Elaboration of a RST Chinese Treebank

Author:
Shuyuan Cao

Advisors:
Mikel Iruskieta (University of Basque Country: UPV-EHU)
Iria da Cunha (Universidad Nacional de Educación a Distancia)

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua

Language Analysis and Processing

January of 2018

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language, Didactics of Language and Literature and Communication, Communications Engineer.

Laburpena

Adimen Artifizialaren (AA) barneko arlo bat izanez, Hizkuntzaren Prozesamenduak (HP) giza-hizkuntzak automatikoki prozesatzea du helburu. Arlo horretako ikasketa anitzetan lorpen emankor asko eman dira. Ikasketa-arlo ezberdin horien artean, diskurtso-analisia gero eta ezagunagoa da. Diskurtsoko informazioa interes handikoa da HPko ikasketetan. Munduko hiztun gehien duen hizkuntza izanda, txinera aztertzea oso garrantzitsua da HPan egiten ari diren ikasketetarako. Hori dela eta, lan honek txinerako diskurtso-egituraz etiketaturiko zuhaitz-banku bat aurkeztea du helburu, Egitura Erretorikoaren Teoria (EET) (Mann eta Thompson, 1988) oinarrituta. Lan honetan, ikerketa-corpusa 50 testu txinatarrez osatu da, eta zuhaitz-bankua hiru etiketatze-mailatan aurkeztuko da: segmentazioa, unitate zentrala (UZ) eta diskurtso-egitura. Azkenik, corpusa webgune batean argitaratu da zuhaitz-bankua kontsultatzeko.
Gako-hitzak: HP, diskurtso-analisia, EET, txinera, corpusa

Abstract

As a subfield of Artificial Intelligence (AI), Natural Language Processing (NLP) aims to automatically process human languages. Fruitful achievements of variant studies from different research fields for NLP exist. Among these research fields, discourse analysis is becoming more and more popular. Discourse information is crucial for NLP studies. As the most spoken language in the world, Chinese occupy a very important position in NLP analysis. Therefore, this work aims to present a discourse treebank for Chinese, whose theoretical framework is Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In this work, 50 Chinese texts form the research corpus and the corpus can be consulted from the following aspects: segmentation, central unit (CU) and discourse structure. Finally, we create an open online interface for the Chinese treebank.

Key words: NLP, discourse analysis, RST, Chinese, corpus

Index

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Information and Motivation

With the development of technology, Artificial Intelligence (AI) is becoming a very popular topic in our daily life. As one of the subfields of AI, Natural Language Processing (NLP) attempts to automatically process human language. Large amount of studies contribute the development of the NLP and get fruitful achievements. Among different research fields, discourse analysis has called much attention during recent years. Discourse analysis is an unsolved problem in this field, although discourse information is crucial for many NLP tasks (Zhou et al., 2014). In particular, the relation between MT and discourse analysis has only recently begun and works addressing this topic remain limited. A shortcoming of most of the existing systems is that discourse level is not considered in the translation tasks, which therefore affects translation quality (Mayor et al., 2009; Wilks, 2009). Notwithstanding, some recent researches indicate that, discourse structure improves MT evaluation (Fomicheva et al., 2012; Tu, Zhou and Zong, 2013; Guzmán et al., 2014). Nevertheless, the MT quality from discourse level still needs to be improved, especially for Asian languages.

Among all the Asian languages, Chinese is the world's most spoken language. It is officially used in China, Singapore, Malaysia and Indonesia. The population that speaks Chinese is more than 17 billion[1]. Due the widely usage of Chinese, the Chinese occupy a very important position in NLP study.

However, until today, the NLP researches for Chinese from discourse level are still few. Therefore, it is important to carry out the research for Chinese from discourse level. Our study aims to analyze Chinese from discourse level. The study will develop a Chinese Treebank with annotated discourse information that can be applied to Chinese NLP researches.

---

[1] Quoted in the web page: Baidu baike-Zhongwen (百度百科-中文， /Baidupedia-Chinese/) [Online] `http://baike.baidu.com/view/48682.htm` [Last consulted: 10 of July, 2017]

1.2 Objectives

As previous mentioned, this work aims to develop a Chinese discourse Treebank. Specifically, we will compile a Chinese corpus with various discourse structures. We will annotate the research corpus and evaluate the annotation results to obtain a high quality annotated corpus. The annotation data will be available.

1.3 Thesis Structure

In this chapter, we have introduced the motivation, objectives and hypotheses of this work. In Chapter 2, we introduce the theoretical framework. In Chapter 3, we focus on the state of the art, where we talk about different but related works. In Chapter 4, we explain the methodology of this work, especially the corpus and the annotation process. In Chapter 5, we evaluate the annotation reliability and make a qualitative analysis for annotation disagreements. In Chapter 6, we conclude our work and look ahead of the future work.

Parts of this dissertation have appeared previously in the following peer-reviewed publications:

Corpus Compilation

- 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*. 315-324.

- Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2016. A Spanish-Chinese Parallel Corpus for Natural Language Processing Purposes. In *Proceedings of Parallel Corpora: Creation and Application International Symposium PaCor2016*. 12.

Segmentation guidelines

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

Chapter 2

Theoretical Framework

In Chapter 2, we will introduce the theoretical framework of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In the first section of this chapter, we will give a general introduction of RST. In the second section of this chapter, we will present some useful tools and problems to annotate RS-trees with Chinese texts.

2.1 Rhetorical Structure Theory (RST)

Rhetorical Structure Theory (RST) by Mann and Thompson (1988) is especially designed for discourse analysis. RST is a theory that describes text discourse structure in terms of Elementary Discourse Units (EDUs) (Marcu, 2000), and also rhetorical relations that can be held between them. EDUs can be Nuclei or Satellites (Satellites offer additional information about Nuclei), denoted by N and S. Mann and Thompson (1988) defined the first 25 relations as the original version of RST. Afterwards an extended version of the list has been provided at RST website[2].The relation can be classified into two types: Nucleus-Satellite (N-S) and Multinuclear (N-N). Table 1 shows all the original relations defined by Mann and Thompson (1988).

[2] `http://www.sfu.ca/rst/` [Last consulted: 29 of December of 2017]

Circumstance	Antithesis	Sequence
Solutionhood	Concession	Joint
Elaboration	Condition	List
Background	Otherwise	Contrast
Enablement	Interpretation	
Motivation	Evaluation	
Evidence	Restatement	
Justify	Summary	
Volitional Cause	Sequence	
Non-Volitional Cause	Contrast	
Purpose		

Table 1. Original classification of RST relations

Moreover, Mann and Thompson (1988) give another relation classification based on subject matter and presentational basis, as Table 2 shows.

Subject Matter	**Presentational**
Elaboration	Motivation
Circumstance	Antithesis
Solutionhood	Background
Volitional Cause	Enablement
Volitional Result	Evidence
Non-Volitional Cause	Justify
Non-Volitional Result	Concession
Purpose	
Condition	
Otherwise	
Interpretation	
Evaluation	
Restatement	
Summary	
Sequence	
Contrast	

Table 2. Relation classification by subject matter and presentational basis

Apart from the RST, for discourse analysis, two methods are also been widely used. One is the discourse theory Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), the other one is a corpus based approach called The Penn Discourse Treebank (PDTB) (Marcus, Santorini and Marcinkiewicz, 1993; Prasad et al., 2008).

SDRT explores the relation between discourse interpretation and discourse coherence. This theory contains several components. Firstly, it creates a language for representing the logical from of discourse and speech. A set of labels represents a discourse, each set stands for a discourse segment. Each label is linked with a representation of its content.

Likewise, the language is assigned a dynamic semantic interpretation. The interpretations of rhetorical relations (e.g. CAUSE, EXPLANATION, CONTRAST, among others) indicate additional content to that given by the lexical semantics of the expressions they connect together.

Secondly, SDRT also offers a logic named *glue logic* that computes the logic form of a discourse by compositional semantics and non-linguistic information. Every discourse segment is connected to another segment by the compositional or the lexical semantics of the expressions.

SDRT can be used to model a wide range of interactions with complex semantics and pragmatics, for instance, word sense disambiguation, questions and responses in dialogue, temporal and causal structures in text and dialogue (Asher and Lascarides, 2003).

The PDTB is a large corpus annotated with discourse structure and discourse semantics. The corpus concentrates on encoding discourse relations and the annotation methodology follows a lexically grounded approach. The following example[3] shows how the discourse relation and their arguments are annotated:

> *(Ex.) Annotation*: Michelle lives in a hotel room, and although she [**drives a canary-colored Porsche**]Arg2, [*she hasn't time to clean or repair it*.]Arg1

The above example shows an annotation of the explicit relation (CONCESSION) between Arg2 and Arg1. In its extended version PDTB 2.0, the sense annotation and the attributions associated with the relation and arguments have also been annotated.

The PDTB can be used for different NLP applications, such as parsing (Prasad, Joshi and Webber, 2010; Stepanov and Riccardi, 2014), information retrieval (Hiong,

[3] Example cited from: `https://www.seas.upenn.edu/~pdtb/index.shtml` [Last consulted: 29th of December, 2017]

Kulathuramaiyer and Labadin, 2012), machine translation (MT) (Meyer and Polakova, 2013; Li, Carpuat and Nenkove, 2014;), etc. In addition, the PDTB is also available for Chinese, and many NLP researches under PDTB have also been applied to Chinese, for instance, Chinese discourse parsing, evaluation of MT, and pos tagging.

RST has been selected as the theoretical framework of this work. Comparing to PDTB and SDRT, 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 and translation strategies (if there are) can be detected in different levels of an RS-tree (da Cunha and Iruskieta, 2010; Iruskieta, da Cunha and Taboada 2015).

2.2 RST Annotation Tools

At the moment, there are two annotation tools for RST. One is the RSTTool (O'Donnell, 2000) and another one is a new released online annotation interface named rstWeb (Zeldes, 2016). In this section, we will introduce the two annotation tools in detail.

2.2.1 RSTTool

The RSTTool[4] (O'Donnell, 2000) is an interface that allows users to annotate the discourse structure of a text in a quick and clear way. It has variant versions for different computer systems. In our work, we use the RSTTool (Mac version) to carry out the study.

The RSTTool is the first annotation interface for discourse annotation under RST. The annotation steps are: (a) segmentation and (b) discourse relations annotation. Figure 1 shows a segmented Chinese text[5] with the RSTTool and Figure 2 shows the4 completely annotated Chinese text by using the tool.

[4] RSTTool: `http://www.wagsoft.com/RSTTool/` [Last consulted: 11 of June, 2017]

[5] The text is a real example from the research corpus. An English translation is offered in the Appendix part. The translation is done by the author.

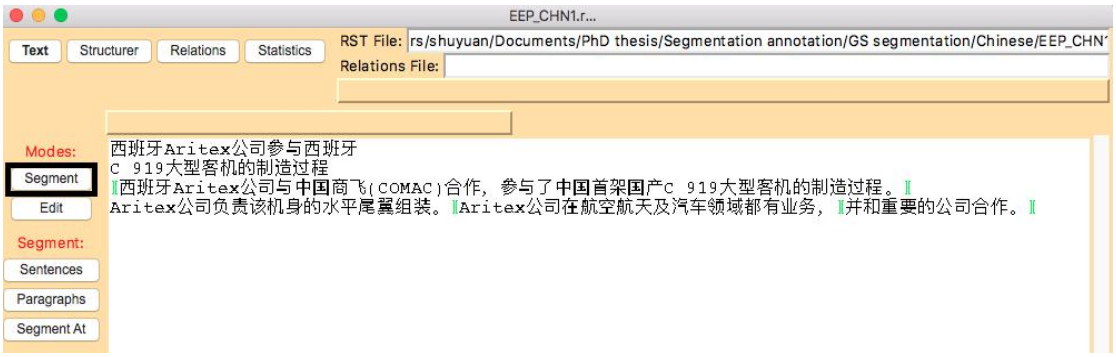

Figure 1. Text segmentation with the RSTTool

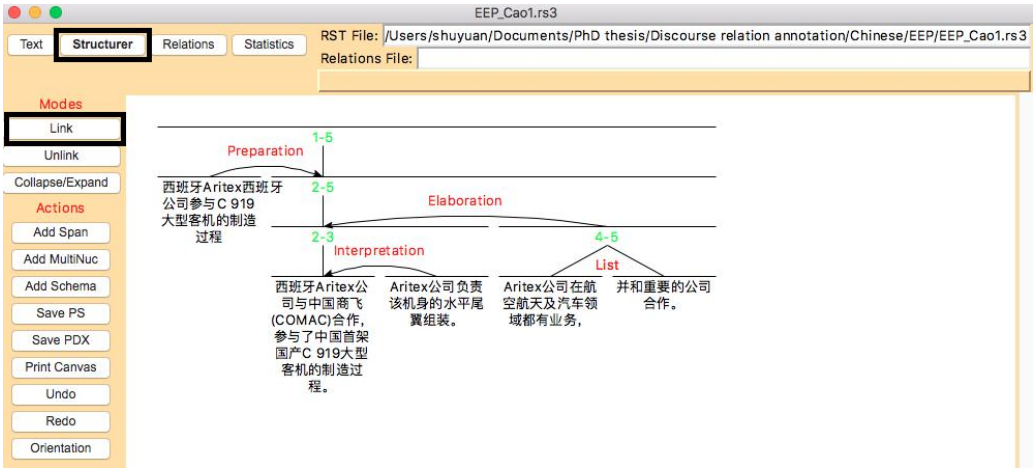

Figure 2. Discourse annotation with the RSTTool

The RSTTool saves the annotation results as XML format. Figure 3 gives the annotation result in XML format of the annotated Chinese text.

