

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2016. A Corpus-based Approach for Spanish-Chinese Language Learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA3)*, 97-106.

Cao Shuyuan, da Cunha Iria, and Bel Nuria. 2016. An analysis of the Concession relation based on the Spanish discourse marker aunque in a Spanish-Chinese parallel corpus. *Procesamiento del Lenguaje Natural*, 56: 81-88.

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2017. Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *EPiC Series of Language and Linguistics*, 2: 315-324.

Cao Shuyuan, and Gete Harritxu. 2018. Using Discourse Information for Education with a Spanish-Chinese Parallel Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'2018)*, 2254-2261.

Carlson Lynn, Marcu Daniel, and Okurowski Mary Ellen. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse Dialogue*, 1-10.

Cui Songren. 1985. *Comparing Structures of Essays in Chinese and English*. Master thesis. Los Angeles: University of California.

da Cunha Iria. 2013. A Symbolic Corpus-based Approach to Detect and Solve the Ambiguity of Discourse Markers. *Research in Computing Science*, 70: 95-106.

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

da Cunha Iria, SanJuan Eric, Torres-Moreno Juan-Manuel, Lloberes Marina, and Castellón Irene. 2012. DiSeg 1.0: The First System for Spanish Discourse Segmentation. *Expert Systems with Applications (ESWA)*, 39(2): 1671-1678.

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

da Cunha Iria, Torres-Moreno Juan-Manuel, Sierra Gerardo; Cabrera-Diego Luis Adrián; Castro Rolón Brenda Gabriela; and Rolland Bartilotti Juan Miguel. 2011. The RST Spanish Treebank On-line Interface. In *Proceedings of Recent Advances in Natural Language Processing (RANLP'2011)*, 698-703.

Eckle-Kohler Judith, Kluge Roland., and Gurevych Iryna. 2015. On the Role of Discourse Markers for Discriming Claims and Premises in Argumentative Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'2015)*, 2236-2242.

Guy Ramsay. 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): 241-58.

Guy Ramsay. 2001. What Are They Getting At? Placement of Important Ideas in Chinese Newstext: A Contrastive Analysis with Australian Newstext. *Australian Review of Applied Linguistics*, 24(2): 17-34.

Hovy Eduard, and Lavid Julia. 2010. Toward a 'Science' of Corpus Annotation: A New Methodology Challenges for Corpus Linguistics. *International Journal of Translation*, 22(1): 13-36.

Imaz Oier, and Iruskieta Mikel. 2017. Deliberation as Genre: Mapping Argumentation through Relational Discourse Structure. In *Proceedings of the 6th Workshop Recent Advances and Related Formalisms*, 1-10.

Iruskieta Mikel, Aranzabe María Jesús, Diaz de Ilarraza Arantza, Gonzalez-Dios Itziar, Lersundi Mikel, and Lopez de Lacalle Oier. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *Proceedings of IV Workshop A RST e os Estudos do Texto*, 40-49.

Iruskieta Mikel, Díaz de Ilarraza Arantza, and Lersundi Mikel. 2014. The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 466-475.

Iruskieta Mikel, da Cunha Iria, and Taboada Maite. 2015. A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2): 263-309.

Iruskieta Mikel, Labaka Gorka, and Desiderato Juliano. 2016. Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts. *Procesamiento de Lenguaje Natural*, 56: 65-72.

Li Yancui, Feng Wenhe, and Zhou Guodong. 2012. Elementary Discourse Unit in Chinese Dsicourse Structure Analysis. *Chinese Lexical Semantics*, 7717: 186-198.

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

Marcu Daniel. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3): 395-448.

O'Donnell Michael. 2000. RSTTool 2.4 − A Markup Tool For Rhetorical Structure Theory. In *Proceedings of First International Conference on Natural Language Generation (INLG'2000)*, 253-256.

Pardo Thiago Alexandre Salgueiro. 2005. *Software vai melhorar compreensão de textos em computadores*. PhD thesis. São Paulo, University of São Paulo.

Pardo Thiago Alexandre Salgueiro, Nunes Maria Maria das Graças V., and Rino Lucia H. M. 2008. Dizer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Lecture Notes in Artificial Intelligence*, 3171:224-234.

Pardo Thiago Alexandre Salgueiro, and Seno Eloize R. M. 2005. Rhetalho: um corpus dereferência anotado retoricamente. *Anais do V Encontro de Corpora*. São Carlos-SP, Brasil.

Prasad Rashmi, Dinesh Nikhil, Lee Alan, Miltsakaki Eleni, Robaldo Livio, Joshi Aravind, and Webber Bonnie. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)*, 2961-2968.

Qiu Wusong. 2010. *Jiyu xiucijiegoulilun de hanyuxinwenpinglun yupianjiegou yanjiu* (基于修辞结构理论的汉语新闻评论语篇研究 *[*Analysis of Discourse Structure in Chinese News Commentaries under Rhetorical Structure Theory*]*). Master thesis. Nanjing: Nanjing Normal University.

Rafalovitch Alexandre, and Dale Robert. 2009. United Nations general assembly resolutions: A six-languages parallel corpus, In *Proceedings of Machine Translation Summit XII*, 292-299.

Resnik Philip, Olsen Mari Broman, and Diab Mona. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2): 129-153.

Stede Manfred, and Neumann Arne. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the International Conference on Language Resources and Evaluation* (*LREC'2014*), 925-929.

Taboada Maite, and Renkema Jan. 2008. *Discourse Relations Reference Corpus* [Corpus]. Simon Fraser University and Tilburg University.

Toldova Svetlana, Pisarevskaya DIna, Ananyeva Margarita, Kobozeva Maria, Nasedkin Alexander, Nikiforova Sofia, Pavlova Irina, and Shelepov Alexey. 2017. Rhetorical relation markers in Russian RST Treebank. In *Proceedings of 6th Workshop Recent Advances in RST and Related Formalisms*, 29-33.

van Dijk Teun A. 1980. *MACROSTRUCTURES: AnInterdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. New Jersey: Lawrence Erlbaum Associations.

Wang Ling, Guang Xiang, Dyer Chris, Black Alan, and Trancoso Isabel. 2013. Mircoblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL' 2013*), 176-186.

Wu Shangyi. 2014. On Application of computer-based corpora in translation. In *Proceedings of 2nd International Conference on Computer, Electrical, and Systems Sciences, and Engineering* (*CESSE' 2014*), 173-178.

Yue Ming. 2006. Discursive Usage of Six Chinese Punctuation Marks. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, 43-48.

Zhou Lanjun, Li Binyang, Wei Zhongyu, and Wong Kam-Fai. 2014. The CUHK Discourse Treebank for Chinese: Annotating Explicit Discourse Connectives for the Chinese Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014)*, 942-949.

Zeldes Amir. 2016. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, 1-5.