EEP_CHN1.rs3

```
<rst>
  <header>
    <encoding name="utf-8" />
    <relations>
      <rel name="circumstance" type="rst" />
      <rel name="solutionhood" type="rst" />
      <rel name="elaboration" type="rst" />
      <rel name="background" type="rst" />
      <rel name="enablement" type="rst" />
      <rel name="motivation" type="rst" />
      <rel name="means" type="rst" />
      <rel name="evidence" type="rst" />
      <rel name="justify" type="rst" />
      <rel name="cause" type="rst" />
      <rel name="result" type="rst" />
      <rel name="purpose" type="rst" />
      <rel name="antithesis" type="rst" />
      <rel name="concession" type="rst" />
      <rel name="condition" type="rst" />
      <rel name="otherwise" type="rst" />
      <rel name="interpretation" type="rst" />
      <rel name="evaluation" type="rst" />
      <rel name="restatement" type="rst" />
      <rel name="summary" type="rst" />
      <rel name="rst" type="rst" />
      <rel name="preparation" type="rst" />
      <rel name="conjunction" type="multinuc" />
      <rel name="disjunction" type="multinuc" />
      <rel name="sequence" type="multinuc" />
      <rel name="contrast" type="multinuc" />
      <rel name="same-unit" type="multinuc" />
      <rel name="list" type="multinuc" />
    </relations>
  </header>
  <body>
    <segment id="1">西班牙Aritex公司参与C 919大型客机的制造过程
</segment>
    <segment id="2">西班牙Aritex公司与中国商飞(COMAC)合作，参与了中国首架国产C 919大型客机的制造过程。</segment>
    <segment id="3">Aritex公司负责该机身的水平尾翼组装。</segment>
    <segment id="4">Aritex公司在航空航天及汽车领域都有业务，</segment>
    <segment id="5">并和重要的公司合作。</segment>
     </body>
</rst>
```

Figure3. Saved annotation result as XML format

2.2.2 rstWeb

rstWeb[6] (Zeldes, 2016) is a new released browser based interface for RST annotations. rstWeb supports for multiple annotated versions of each document, administration for user assignments, projects and guideline links. Figure 4 shows a segmented Chinese text with rstWeb and Figure 5 shows a discourse structure annotated Chinese text by using rstWeb.

[6] rstWeb: `https://corpling.uis.georgetown.edu/rstweb/info/` [Last consulted: 06 of July of 2017]

Figure 4. A segmented Chinese text by using rstWeb

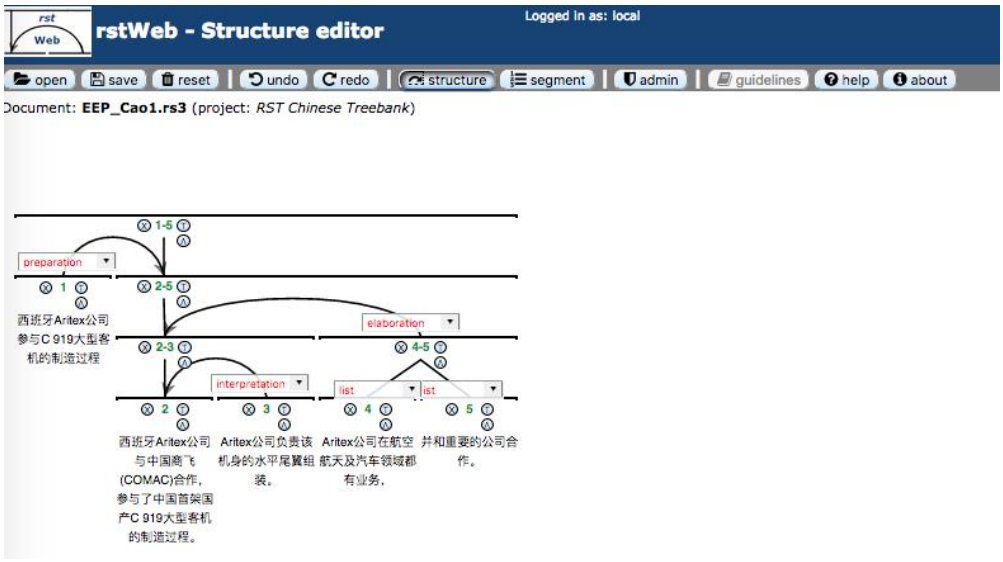

Figure 5. An annotated Chinese by using rstWeb

From Figure 5 we can see that the discourse annotation by using rstWeb is similar to the RSTTool discourse annotation result. Regarding the saved result with rstWeb, it also gives us the output as XML format. In addition, rstWeb can give an auto-fitted screenshots of analyzes. Figure 6 shows a saved annotation result in XML format with rstWeb. Figure 7 shows an auto-fitted screenshot of analyzes.

As a newly released browser based annotation interface, rstWeb has its advantages comparing to the RSTTool. For example, the RSTTool does not support all the Asian languages as indicated in its webpage while rstWeb supports all the Asian languages. Our study starts earlier than the release of rstWeb, therefore, we use the RSTTool as the annotation tool.

RST Chinese Treebank_EEP_Cao1.rs3_local.rs3

```
<rst>
        <header>
                <relations>
                        <rel name="antithesis" type="rst"/>
                        <rel name="background" type="rst"/>
                        <rel name="cause" type="rst"/>
                        <rel name="circumstance" type="rst"/>
                        <rel name="concession" type="rst"/>
                        <rel name="condition" type="rst"/>
                        <rel name="conjunction" type="multinuc"/>
                        <rel name="contrast" type="multinuc"/>
                        <rel name="disjunction" type="multinuc"/>
                        <rel name="elaboration" type="rst"/>
                        <rel name="enablement" type="rst"/>
                        <rel name="evaluation" type="rst"/>
                        <rel name="evidence" type="rst"/>
                        <rel name="interpretation" type="rst"/>
                        <rel name="justify" type="rst"/>
                        <rel name="list" type="multinuc"/>
                        <rel name="means" type="rst"/>
                        <rel name="motivation" type="rst"/>
                        <rel name="otherwise" type="rst"/>
                        <rel name="preparation" type="rst"/>
                        <rel name="purpose" type="rst"/>
                        <rel name="restatement" type="rst"/>
                        <rel name="result" type="rst"/>
                        <rel name="same-unit" type="multinuc"/>
                        <rel name="sequence" type="multinuc"/>
                        <rel name="solutionhood" type="rst"/>
                        <rel name="summary" type="rst"/>
                </relations>
        </header>
        <body>
                <segment id="1" parent="7" relname="preparation">西班牙Aritex公司参与C 919大型客机的制造过程</segment>
                <segment id="2" parent="6" relname="span">西班牙Aritex公司与中国商飞(COMAC)合作，参与了中国首架国产C 919大型客机的制造过程。</
segment>
                <segment id="3" parent="2" relname="interpretation">Aritex公司负责该机身的水平尾翼组装。</segment>
                <segment id="4" parent="9" relname="list">Aritex公司在航空航天及汽车领域都有业务，</segment>
                <segment id="5" parent="9" relname="list">并和重要的公司合作。</segment>
                <group id="6" type="span" parent="7" relname="span"/>
                <group id="7" type="span" parent="8" relname="span"/>
                <group id="8" type="span" />
                <group id="9" type="multinuc" parent="6" relname="elaboration"/>
  </body>
</rst>
```

Figure 6. Saved annotation result with rstWeb

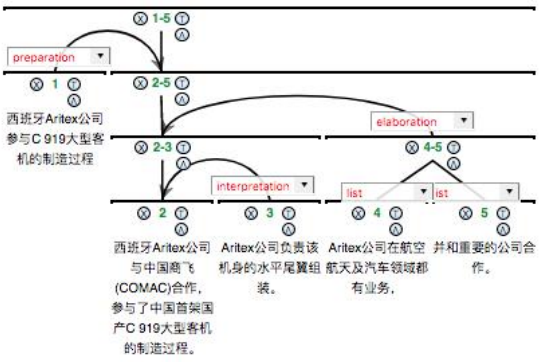

Figure 7. An auto-fitted screenshot of RST analysis by using rstWeb

2.3 RST Applications for NLP

RST has been used for several successful NLP tasks (Taboada and Mann, 2006), and especially for a large number of computational applications, including parsing, information extraction, MT, etc.

- Parsing

Parsing is the process of analysing a string of symbols and conforms to the rules of a formal grammar in NLP study. Large amounts of works address this topic with RST. Marcu (1997) uses discourse markers (DM) as relations' indications to develop an algorithm to parse the discourse structure of texts. Hanneforth, Heintze and Stede (2011) combine a surface based approach to discourse parsing with an explicit rhetorical grammar to construct an under-specified representation of possible discourse structures. Heilman and Sagae (2015) present a fast sift-reduce RST discourse segmenter and parser, which achieves near state-of-the-art accuracy and processes PDTB documents successfully. Surdeanu et al. (2015) develop two discourse parsers by using RST, one is based on the top of constituent-based syntax, and the other one uses dependency-based syntax. The first experiment exploiting different views of the data and related tasks to improve text level multilingual discourse parsing with RST is presented by Braud, Plank and Søgaard (2016). CODRA is a parser for performing rhetorical analysis in the RST framework (Joty, Carenini, and Ng, 2015). Other parsing works based on RST are the following: (i) for Spanish (da Cunha, 2016), (ii) for Catalan (da Cunha et al., 2016), (iii) for Basque (Iruskieta and Zapirain, 2015), and (iv) for Arabic (Mathkour, Touir and Al-Sanea, 2008).

- Information extraction (IE)

Information extraction (IE) is the task to automatically extract structured information from unstructured and semi-structured machine-readable documents. IE processes human language texts by means of NLP. Regarding IE and RST, Moens and de Busser (2002) propose a system for creating legal summaries by the identification of rhetorical structure in court decisions. Shinmori et al. (2002) analyze the rhetorical structure of the patent description in order to extract the claims in Japanese patents. Li (2010) presents a system that automatically extracts the rhetorical structure of a text to make the summarization under RST. The automatic summarization studies for Spanish are the works of da Cunha (2008), da Cunha, Wanner, and Cabré (2007), and Bengoetxea and Iruskieta (2018). Bengoetxea, Atutxa and Iruskieta (2017) use machine learning based approach to develop an automatic system to extract the main information of Basque scientific texts under RST. Besides, Otegi et al. (2017) design a multilingual tool (ANALHITZA) to process written texts in Basque, Spanish and English, that can be used to explore the indicators of the central unit (CU) or the signals of some discourse relations under RST or other approach.

- Machine translation (MT)

Machine translation (MT) explores the use of software to translate text or speech from one language to another. Fomicheva, da Cunha and Sierra (2012) evaluate the MT between Spanish and English under RST. They use a Spanish-English corpus to evaluate two MT systems via the discourse strategies. Guzmán et al. (2014) carry out a similar work using RST as the framework, comparing the output of MT and a human reference. Tu, Zhou and Zong (2013) present a RST-based translation framework for modelling semantic structures in translation model, so as to maintain the semantically functional integrity and hierarchical relations of EDUs during translating (Iruskieta, da Cunha and Taboada, 2015).

Chapter 3

State of the Art

RST has been applied to different languages. In Chapter 3, we will present detailed information of RST-based treebanks for different languages, for instance, what are the original sources of the corpus, the statistic information about the corpus and the topic of the corpus. Then, we will discuss different but related researches about Chinese discourse analysis.

3.1 RST based Treebanks

Several corpora for different languages have been annotated under RST. Authors of these corpora have established their own segmentation criteria for different discourse analysis tasks.

(i) English

The best known-annotated RST corpus for English is the RST Discourse Treebank (Carlson, Marcu and Okurowski, 2001)[7]. Totally 385 texts are selected, the texts which are are journalistic texts. The topics of the texts are culture, economy and editorials among others.

The Discourse Relations Reference Corpus (Taboada and Renkema, 2008)[8] is another RST Treebank for English. This corpus contains 65 texts. The genres of the texts are journal articles, advocacy letter and review texts. The topics of the corpus are economy, language, social service among others.

(ii) German

The corpus for German by using the RST is The Potsdam Commentary Corpus (Stede and Neumann, 2014)[9]. The corpus includes 220 German newspaper

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

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

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

commentaries with topic of politics. The corpus is extracted from the online newspaper *Märkische Allgemeine Zeitung* and *Tagesspiegel* and contains 44,000 words.

(iii) Spanish

The corpus annotated in the RST framework for Spanish is The RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011) [10]. The corpus contains 267 texts and 52,746 words. The texts in this corpus are only specialized texts, such as scientific articles, conference papers, articles and reports in magazines. The texts have been divided into 9 domains: astrophysics, earthquakes, engineering, economy, linguistics, medicine, psychological and sexuality.

(iv) Basque

The RST discourse analysis for Basque is presented in The RST Basque Treebank (Iruskieta et al., 2013)[11]. This corpus is a public corpus that can be used for Basque NLP tasks. It includes abstracts from three specialized domains: medicine, terminology and science and literary reviews. 88 documents have been selected in the corpus.

(v) Portuguese

Two annotated corpus by using RST exist for Portuguese: The CorpusTCC (Pardo, Nunes and Rino, 2008) and Rhetalho (Pardo and Seno, 2005)[12]. The CorpusTCC is built for the detection of linguistic patterns and indication of rhetorical relations. This corpus contains 100 Brazilian Portuguese scientific texts, including 53,000 words. Rhetalho is a corpus designed for parser evaluation and it consists of 40 texts, 20 from computer science domain and 20 from the online newspaper Folha de São Paulo. The total words of the corpus are around 5,000 words.

(vi) Russian

The Russian RST Treebank[13] is designed for the Russian discourse analysis by Toldova et al. (2017). The corpus aims to annotate texts of four genres and domains: science, popular science, news stories, and analytic journalism. Currently, 73 annotated

---

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

[11] `http://ixa2.si.ehu.es/diskurtsoa/en/` [Last consulted: 10 of January of 2018]

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

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

texts are included in the corpus, most of the annotated texts are news stories. 44,685 tokens are included in the already annotated 73 texts.

(vii) Basque and Spanish

The RST Basque-Spanish DELIB Treebank (Imaz and Iruskieta, 2017)[14] is an annotated bilingual RST corpus for Basque and Spanish. The corpus is an extended version of the RST Basque Treebank. 100 texts in Basque and other 100 Spanish texts are included in this corpus. The corpus involves 8,900 words for the Basque subcorpus and 11,166 words for the Spanish subcorpus.

(viii) English, Spanish and Basque

The trilingual RST corpus is The Multilingual RST Treebank (Iruskieta, da Cunha and Taboada, 2015)[15]. The parallel corpus includes 45 texts (15 texts for each language), the English subcorpus contains 5,706 words, the Spanish subcorpus contains 6,324 words and the Basque subcorpus contains 4,800 words. The main topic of this corpus is terminology research.

3.2 Corpus based Chinese Discourse Analysis

The earlier Chinese discourse analysis is the Penn Chinese Discourse Treebank (CDTB) (Xue, Xia, Chiou, and Palmer, 2006), which follows The Penn Discourse Treebank (PDTB) (Marcus, Santorini and Marcinkiewicz, 1993; Prasad et al., 2008) annotation criteria. This corpus contains CTB-I and CTB-II[16]. The corpus can be used for different NLP tasks, such as word segmentation information, part-of-speech (POS) information, parsing information, and grammar extraction. Currently, the corpus is partly accessible. The texts of this corpus are mainly taken from *XinHua* newswire articles, Hongkong News and *Sinorama*[17]. The topics of the corpus are various, such as general politics, culture, economy, travel, etc.

---

[14] `http://ixa2.si.ehu.es/diskurtsoa/rstfilo/index.php` [Last consulted: 01 of December of 2017]

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

[16] Due to the statement of authors, CTB-I is released by LDC as Chinese Treebank Versions 1.0 and 2.0. CTB-II is included in Chinese Treebank Version 3.0. In 2013, they publish the 8th version and name it as The Chinese Discourse Treebank. More information can be consulted: `https://catalog.ldc.upenn.edu/LDC2013T21` [Last consulted: 06 of July of 2017].

[17] Sinorama is a magazine from China Taiwan Province.

The Sinica Treebank[18] is created by Huang et al. (2014). Its first version was released in 1997. Currently, the Sinica Treebank has its third version and includes 61,087 trees (361,834 words). There are 1,000 tree structures open to the public for academic research. This corpus has been tokenized and offers word segmentation information, POS information, syntax information, and semantic information. The Sinica Treebank uses the texts from Sinica Corpus (Chen et al., 1996) and the topics of the texts are different, for instance, politics, travelling, sports, society, etc.

The Discourse Treebank for Chinese[19] is another project for Chinese discourse analysis and was created by Zhou et al (2014). They annotated explicit intra-sentence discourse connectives, their corresponding arguments and senses for all 890 documents of the Chinese Treebank 5, by adopting the annotation scheme of PDTB.

Regarding RST based Chinese discourse treebank, there are two related works so far. Yue (2006) creates the Caijingpinglun Corpus (CJPL) under RST. The CJPL corpus contains 40,000 Chinese financial news commentaries, and about 80 million words. Yue (2006) annotates relations between sentences (inter-sentence) and within a sentence (intra-sentence) to analyze the Chinese rhetorical structure. Qiu (2010) annotates 10 Chinese news commentaries under RST to explore the characters of Chinese discourse structure. The corpus contains 12,538 words. However, some limitations exist for the two works. Firstly, none of the works is available to the public[20]. Secondly, for both corpora, the single source cannot guarantee the discourse diversity of the texts. The genre and the topic of the texts in both corpora are simple. The three aspects affect the quality of the discourse structure. A corpus with a high quality for discourse analysis requires texts of different topics and genres from different sources. Thirdly, authorization of texts. The authors donot mention if they have permission to use the texts for their studies. Fourthly, few texts have been annotated for Chinese discourse analysis. Although the corpus of Yue (2006) selects 40,000 Chinese financial news commentaries, the author only annotates 90 commentaries. The corpus of Qiu (2010)

---

[18] Sinica Treebank: `http://rocling.iis.sinica.edu.tw/CKIP/engversion/treebank.htm` [Last consulted: 06 of July of 2017]

[19] Though Zhou et al. (2014) declare that their Treebank is open to the public in their paper; we did not find it after searching in the Internet. We wrote to them requesting the related information, but they have not sent a response.

[20] The work of Yue (2006), we wrote to her requesting the related information, but she has not sent a response. For the work of Qiu (2010), we cannot find the contact information, neither the information of the supervisor.

only contains 10 annotated texts. Lastly, none of the works mentions the evaluation of the annotation quality, they do not relate any inter annotator agreement..

Chapter 4

Methodology

In Chapter 4, we will explain how we carry out the study. In the first section of the chapter (Section 4.1), we will focus on the construction of the research corpus. We will talk about the considered characteristics for the development corpus, the statistic information of the corpus, the applications of the corpus, etc. Secondly, each research step will be introduced in the following sections:

Discourse segmentation. Segmentation is an important step for NLP study, including RST studies, because it can affect the discourse annotation. The section 4.2 will introduce how we elaborate our segmentation criteria.

Central Unit (CU) annotation. Central Unit is the key information of a text, studies of CU benefits the RST annotation tasks. The section 4.3 will describe the methodology of the CU annotation.

Discourse structure. Discourse relation annotation is the most important step under RST, it can reflect the all the discourse information (order of EDUs, signals, discourse relations, etc.) of a text. The section 4.4 will explain in detail the discourse relations' annotation.

The section 4.5 will conclude this chapter and look ahead of the following chapter.

4.1 Corpus Compilation

The corpus compilation is one of the fundamental research steps for this study. Complexity of discourse structure and heterogeneity are the main characteristics taken into account for the 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 have selected 50 Chinese texts to form our research corpus. The genres of the texts are: (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. Table 3 shows the genres and sources of the corpus.

The corpus includes texts related to 7 domains: (a) terminology (15 texts), (b) culture (6 texts), (c) language (8 texts), (d) economy (7 texts), (e) education (4 texts), (f) art (5 texts), and (g) international affairs (5 texts). Table 3 is the conclusion of the corpus genre and source information.

Genre	**Texts**	**Original publication**
Abstract of research paper	15	International Conference about Terminology (1997)
News	15	Shanghai Miguel Cervantes Library, Chamber of Commerce and Investment of China in Spain, Spain Embassy in Beijing, Confucius Institute Foundation in Barcelona
Advertisement	13	Shanghai Miguel Cervantes Library, Spain-China Council Foundation, Beijing Cervantes Institute, Granada Confucius Institute
Announcement	7	Spain Embassy in Beijing, Confucius Institute Foundation in Barcelona, Beijing Cervantes Institute
Total	50	

Table 3. Corpus source and genre information

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

Finally, we have created an online interface to access the corpus: `http://ixa2.si.ehu.es/rst/zh/`. Moreover, users can also download the texts of the corpus. Figure 8 is a screenshot of the webpage of the corpus.

Figure 8. The website of the RST Chinese Treebank

4.2 Discourse Segmentation

Segmentation is a crucial step of discourse analysis, since it can affect the result of the relational discourse structure. Moreover, discourse segmentation can be useful for different NLP tasks, for instance, the evaluation of automatic segmentation systems, and the development of discourse parsers and automatic summarizers. The segmentation tool that we use in this work is the RSTTool (O'Donnell, 2000). An entire text can be divided into various independent EDUs once the segmentation step is finished with the RSTTool. Figure 9 includes an example of a segmented text from the corpus.

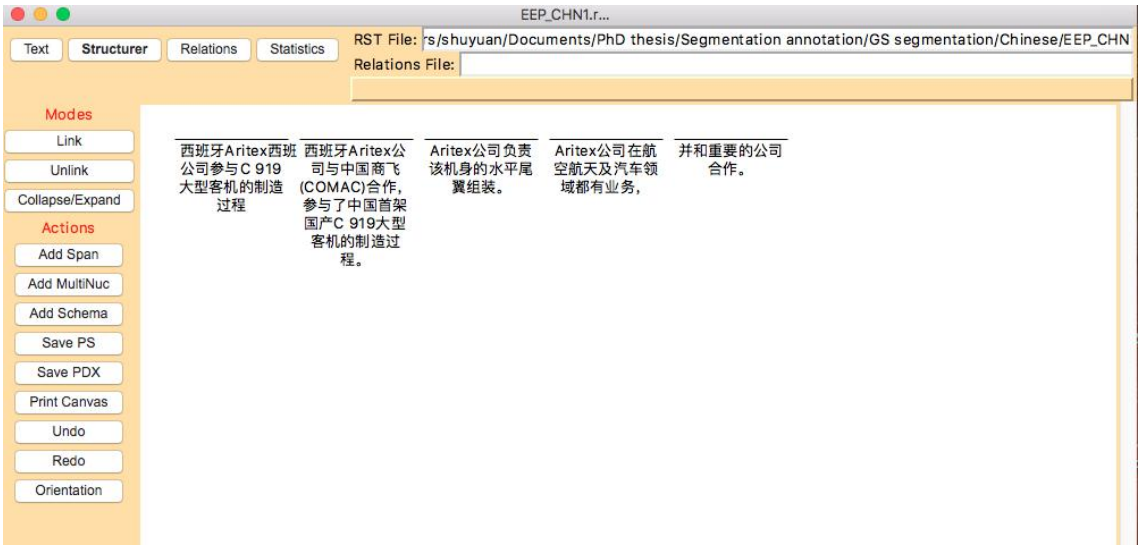

Figure 9. Independent EDUs of a segmentated text with RSTTool

First of all, we elaborate a preliminary discourse segmentation criteria proposal for 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).

The following segmentation criteria are used in our work:

- Paragraphs and line breaks. In our study, a line break will be taken as an independent EDU to segment the titles (and subtitles).

  (Ex.1) Text name: FCEC1

  Text: [亲爱的朋友们，] [...]

  English: [Dear friends,] [...][21]

  Explanation: The Chinese passage starts with a greeting, it is followed by a comma and there is a line break.

- Sentences and periods. In our study, the period marks the end of an independent EDU.

  (Ex.2) Text name: ICP4

  Text: [塞万提斯学院正式教师职位招聘在西班牙媒体上公布。] [同时也在塞万提斯学院网站发布信息。

  English: [Cervantes Institute official professor recruitment notice publishes on Spanish media.] [Meanwhile, also publishes on the Cervantes Institute webpage.]

  Explanation: After the word "*gongbu*" (公布) ('publish'), there is a period, followed by another sentence.

- Question mark and exclamation mark. Both marks are signals of a sentence boundary.

  (Ex.3) Text name: TERM34

  Text: [区分界限在哪里？] [区分表语及非表语的关键在哪里？] [涉及文字关系、背景联系、物主关系还是其它方面？]

  English: [Distinguish boundary in where?] [Distinguish predicative and non-predicative of key in where?] [About characters relation, background relation, possessive relation or other aspect?]

[21] In this work, we give an English literal translation for all examples in Chinese to let the readers understand.

Explanation: At the end of each sentence, there is a question mark.

- Other EDUs should have a main verb or an adjunct verb phrase[22]. This is a basic segmentation criterion and segmentation criteria below should follow this rule.

(Ex.4) Text name: CCICE3

Text: [10 月份，西班牙财政部共**筹集** 143.99 亿欧元，共拍卖国债四次。]

English: [The month of October the Treasury raised 14.399 million in four issues.]

Explanation: The Chinese word "*chouji*" (筹集) is a verb and means 'raise' in English.

- Discourse Marker (DM)[23], verb and comma. If there is a DM at the beginning of a sentence and, this sentence is divided into two parts by a comma (each one including a verb), both parts are considered independent EDUs.

(Ex.5) Text name: TERM31

Text: [**由于**经常使用词法句型模式，] [用以分析文本或者至少说明性略语较为合适。]

English: [Due to often uses morph-syntax models,] [to analyze texts or at least illustrative abbreviations.]

Explanation: The Chinese DM "*youyu*" (由于) ('due to') is placed at the beginning of the first EDU, and a comma is included in the sentence. Besides, the first EDU includes the Chinese verb "*shiyong*" (使用) ('use'), while the second EDU includes the verb "*fenxi*" (分析) ('analyze').

(Ex.6) Text name: TERM19

Text: [此时，标准不但会失效，] [**而且**也不能发挥作用。]

---

[22] In RST clauses (adverbial clauses) are considered EDUs, except for complement clauses (Mann and Thompson, 1988).

[23] Schiffrin (2001: 54) indicates: "Discourse markers (DMs) involve linguistic items that in cognitive, expressive, social and textual domains." Also, Portolés (2001) explains that DMs are invariable linguistic units that depend on the following aspects: (a) distinct morpho-syntactic properties, (b) semantics and pragmatics and (c) inferences made in the communication. Eckle-Kohler, Kluge and Gurevych (2015) give a more specific definition of DMs from the textual level that DMs are used to signal discourse relations in a text segment. In our study, we follow the definition of Eckle-Kohler, Kluge and Gurevych (2015).

English: [In this condition, standardization not only ceases to be effective,] [but also could not play its role.]

Explanation: The Chinese DM "*er*" (而且) ('but also') appears after a comma in the sentence. In addition, verbs are included in both EDUs: "*shixiao*" (失效) ('lose effectiveness') in the first EDU, and "*fahui*" (发挥) ('exert') in the second EDU.

- Semicolon plus adjunct verb phrase.

(Ex.7) Text name: TERM34

Text: [例如，形容词 marginal（边上的）在英语中可用于参照语和谓语，例如"边缘注释(marginal not)"以及"边缘个案(marginal case)"；] [相反，在"名词非表语性形容词"一类中，尽管采用了形容词的定义，但是与名词发挥的作用类似，比如：linguistic difficulties（语言上的困难）/language difficulties（语言困难）。]

English: [For example, adjective marginal (something besides) in English can be used referential and predicate, for example, "marginal note" and "marginal case";] [in contrast, in "noun but not predicative adjective" category, although adapts adjective definition, with noun works function similar, such as, linguistic difficulties/language difficulties]

Explanation: A semicolon separates the text into two parts, and each EDU includes a Chinese verb: the verb "*yong*" (用) ('apply to') in the first EDU and the verb "*shiyong*" (使用) ('use') in the second EDU.

- Parenthetical and dash. Only when a parenthetical unit does not modify a noun neither an adjective and it includes a verb, it is an independent segment; if within the parenthetical unit there are coordinated parts, the coordinated parts are also segmented[24].

(Ex.8) Text name: TERM18

Text: [确实，术语数据库的设计和管理无论在理论和方法论] [**(**如何表示一个术语？] [有最简单的表达方法吗？] [术语之间如何分类？**)**] […]

[24] This criterion only exists in our work; the mentioned Chinese segmentation works have overlooked this segmentation criterion.

English: [Indeed, the design and management of the terminology database no matter in theory and methodology,] [(how to express a terminology?] [is there the easiest way to express?] [how to distinguish among terminologies?]) […]

Explanation: The parenthetical unit does not modify its previous part; it should be an independent segment. The sentences "*ruhe biaoshi yige shuyu*?" (如何表示一个术语？) (How to express a term?), "*you zuijiandan de fangfa ma*?" (有最简单的方法吗？) (Is there the easiest way to express?) and "*shuyu zhijian ruhe fenlei*?" (术语之间如何分类？) (How to distinguish among terminologies?) include a verb and are coordinated parts in this parenthetical unit with verbs and question marks.

- Coordination and ellipsis with verbs. Coordinated clauses with verbs are considered independent EDUs (even they include a null subject).

(Ex.9) Text name: TERM25

Text: […] [自 1994 年以来**我们**在德武斯特大学进行法律领域专业文件的翻译工作，] [**我们**希望能按照实际情况呈现出这些年工作中碰到的问题以及取得的成就。] […]

English: [From 1994 until now we in Deusto University carry out law campus professional document of translation works,] [we hope can follow real situation present these years works encounter problems and achievement] […]

Explanation: In the Chinese text, the two coordinated clauses include verbs ("*jinxing*" [进行] ['to carry out'] and "*xiwang*" [希望] ['hope']).

- Relative, modifying and appositive clauses. Relative clauses, clauses that modifies a noun or adjective or appositive clauses are not considered independent EDUs.

(Ex.10) Text name: BMCS5

Text: [比如我们在很多网站上都能找到有关网络属于词汇的文章，**上面会提出一些命名建议**。]

English: [For example, we in many websites could find about Internet terminology of articles, where gives some dominate suggestions.]

Explanation: The part after comma is "*shangmian hui tichu yixie mingming jianyi*" (上面会提出一些命名建议) refers to those articles in the websites, which is related with the part before comma.

- Reported speech. In this study, we do not consider reported speech as an independent EDU.

(Ex.11) Text name: TERM29

Text: [据西班牙财政部在官网发布的消息显示，该机构将在本周二拍卖 6 至 12 月到期的短期国债，] [预期拍卖 40 亿至 50 亿欧元。]

English: [According to Spanish Ministry of Finance on official website of the agency publish the notice shows, the agency will on this Tuesday be auctioned from June to December short-term treasury bonds,] [expected auction 4 billion to 5 billion euros.]

Explanation: In the Chinese message, the content *gaijigou jiangzai benzhouer paimai 6 zhi 12 yue daoqi de duanqi guozhai* (该机构将在本周二拍卖 6 至 12 月到期的短期国债) and the content *yuqi pamai 40yi zhi 50yi ouyuan* (预期拍卖 40 亿至 50 亿欧元) are reported speech of their previous part, which is *genju xibanya caizhengbu guanwang xianshi* (根据西班牙财政部官网显示) ('According to the Spanish Ministry of Finance office website shows'). In this case, the attributed parts are not considered as the independent EDUs, we only segment within the attributed parts.

- Truncated EDUs. For the cases of truncated EDUs, we use the non-relation label of Same-unit (Carlson, Marcu and Okurowski, 2003) (see Figure 10).

(Ex.12) Text name: CCICE3

Text: [确实，术语数据库的设计和管理无论在理论和方法论] [(如何表示一个术语？] [有最简单的表达方法吗？] [术语之间如何分类？)] [乃至信息学范围内都带来了种种疑问] [(术语数据库应采用哪种结构？] [应考虑到哪些联系？] [字典应统一成什么样？)。]

English: [Indeed, the design and management of the terminology database no matter in theory and methodology,] [(how to express a terminology?] [is there the easiest

way to express?] [how to distinguish among terminologies?)] [and even information scope within brings all kinds of questions] [(Term database should adopt which kind structure?] [Should consider which relations?] [Dictionary should be unified as what?)]

Explanation: The Chinese text shows, the EDU(5-8) and the EDU(9-12) consist of a complete sentence. Meanwhile, the EDU(6-8) and the EDU(10-12) are the inserted parts of the Chinese sentence.

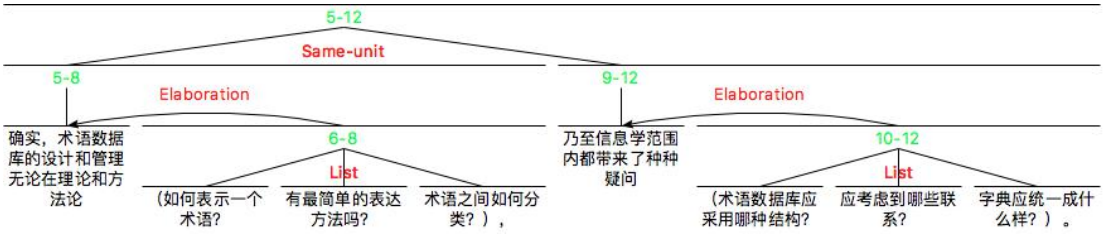

Figure 10. Case of *Same-unit* in the corpus

The segmentation information is available in the website. All the segmented texts can be consulted from there, as Figure 11 presents.

HOME | RELATIONS | RELATIONS IN TREES | EDUS | SEARCH | SEARCH SPANISH | REFERENCES | PRIVATE

ICP_CHN1-GS.rs3 (13)

EDU	Segment	Tagger	Central Unit
1	学院介绍	GS	
2	塞万提斯学院创建于1991年，	GS	
3	旨在推动西班牙语教学、传播西班牙及其他西班牙语国家的文化。	GS	
4	塞万提斯学院的总部设在马德里及西班牙著名作家米盖尔·塞万提斯的故乡阿尔卡拉·德·埃纳雷斯（马德里大区）。	GS	
5	塞万提斯学院目前在世界四大洲拥有70多所分院，北京塞万提斯学院是这些学院中最重要的之一，它是中国的第一所分院。	GS	
6	另外，塞万提斯学院还负责：	GS	
7	• 组织西班牙语水平认证考试（DELE），对学员颁发官方学位证书、证明	GS	
8	• 开设西班牙语课程	GS	
9	• 开设对西班牙语教师的培训课程	GS	
10	• 为西班牙语语言文学研究者的研究活动提供支持	GS	
11	• 与其他机构合作组织文化活动	GS	
12	塞万提斯学院由西班牙及西班牙语美洲的学术界、文化界、文学界的代表人物领导工作。	GS	
13	北京塞万提斯学院与艺术馆、画廊、出版社及中国、西班牙、拉丁美洲其他各文化机构合作组织各类文化活动。	GS	

Figure 11. A segmented text in the website

4.3 Central Unit

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 (Cao, da Cunha, and Iruskieta,

2016). CU can be applied to different NLP studies, for example, automatic summarization, development of intelligent systems (Iruskieta, Labaka and Desiderato, 2016) and sentiment analysis (Alkorta, Gojenola, Iruskieta and Pérez, 2015). Genre, domain and discourse structure determine the position of the CU in a text; thus, by consulting the CU of the texts in the corpus, users can know how to organize the information of texts in different genres and domains. A good translation of the main topic or CU is also fundamental for a MT system.

The studies on CU within RST are the following: i) Iruskieta et al. (2013) annotated and harmonized manually some CUs of the RST Basque Treebank, ii) Iruskieta, Díaz de Ilarraza and Lersundi (2014) analyze indicators (nouns, verbs and other word categories) that indicate the CU of a rhetorical structure and show the correlation in the agreement of the CU with the agreement in rhetorical relations. Afterwards, CU automatic annotation system is created: iii) a rule-based system is created to detect the CUs for Basque scientific abstracts (Iruskieta et al., 2015) and also for Brazilian Portuguese and Basque texts (Iruskieta, Labaka and Antonio, 2016) and iv) a machine learning based CU annotation system (Bengoetxea, Atutxa and Iruskieta, 2017)[25].

According to van Dijk (1980), language users are able to summarize discourses, expressing the main topics of the summarized discourse. In our study, based on the natural human understanding, we annotate the CU of each text of the corpus. Moreover, we have extracted all the possible words that can represent the CU.

Figure 12 presents the CU of the annotated Chinese text in the corpus.

[25] The system can be consulted at: `http://ixa2.si.ehu.es/CU-detector`. [Last consulted: 16th of January, 2018]

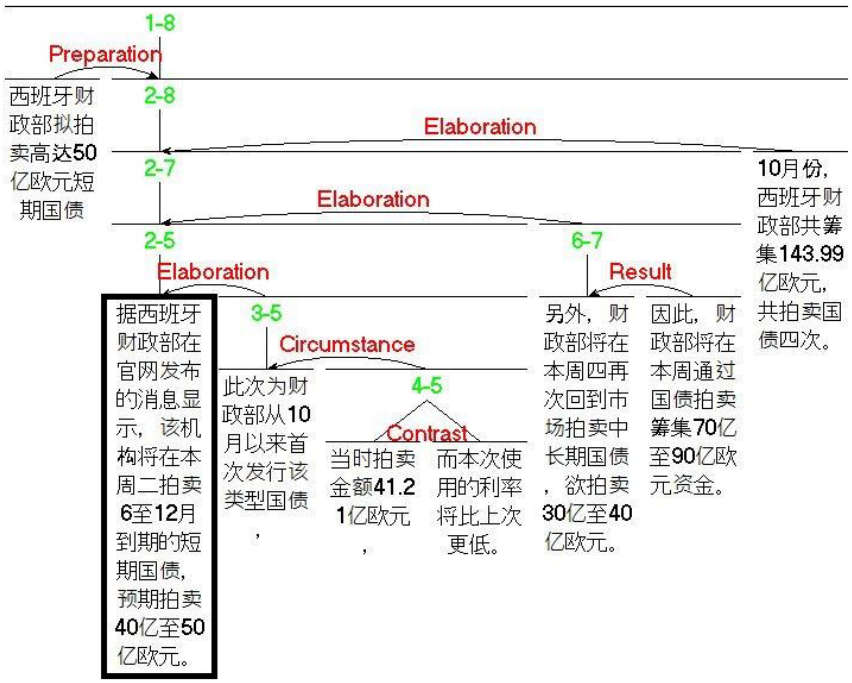

Figure 12. CU of the annotated text (CCICE3_CHN)

In Figure 12, for the Chinese text, all the arrows are also point to the EDU2. Therefore, the main idea in the Chinese text is "*ju xibanya caizhengbu zai guanwang xianshi, gai jigou jiang zai benzhouer paimai 6 zhi 12 yue daoqide duanqiguozhai, yuqi paimai 40 yi zhi 50 yi ouyuan* (据西班牙财政部在官网发布的消息显示，该机构将在本周二拍卖 6 至 12 月到期的短期国债，预期拍卖 40 亿至 50 亿欧元。)[26]".

For the CU annotation, we follow the method proposed by Iruskieta (2015), which is especially designed for the CU annotation under RST. Firstly, we confirm the topic of the text statement. Secondly, we find the purpose of the text. Thirdly, we explore the method mentioned in the text. Fourthly, we find the results of the text. Lastly, we check the conclusion of the text.

4.4 Discourse Structure

Discourse structure annotation is one of the most difficult challenges for annotation works (Hovy and Lavid, 2010). In this study, we use intra-sentence annotation style and inter-sentence annotation style. Intra-sentence annotation means that we annotate the discourse relation within a segmented sentence. Inter-sentence annotation means that the discourse relations will be defined between the sentences. For the text annotation,

[26] English literal translation: According to Spanish Ministry of Finance on official website of the agency publish the notice shows, the agency will on this Tuesday be auctioned from June to December short-term treasury bonds, expected auction 4 billion to 5 billion euros.

we follow the annotation guideline proposed by Pardo (2005), firstly we annotate the relations within the segmented sentences; secondly, we identify the relations between the sentences within a paragraph. Lastly, we find the relations between paragraphs. We use RSTTool to finish the discourse annotation task. Figure 13 shows an annotated Chinese text with RSTTool.

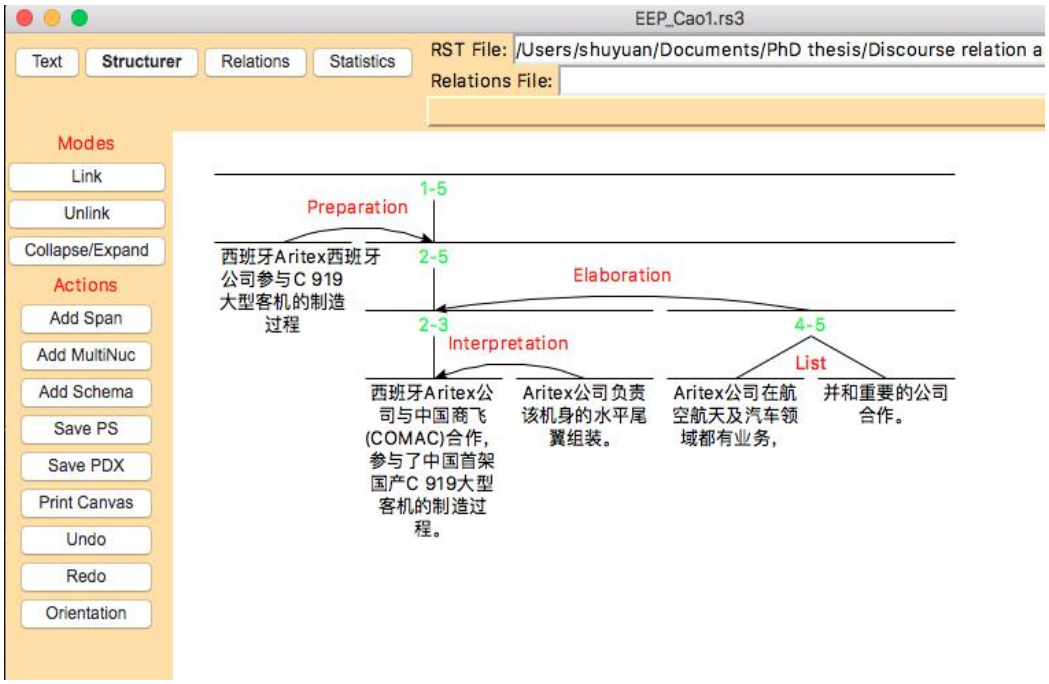

Figure 13. An annotated text with discourse relations under RSTTool

In Figure 13, we can see that the text is being annotated by intra-sentence annotation style and inter-sentence annotation style. EDU2, EDU3 and EDU(4-5) are three independent sentences, three discourse relations have been defined for the EDUs. EDU2 and EDU3 hold a INTERPRETATION relation, and EDU(2-3) and EDU(4-5) contain a ELABORATION relation. EDU4 and EDU5 are two parts of a complete sentence, the relation that included between EDU4 and EDU5 is LIST.

The discourse relations that we use are in the following table (Table 4). Totally, 26 relations have been selected in this study. The 21 relations are N-S relations, and the other 5 relations are N-N relations. The used relations are presented in the RST webpage[27].

[27] `http://www.sfu.ca/rst/01intro/intro.html` [Last consulted: 29 of December of 2017]

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

Table 4. Selected discourse relations for annotation

Here for each selected relation, we will give an example from the corpus and its English literal translation[28].

Category: N-S

(1) ANTITHESIS

Nuclear: The author favours the idea.

Satellite: The author disfavours the idea.

Text Name: TERM30

Chinese: [当今术语管理系统的使用减轻了专业术语因归档带来的收存、文件修复以及展示的工作负担。]$_S$ [但不管何种情况下，术语汇集、分析和确认工作都由专业人员担任。]$_N$

English: [Currently terminology management system use mitigates professional terminology brings archiving, file repair, and presentation burdens.]$_S$ [However, in any case, terminology collection, analysis and validation work by professional staff serve.]$_N$

(2) BACKGROUND

Nuclear: The understanding has already been inserted in the text.

[28] The explanations of the discourse relations are extracted from RST webpage, but all the examples are from the research corpus. In addition, to show some inter-sentence relations, the segmentation may not follow the segmentation criteria.

Satellite: Text for getting the understanding.

Text Name: TERM23

Chinese: [在爱尔兰语术语委员会(An Coiste Téarmaíochta)的协助下，Fiontar 和 VOCALL 正逐渐满足大学教育和职业培训中建立爱尔兰语术语库的需求。]$_S$ [此次报告将研究这两个组织在创建新术语过程中使用的方法，]$_N$

English: [With An Coiste Téarmaíochta help, Fiontar and VOCALL are increasingly meeting university education and vocational training establish Irish termbase needs.]$_S$ [The report will examine these two organizations during the establishment new terms process of usage methods.]$_N$

(3) CAUSE

Nuclear: A situation.

Satellite: Another situation that causes that one.

Text Name: ICP7

Chinese: [由于 MOPAC 的支持，]$_S$ [塞万提斯学院的图书馆网络(RBIC)得以增加这一有关信息移动服务的功能。]$_N$

English: [Due to MOPAC support,]$_S$ [Cervantes Institute of library network (RBIC) is able to increase information mobile service of functionality.]$_N$

(4) CIRCUMSTANCE

Nuclear: Text shows the ideas or the events that occur in the interpretive text.

Satellite: An interpretive context of situation or time.

Text Name: TERM51

Chinese: [在详细描述各个具体要素之前，]$_S$ [我们需首先确认与地理术语相关的地名的概念。]$_N$

English: [Before in detail describing each element,]$_S$ [we need to firstly confirm geographical terms related to place names of concept.]$_N$

(5) CONCESSION

Nuclear: A situation confirmed by the author.

Satellite: Another situation inconsistent but also affirmed by the author.

Text Name: TERM34

Chinese: [在很多情况下，要找到巴斯克语对应临近语中的关系形容词，需要经过多个步骤 (Ensunza, 1989; Loinaz, 1995)。]$_N$ [尽管如此，从某种程度上来说，

选择何种办法很大程度上仍取决于作者或者译者良好的判断力、直觉以及审美。]S

English: [In many cases, to find the Basque language corresponded related language of relational adjective, it requires several steps (Ensunza, 1989; Loinaz, 1995).]N [Although, to some extent, the choice of approach still largely depends on the good judgment, intuition and aesthetic of the author or translator.]S

(6) CONDITION

Nuclear: Action or situation whose occurrence results from the occurrence of the conditioning situation.

Satellite: A condition situation.

Text Name: BMCS3

Chinese: [若您希望进行全面集中的语言学习或者您希望短时间内提高您的语言水平，]S [紧凑课程是一个很好的选择。]N

English: [If you want completely focus on language study or you want in short time improve your language level,]S [intensive course is a good choice.]N

(7) ELABORATION

Nuclear: The basic information.

Satellite: An additional information of the basic information.

Text Name: EEP7

Chinese: [9 月 18 日，由西班牙驻华大使馆和卢米埃影城主办的第六届西班牙电影节在北京侨福芳草地开幕。]N [电影节持续至 23 日，展映了 7 部西班牙最近几年出品的精彩影片。]S

English: [September 18th, hosted by Spanish Ambassador and Lumiere Studios the 6th Spanish Film Festival in Beijing Parkview Green Center.]N [The film festival lasted until 23rd, showing 7 Spanish recent years' best films.]S

(8) ENABLEMENT

Nuclear: An action.

Satellite: The information aims to perform the action.

Text Name: BMCS4

Chinese: [在课程中，你将熟悉考试中的各项内容，集中复习听说读写四个部分。]N [根据塞万提斯学院教学大纲，有针对性地复习各级别对应的语法和词汇内容。]S

English: [In course, you will be familiar with the exam of contents, focus on listening, speaking, reading and writing four parts.]$_N$ [According to Cervantes Institute syllabus, targeted review of each level corresponded grammar and vocabulary content.]$_S$

(9) EVIDENCE

Nuclear: A claim.

Satellite: Information that increases the reader's belief in the claim.

Text Name: TERM34

Chinese: [在任何情况下，各种语言中都有形容词可以涵盖上述三种类别。]$_N$ [为了更好地说明，我们将使用利维（Levi）的例子：musical voice（音乐般的声音）、musical criticism（音乐评论）、musical comedy（音乐剧）。]$_S$

English: [In any case, in various languages have adjectives can cover the above three categories.]$_N$ [To better explain, we will use Levi's example: musical voice, musical criticism, musical comedy.]$_S$

(10) EVALUATION

Nuclear: A situation.

Satellite: An evaluative comment about this situation.

Text Name: FCEC1

Chinese: [目前中国已经成为世界第二大经济强国，]$_N$ [或许是世界上经济最具活力的国家。]$_S$

English: [Currently China has become the world's second largest economy country,]$_N$ [perhaps is the world's most dynamic economy country.]$_S$

(11)INTERPRETATION

Nuclear: A situation.

Satellite: An interpretation of the situation.

Text Name: TERM30

Chinese: [从某种程度上来说，撰写专业的书面内容与术语的“技术生产”紧密相连，]$_N$ [这是一种与机器运作原理一致的话语模式。]$_S$

English: [To some extent, the writing professional of written content with the “technical production” closely associated with,]$_N$ [it is a with machines work theory accordance of discourse model.]$_S$

(12) JUSTIFY

Nuclear: A text.

Satellite: Information that supports the writer's right to express the text.

Text Name: TERM30

Chinese: [但是对于其它鲜有人使用的语言来说，情况就不同了。]$_N$ [术语管理有时仅与语言学规划的政策相关，还有时仅包含个人情绪。]$_S$

English: [But to other very few used languages, the situation is different.]$_N$ [Terminology management is sometimes only with linguistic planning policies related, and sometimes contains personal emotions.]$_S$

(13) MEANS

Nuclear: An event or an idea.

Satellite: A way to make that event or idea becomes true.

Text Name: EEP4

Chinese: [论坛旨在重申"新丝绸之路"的倡议。]$_N$ [尤其是通过推动各社会团体、"智库"、公司和政府组织间对话交流来"促进亚欧的共同繁荣"。]$_S$

English: [The forum aims to reiterate the initiative of "New Silk Road".]$_N$ [In particular, through the promotion of dialogues and exchanges among various social groups, "think tanks", corporations and government organizations to "promoting the common prosperity of Asia and Europe".]$_S$

(14) MOTIVATION

Nuclear: An action.

Satellite: Information increases the reader's desire to perform the action.

Text Name: TERM39

Chinese: [对于上述小语种，并没有足够完善的词汇及专业术语资源来帮助学生进行学习。]$_S$ [我们为上述小语种人群收集了一个多语种词汇表用于这些小语种使用和教学，这是作为多媒体教学 CALL 的一部分。]$_N$

English: [For the above mentioned minority language, there is not enough vocabulary and professional terminology resources to help students to study,]$_S$ [We for above mentioned minority languages people have collected a multilingual glossary for these minority languages use and teaching, this as Multimedia Teaching Project CALL of part.]$_N$

(15) OTHERWISE[29]

Nuclear: An action or situation whose occurrence results from the lack of occurrence of the conditioning situation.

Satellite: Conditioning situation.

Text Name: RST webpage example

Chinese: [项目负责人应立刻为修改的手册提交修改后的条目。]$_N$ [否则，将使用现存条目。]$_S$

English: [Project leaders should immediately for the revised brochure to submit the entries.]$_N$ [Otherwise, it will use the existing entries.]$_S$

(16) PURPOSE

Nuclear: An intended situation.

Satellite: The intent behind the situation.

Text Name: ICEG2

Chinese: [欢迎浏览格拉纳达大学孔子学院过去举办过的讲习班的纪录，]$_N$ [以便您更好地参与其中。]$_S$

English: [Welcome visit Granada University Confucius Institute in the past held workshops record,]$_N$ [so that you could better participate in it.]$_S$

(17) PREPARATION

Nuclear: Content to be presented.

Satellite: Content that prepares the reader to expect and interpret the content to be presented.

Text Name: FICB4

Chinese: [1. 主办单位]$_S$ [西班牙加泰罗尼亚华侨华人社团联合总会]$_N$

English: [1. Organization Unit]$_N$ [Spain Catalonia Overseas Associations]$_S$

(18) RESTATEMENT

Nuclear: A situation.

Satellite: A reexpression of the situation.

Text Name: EEP3

[29] After the annotation work, we realize that there is no OTHERWISE relation in the corpus. The example of the OTHERWISE relation is extracted from the RST webpage. We translate the English example into Chinese.

Chinese: [此次联委会中谈到的最重要的几项内容为基建材料的合作，]$_N$ [准确地说是西班牙 ALBA 光源公司和上海光源公司间已有的协议及与 GTC 公司有关宇航用品的合作。]$_S$

English: [This time the Joint Committee meeting mentions the most important content is infrastructure materials cooperation,]$_N$ [Precisely speaking, Spanish ALBA Light Company and Shanghai Guangyuan Company already signed agreements and with GTC company related aerospace supplies cooperation.]$_S$

(19) RESULT

Nuclear: A situation.

Satellite: Another situation which is caused by that one.

Text Name: CCICE5

Chinese: [据悉，Inditex 集团主席与中国环保部部长在北京的一次会面上探讨引入生态店。]$_N$ [因此，中国将可能成为 Inditex 第一个拥有绿色生态店的市场。]$_S$

English: [It says that, Inditex Group Chairman with China Environmental Ministry Minister on Beijing meeting discuss the introduction of an ecological shop.]$_N$ [As a result, China will probably be Inditex first green ecological market.]$_S$

(20) SOLUTIONHOOD

Nuclear: A situation or method supporting full or partial satisfaction of the need.

Satellite: A question, request.

Text Name: TERM28

Chinese: [我们放弃了术语处理中通常使用的办法，]$_S$ [将研究建立在三个方面：信息资源、语料语言学特性以及翻译学研究。]$_N$

English: [We have given up terminology handling general approach,]$_S$ [will research based on three aspects: information resources, corpus linguistics and translation studies.]$_N$

Category: N-N

(21) CONJUNCTION

Nuclear: A situation or an action.

Nuclear: Another situation or another action that happens at the same time.

Text Name: ICEG1

Chinese: [它名下的图书馆中将会存有汉办捐赠的图书，]$_N$ [同时也会源源不断地补充添加新的书籍。]$_N$

English: [Its under name of library will have *Hanban* donated books,]$_N$ [At the same time also continue to add new books.]$_N$

(22) CONTRAST

Nuclear: One option.

Nuclear: The other option.

Text Name: CCICE1

Chinese: [具体而言，Mapfre 北美市场的保护费为 21.03 亿欧元，而南美市场为 18.23 亿欧元。]$_N$ [而去年情况相反，北美收入为 15.73 亿欧元，南美收入为 20.95 亿欧元。]$_N$

English: [Specifically, Mapfre North American market of protection fee is 2,103 billion euros, compared to 1.82 billion euros for the South American market.]$_N$ [In contrast to last year's situation, North American revenues were 1.573 billion euros and South American revenues were 2.095 billion euros.]$_N$

(23) DISJUNCTION

Nuclear: An alternative.

Nuclear: Another alternative.

Text Name: ICP7

Chinese: [读者用户们越来越频繁地使用智能手机来获取学术、工作以及科研领域的有效信息，]$_N$ [查收或回复最新的电子邮件，]$_N$ [亦或是了解图书馆的概况。]$_N$

English: [Readers are increasingly use smartphones to get academic, works and research effective information,]$_N$ [check or reply to the latest email,]$_N$ [or understand library's general information.]$_N$

(24) LIST

Nuclear: An item.

Nuclear: A next item.

Text Name: ICP3

Chinese: [她孜孜不倦地推动本土设计师的发展，]$_N$ [并且在竞争激烈的时尚界为他们创造了获取经验的平台。]$_N$

English: [She tirelessly promote local designers' development,]$_N$ [and in the competitive highly fashion industry for them to create a get experience platform.]$_N$

(25) SEQUENCE

Nuclear: An item.

Nuclear: A next item.

Text Name: FICB3

Chinese: [选手首先进行自我介绍，]$_N$ [然后随机选取一套与中国文化相关的问题进行现场问答，]$_N$ [最后一个环节选手则展示与中国才艺相关或者含有中国元素的才艺。]$_N$

English: [The contestants firstly gave a self-introduction,]$_N$ [then randomly selected a set of Chinese culture related questions to answer,]$_N$ [the last part is to show of Chinese talent related or contains Chinese elements talent.]$_N$

All the annotated texts can be consulted in the website. We give the annotation results as 3 forms: rs3, text and image. Figure 14 shows the how to consult the annotated texts from the corpus[30]

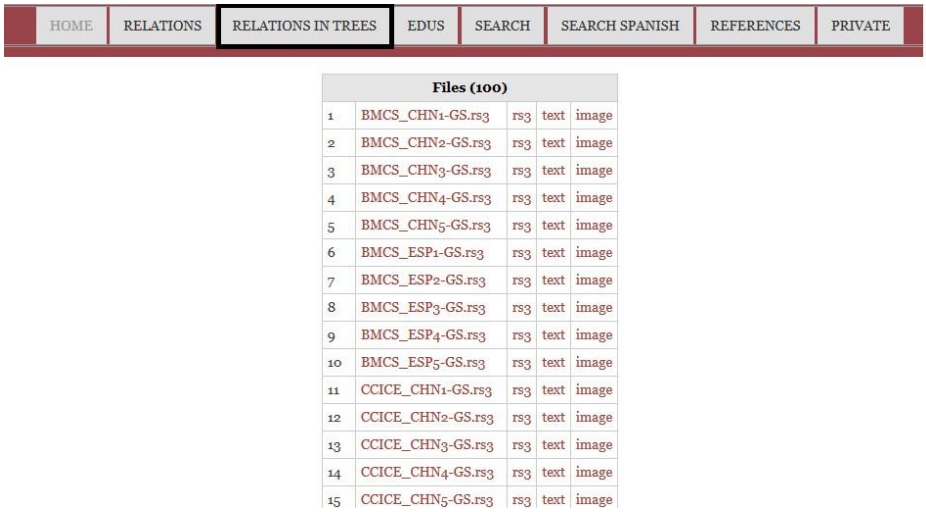

Figure 14. Corpus consultation with different ways

From Figure 14, we can see that, under the "RELATIONS IN TREES" column, users can consult the annotated texts by 3 different options: rs3, text and image. In addition, users can also consult each selected relation from the website (see Figure 15).

[30] Due to the space limitation, Figure 14 shows parts of the website.

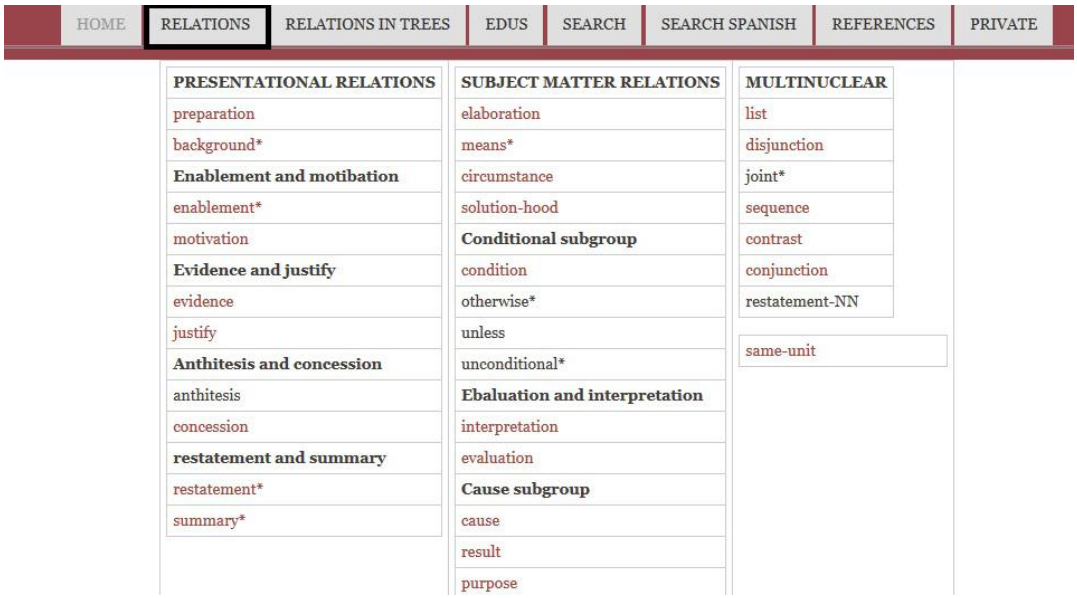

Figure 15. Consultation of each selected relations

In our website, as Figure 15 presents, under "RELATION" column, users can find each selected relation independently. Under each relation, all the texts that contain the corresponded relation can be found.

4.5 Evaluation Method

After finishing all the annotation tasks (segmentation, CU and discourse structure), we select Kappa as the evaluation method for segmentation and CU. Kappa has been used to measure the annotation agreement for the previous RST studies (Iruskieta, Diaz de Ilarraza and Lersundi, 2015; Cao et al., 2017). Kappa gives the agreement of annotation as follows:

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

In our work, P(A) represents the actual observed agreement, and P(E) represents chance agreement.

For the discourse annotation evaluation, we follow a qualitative evaluation method proposed by Iruskieta, da Cunha and Taboada (2015). All the detailed information about evaluation will be presented in the next chapter (Chapter 5).

4.6 Chapter Overview

In this chapter, we have introduced the methodology of the study. We elaborate the four steps to carry out the study: (i) corpus construction, (ii) segmentation annotation, (iii) CU annotation, and (iv) discourse structure annotation.

For the corpus compilation, we present the detailed information of the sources, genre and topics. For each annotation step, we list the annotation criteria and examples. In the following chapter, we evaluate the annotation quality and analyze the annotation disagreements.

Chapter 5
Annotation Evaluation and Analysis

The measurement of the annotation agreement reflects if the annotation is reliable. In this chapter, we will give the annotation evaluation and for each annotation step and the analysis for the annotation disagreements. In the Section 5.1, we will evaluate the segmentation annotation. We will calculate the accuracy of the annotation by using Kappa. We will explore the reason of the annotation disagreements. In the Section 5.2, our evaluation will go for CU annotation. We will also use Kappa to see the annotation results. Besides, we will conclude some words as the signal of the CU in our corpus. For the annotation disagreement, we will make a qualitative analysis. In the Section 5.3, we will evaluate the discourse relation annotation. Following the work by Iruskieta, da Cunha and Taboada (2015), we will apply this qualitative analysis for our discourse annotation. Four elements will be measured: Nuclearity (N), Relation (R), Composition (C), and Attachment (A). The inter annotator agreement will be measure with F-measure, using standard measurement for discourse relations. The Section 5.4 will summarize the information of this chapter.

5.1 Segmentation annotation evaluation and analysis

In this work, we use Cohen Kappa to measure inter-annotator agreement between the two corpus annotators (A1 and A2). The two annotators are Chinese linguistics. As we said in the previous chapter (Chapter 4), 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 5 includes the statistics used to measure the agreement between both annotators.

Other discourse evaluation measures have been employed to address the problematic of discourse evaluation measures. See Fournier (2013), and Sidarenka, Peldszus and Stede (2015) for further details.

Annotator		**A2**		**Total**
		Yes	No	
A1	Yes	765	101	866
	No	204	1888	2092
Total		969	1989	2958

Table 5. Segmentation cross tabulation

Table 6 includes the Kappa agreement results regarding each part of the corpus. The highest agreement between both annotators is 0.815, and the lowest agreement is 0.616. The agreement for the whole corpus is 0.76, which means the preliminary segmentation criteria are reliable for Chinese.

Corpus Source	**Kappa Agreement**
TERM	0.815
BMCS	0.719
CCICE	0.744
EEP	0.711
FCEC	0.711
FICB	0.616
ICP	0.759
ICG	0.705
Total	0.76

Table 6. Kappa results regarding each part of the corpus

After obtaining the evaluation of segmentation results, we analyze the disagreement sources between both annotators to establish the gold standard segmentation for our corpus. The following cases summarize the segmentation errors and include an example of the final segmentation decision:

- **Title**

A1: [2.] [术语构建] (×)

[2.] [Terminology construction]

A2: [2. 术语构建] (√)[31]

[2. Terminology construction]

Analysis: A1 has divided the title into two parts due to the period. However, we do not segment any element in a title or subtitle.

- **Comma + DM + verb**

A1: [这些内容不仅丰富了术语内容，] [**同时**还引起了一些术语基本定义的争论。] (√)

[These things have enriched the content of terms,] [**meanwhile** also cause some debates of the basic definition of terminology.]

A2: [这些内容不仅丰富了术语内容，**同时**还引起了一些术语基本定义的争论。] (×)

[These things have enriched the content of terms, **meanwhile** also cause some debates of the basic definition of terminology.]

Analysis: A1 has divided the sentence into two parts due to the comma. This segmentation is correct, because the discourse marker "*tongshi*" (同时) ('meanwhile') appears after the coma. Besides, the two parts have the same subject, and there is a verb "*fengfu*" (丰富) ('enrich') in the first EDU and another verb "*yinqi*" (引起) ('cause') in the second EDU.

- **Colon**

A1: [各种语言中唯一一致的命名参照物的情况是：] [术语均从英语中来。] (√)

[For all languages the only consistent reference is:] [all terminologies **come** from English.]

A2: [各种语言中唯一一致的命名参照物的情况是：术语均从英语中来。] (×)

[31] In this work, we use "√" to represent the correct segmentation and "×" to represent the incorrect segmentation.

[For all languages the only consistent reference is**:** all terminologies **come** from English.]

Analysis: A1 has divided the sentence into two parts due to the colon. In the preliminary version of segmentation criteria, colon was not considered; therefore, there is a disagreement regarding this punctuation mark between both annotators. We decide to segment the part after colon, because both EDUs include verbs: "*mingming*" (命名) ('to give name / dominate') in the first EDU and "*lai*" (来) ('come') in the second EDU.

- **Temporal adverb clause + comma + verb clause**

A1: [**当**上述内容均能在同一片文章中准确**描述**后**,**] [我们便能做到**建立**巴斯克语的"法律论述体系"。] (√)

[**When** all the previous mentioned can be **described** in the same passage,] [we can **establish** the "legal discourse system" for Basque.]

A2: [**当**上述内容均能在同一片文章中准确**描述**后，我们便能做到**建立**巴斯克语的"法律论述体系"。] (×)

[**When** all the previous mentioned can be **described** in the same passage, we can **establish** the "legal discourse system" for Basque.]

Analysis: A1 has divided the sentence into two parts due to the comma. The temporal adverb "*dang*" (当) ('when') and the comma can be considered as a segmentation boundary, because both EDUs include a verb: "*miaoshu*" (描述) ('describe') in the first EDU and "*jianli*" (建立) ('establish') in the second EDU.

- **Wrong EDU without verbs**

A1: [**包括** 12 副绘画作品和 2 副达利的原创作品，][以及 205 份杂志、报纸及宣传单。] (×)

[**Including** 12 paintings and 2 original works of Dalí,] [and 205 magazines, newspapers and advertisements.]

A2: [**包括** 12 副绘画作品和 2 副达利的原创作品，以及 205 份杂志、报纸及宣传单。] (√)

[**Including** 12 paintings and 2 original works of Dalí, and 205 magazines, newspapers and advertisements.]

Analysis: A1 has divided the sentence into two parts because it is a coordinated sentence. However, the segmentation of the annotator A1 is not correct because there is no verb in the second EDU. The only verb in this sentence is "*baokuo*" (包括) ('include').

Based on the error analysis, we have improved our segmentation criteria. Meanwhile, we carry out a debate between discourse experts and, taking our segmentation criteria into account, we have chosen the best segmentation option in case of disagreement.

Hence, we have created the gold standard segmented corpus for Chinese. This gold standard will be the basis for the discourse annotation of the corpus.

Table 7 shows the final criteria used for the discourse segmentation. We have divided the segmentation criteria into two types: EDU criteria and Non-EDU criteria.

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 adjunct verb clauses	

Table 7. Final discourse segmentation criteria

5.2 Central Unit (CU) annotation evaluation and analysis

CU is the key information of a text and is an important aspect for information retrieval study. The CU annotation can check annotators' natural comprehension ability of a text.

Two linguists (A1 and A2) annotated the CU for each segmented text in the corpus. For each segmented text, the annotators have identified which EDUs are the CUs of the text. Same as the segmentation evaluation, we also use Kappa to measure the agreement between the two annotators. Table 8 shows the agreement between the two annotators.

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

Table 8. Evaluation result of the CU annotation

From Table 8, we can see that the final evaluation result (0.977) proves that the CU annotation between the two annotators is almost perfect. Table 9 gives the annotation results of each part of the research corpus.

Texts name	Kappa result
TERM	0.974
BMCS	0.962
CCICE	0.952
EEP	0.951
FICB	1
FCEC	1
ICEG	1
ICP	0.987

Table 9. CU annotation results of each part

From Table 9 we can see that the annotation agreements of three parts (FICB, FCEC and ICEG) are totally perfect. The annotation agreements of rest parts are almost perfect (from 0.95 to 0.98). In addition, we have analyzed the texts with lower annotation agreements (< 0.8) to explore the causes. Table 10 shows the lowest results of CU annotation.

Texts name	Kappa result
BMCS2	0.743
EEP4	0.333
TERM25	0.764
TERM29	0.764

Table 10. Lowest results of CU annotation

(1) *Text Name*: BMCS2

Annotator1: [我们所有的老师都是西班牙语为母语的教师，]EDU_1 [受过专业对外西班牙语教学(ELE)资格培训，]EDU_2 [并具有在中国教学的丰富经验。]EDU3 [我们的教材为西班牙原版教材，]EDU_4 [内容新颖，]EDU_5 [适用于中国学生学习。]EDU_6

English: [We all teachers are Spanish native professors,]EDU_1 [trained in professional Spanish language teaching (ELE),]EDU_2 [and has in China rich teaching experience.]EDU_3 [Our teaching materials are Spain original materials,]EDU_4 [content novel,]EDU_5 [adequate to Chinese students learn.]EDU_6

Annotator2: [我们所有的老师都是西班牙语为母语的教师，]EDU_1 [受过专业对外西班牙语教学(ELE)资格培训，]EDU_2 [并具有在中国教学的丰富经验。]EDU_3

English: [We all teachers are Spanish native professors,]EDU_1 [trained in professional Spanish language teaching (ELE),]EDU_2 [and has in China rich teaching experience.]EDU_3

Analysis: For this text, the annotator A1 annotates more EDUs as CUs than A2. This text talks about two parts: the teacher and the teaching materials. Therefore, we think the teaching material information is as important as the teacher information and the second one need to be considered as the CU also. The annotation of A1 is correct.

(2) *Text Name*: EEP4

Annotator1: [论坛旨在重申“新丝绸之路”的倡议，]EDU_1 [尤其是通过推动各社会团体、“智库”、公司和政府组织间对话交流来“促进亚欧的共同繁荣”。]EDU_2

English: [The Forum aims to reaffirm “New Silk Road” the initiative,]EDU_1 [Especially through the promotion of social groups, “think tanks”, companies and government organizations, dialogue and exchange “to promote Asia and Europe common prosperity ”.]EDU_2

Annotator2: [10 月 28 日和 29 日，由国务院发展研究中心、国际关系和可持续发展中心、中国驻西班牙大使馆和和托雷多国际和平中心共同主办的第二届“丝路国际论坛 2015 年会”在马德里召开。]EDU_1

English: [October 28 and 29, by the State Council Development Research Center, International Relations and Sustainable Development Center, the Chinese Embassy in

Spain and Torredo International Peace Center co-sponsored the second "Silk Road International Forum 2015" Held in Madrid.]EDU_1

Analysis: The two annotators select different EDUs as the CUs. The annotator A1 thinks the aim of the forum is the main information while the annotator A2 considers the introduction of the forum is the main information of the text. The text concentrates on the introduction of the forum. Therefore, we think the annotation of A2 is adequate.

(3) *Text Name*: TERM25

Annotator1: [因此，近年来我们的工作目标在于将翻译过程中使用的各个方法（合理的术语使用、创建新的术语条目）]EDU_1 [以及巴斯克语必须能深层次融会贯通的各法律体系内容（西班牙、法国以及欧盟的法律）整合在一个文档中，]EDU_2

English: [Thus, in recent years, our goal is to translate the various methods used in the process (rational use of terms, to create a new term entry)]EDU_1 [as well as the legal system must be capable of deep content Basque mastery (Spain, France and EU law) integrated in a document,]EDU_2

Annotator2: [我们希望能按照实际情况呈现出这些年工作中碰到的问题以及取得的成就。]EDU_1

English: [We hope that we will be able to show the problems encountered and the achievements we have achieved in these years' work according to the actual situation.]EDU_1

Analysis: The text talks about the terminology translation. The main information of the text falls on how to translate the terminologies. The annotation of the annotator A1 reflects this main idea of the text. The annotation of annotator A2 is the part of the whole text, but cannot represent the main information of the text.

(4) *Text Name*: TERM29

Annotator1: [这也促使我们在进行专项研究时，不仅要兼顾上述理论原则，]EDU_1 [还应考虑在术语和信息学方面采用不同的方法论。]EDU_2 [同时，我们还应该面对上述问题，进行术语研究并整合相关结国。]EDU_3

English: [This prompted us to conducting specific research, not only to take into account the above theoretical principles,]EDU1 [also should consider in terms of terminology and informatics using different methodologies.]EDU2 [At the same time,

we also should face these problems, conduct research and integration of related terminology knot country.]EDU3

Annotator2: [自从计算机实现了语言信息存储及加工功能，术语便从未停止其适应各种技术创新的脚步，]EDU_1

English: [Since the computer to achieve the language information storage and processing functions, the term will never stop its adaptation to the pace of technological innovation,]EDU_1

Analysis: The annotation of the annotator A2 shows a phenomenon, not the key information of the text. The key information in this text is the problem of the research and how to solve the problem. Thus, the annotation of the annotator A1 is correct.

Lastly, based on the previous analysis, we also conclude the words that can be considered as the CU symbols. The indicators are listed in Table 11.

Noun	Verb	Proper Noun	Preposition	Pronoun	Conjunction
任务 (task)	旨在 (purpose)	我们 (we/us/ourselves)	自..以来 (since)	本/此 (this)	不仅..同时 (not only..but also)
目标 (goal)	传播/推广 (diffusion)		为 (for)		并/以及/还 (and/also)
	提供 (offer)		据 (based on)		
	阐述 (state)				
	描述 (describe)				

Table 11. The indications for CU annotation in Chinese texts

Here we give an example of each word in the corpus and the English literal translation for each example.

(Ex.1) Word (occurrences in total): 任务 (task) (1)

Text Name: FECE2

Chinese: 西中理事基金会是一家非营利性机构，它创建在 2004 年，主要**任务**是推广中国和西班牙两国间的外交关系，改善和提高西班牙在中国的形象和地位。

English: Spain-China Council Foundation is a non-profit organization, founded in 2004, main **task** is to promote China and Spain two countries of diplomatic relation, improve and enhance Spain in China of image and position.

(Ex.2) Word (occurrences in total): 旨在 (purpose) (5)

Text Name: TERM40

Chinese: 本文**旨在**描述系统用户在进行某专题信息内容复原时使用的语言工具：法律分类工具。

English: This article **purpose** describe system users for restoring a thematic message using language tools: law classification tool.

(Ex.3) Word (occurrences in total): 目标 (goal) (2)

Text Name: TERM50

Chinese: 该报告的**目标**在于展示该所大学的工作小组研究成果。

English: The report of **goal** is to display this university of research group achievements.

(Ex.4) Word (occurrences in total): 传播 (diffusion) (1)

Text Name: ICP3

Chinese: 北京塞万提斯学院除了开设西班牙语课程外，还负责**传播**西班牙及拉丁美洲文化。

English: Beijing Cervantes Institute in addition to offering Spanish courses, also responsible for the **diffusion** of Spanish and Latin America culture.

(Ex.5) Word (occurrences in total): 推广 (diffusion) (3)

Text Name: BMCS1

Chinese: 自中心成立以来，我们始终致力于通过开设适应不同学习需求的西班牙语课程、组织各种形式的文化活动和沙龙作坊, 以及面向公众的米盖尔·德·塞万提图书馆来**推广**西班牙语言教学及宣传西班牙语国家的文化。

English: Since its inception, we always dedicate to through offering adapted to different study needs of Spanish courses, the organization of various forms of cultural activities and salon workshops, and to the public of Miguel de Cervantes Library **diffusion** Spanish language teaching and broadcast Hispanic countries of culture.

(Ex.6) Word (occurrences in total): 提供 (offer) (2)

Text Name: BMCS4

Chinese: 作为 DELE 考试的主办机构，塞万提斯学院为想要通过考试获取水平证书的考生**提供**优质的考前准备课程。

English: As DELE exam organizer, the Cervantes Institute for (who) want to pass exams to get the level certificate candidates **offer** high-quality preparation courses.

(Ex.7) Word (occurrences in total): 阐述 (state) (1)

Text Name: TERM19

Chinese: 本次报告将借助加泰罗尼亚语术语标准实施中获得的经验，**阐述**建立术语规范的必要性，同样还会讨论面临的一些困难，并对当今社会中的这一形势提出一些构想。

English: This report will draw on the Catalan terminology standards implementation gained experiences, to **state** establish terminology standards the necessary, as well as discuss faces the difficulties, and to today's society of this situation to give some ideas.

(Ex.8) Word (occurrences in total): 描述 (describe) (4)

Text Name: TERM32

Chinese: 此份报告试图**描述**如何创造计算机科技支持下的术语构建器，并将其变为术语组中不可缺少的一部分。

English: This report aims to **describe** how to create computational technology support of terminology construction system, and makes it convert terminology group as integral part.

(Ex.9) Word (occurrences in total): 我们 (we/us/ourselves) (10)

Text Name: BMCS2

Chinese: **我们**所有的老师都是西班牙语为母语的教师，受过专业对外西班牙语教学(ELE)资格培训，并具有在中国教学的丰富经验。

English: **We** all of teachers are Spanish native-speaking teachers, have received ELE certificate training, and have in China teaching of rich experiences.

(Ex.10) Word (occurrences in total): 自. . 以来 (since) (2)

Text Name: EEP8

Chinese: **自** 1987 年创建**以来**，由卡勒斯·玛格拉内尔指挥的古典管弦音乐团从中世纪到十九世纪以来为西班牙音乐遗产的传承做出了许多研究性音乐学的工作。

English: **Since** 1987 founded, by Carles Magraner led of classical orchestra, from the Middle Ages to the 19th century for the Spanish musical heritage of legacy has made much research on musicology.

(Ex.11) Word (occurrences in total): 为[32] (for) (7)

Text Name: FICB5

Chinese: 第四届亚洲旅游国际会议于 11 月 23 日和 24 日在巴塞罗那举行，今年的主题 “**为**亚洲游客打造品质旅游”。

English: The 4th Asian Tourism International Conference was held on November 23rd and 24th in Barcelona, this year’s topic is “**for** Asian tourists to establish quality travel”.

(Ex.12) Word (occurrences in total): 据 (based on) (3)

Text Name: CCICE2

Chinese: **据**西班牙财政部在官网发布的消息显示，该机构将在本周二拍卖 6 至 12 月到期的短期国债，预期拍卖 40 亿至 50 亿欧元。

English: **Based on** the Spanish Ministry of Finance released on the official website of the message, the agency will on Tuesday auction from June to December short-term treasury bonds due, is expected to auction 4 billion to 5 billion euros.

(Ex.13) Word (occurrences in total): 本[33] (this) (6)

Text Name: TERM39

Chinese: **本**文提出了一个有关建立多语种术语库的方法论。

English: **This** paper presents a methodology for establishing a multilingual terminology database.

---

[32] The Chinese word *wei* (为) contains different meanings, and can be other phrase if combined with other word. Among all the annotated CUs, this word appear 17 times, but only 7 CUs include this word as the meaning of “for”.

[33] The Chinese word *ben* (本) can be other phrase if combined with other word. Among all the annotated CUs, this word appears 11 times, but only 6 CUs include this word as the meaning of “this”.

(Ex.14) Word (occurrences in total): 此[34] (this) (6)

Text Name: TERM18

Chinese: **此**份报告旨在从语言学和社会学的角度批判性的评价上述趋势。

English: **This** report aims to from a linguistic and sociological perspective critically evaluate the above mentioned trends.

(Ex.15) Word (occurrences in total): (不仅仅)...同时[35] (not only..but also) (2)

Text Name: ICP5

Chinese: 在我们学院学习西班牙语，**不仅仅**是学习语言本身，**同时**也是学习西班牙语世界的文化。

English: In our institute studying Spanish, is **not only** about learning the language itself, **but also** about learning Spanish-speaking world culture.

(Ex.16) Word (occurrences in total): 并 (and) (4)

Text Name: FICB2

Chinese: 2015 年 10 月 25 日，巴塞罗那孔子学院组织**并**举办了为期一天的本土汉语教师培训。

English: On 25th of October of 2015, Barcelona Confucius organized **and** conducted a one-day native Chinese teachers training.

(Ex.17) Word (occurrences in total): 以及 (and) (5)

Text Name: TERM34

Chinese: 本研究将着重介绍巴斯克语运用上述哪种方式进行表达，**以及**每种方法的区分界限在哪里（至今为止）。

English: This study will highlight the introduction of the Basque language uses above mentioned which methods to express, **and** each method of the distinction between in where (so far).

(Ex.18) Word (occurrences in total): 还 (and also) (4)

Text Name: ICP3

---

[34] The Chinese word *ci* (此) can be other phrase if combined with other word. Among all the annotated CUs, this word appears 9 times, but only 6 CUs include this word as the meaning of "this".

[35] The Chinese phrase *bujinjin...tongshi* (不仅仅...同时) is formed of two words: *bujinjin* (不仅仅) ('not only') and *tongshi* (同时) ('but also' / 'meanwhile'). In the Chinese expression, when expressing 'not only...but also', the word *bujining* (不仅仅) ('not only') can be erased, and the meaning in the context doesnot change. Among all the annotated CUs, the phrase *bujinjin...tongshi* (不仅仅...同时) appears once, and the only word *tongshi* (同时) appears once, which equivalents to *bujinjin...tongshi* (不仅仅...同时) under its content.

Chinese: 北京塞万提斯学院除了开设西班牙语课程外，还负责传播西班牙及拉丁美洲文化。

English: Beijing Cervantes Institute in addition to offering Spanish courses, **and also** responsible for the dissemination of Spanish and Latin American culture.

5.3 Discourse relation annotation evaluation and analysis

Discourse elements are reflected through the discourse relations annotations, for instance, the nuclearity order, number of DMs, definition of relations, etc. Two Chinese native-speaking annotators participate in the discourse relation annotation task.

Currently, under RST, for discourse annotation evaluation, two methods exist. One method is a quantitative analysis created by Marcu (2000). To quantify the agreement between the rhetorical analysis (EDUs, spans, nuclearity and rhetorical relations) done by annotators is the main character of this quantitative evaluation. However, da Cunha and Iruskieta (2010) and van der Vliet (2010) indicate some limitations of Marcu's methods:

(a). Factor confliction. The evaluated discourse elements, nuclearity and relation, are not independent of each other.

(b). Deficiencies in the descriptions. The description of comparison and weight used for agreement in certain discourse relations are still need to be improved.

Another evaluation method is a qualitative analysis created by Iruskieta, da Cunha and Taboada (2015). The qualitative evaluation method quantifies linguistic data for rhetorical structure, meanwhile this method also shows linguistic features affecting rhetorical structure. This is the first study provides a rigorous qualitative methodology for comparing of rhetorical structures. This method measures the agreement in rhetorical relations based on the following factors: constituent (C), attachment point (A) and the definition of relation (R), and solves the limitations of quantitative evaluations. Moreover, a qualitative description of agreement and disagreement can be provided under this qualitative method by means of types of agreement and sources of disagreements (disagreement of annotator and disagreements of language[36]). The statistics method used in this qualitative analysis is F-measure. Based on a corpus text,

[36] The disagreement of language is used for the bilingual / multilingual parallel corpus, not for the single language corpus. Therefore, in this study, we are not going to give the description of disagreement of language.

Figure 16 shows the annotation of A1 and Figure 17 shows the annotation of A2. Table 12 explains how we compare the annotations using this qualitative analysis. In the "Qualitative Evaluation" column, we use a "√" to represent an instance of agreement, and a "×" to show a disagreement. The last two columns show the result of the types of agreement (Agree) or the disagreement sources (Disagreement).

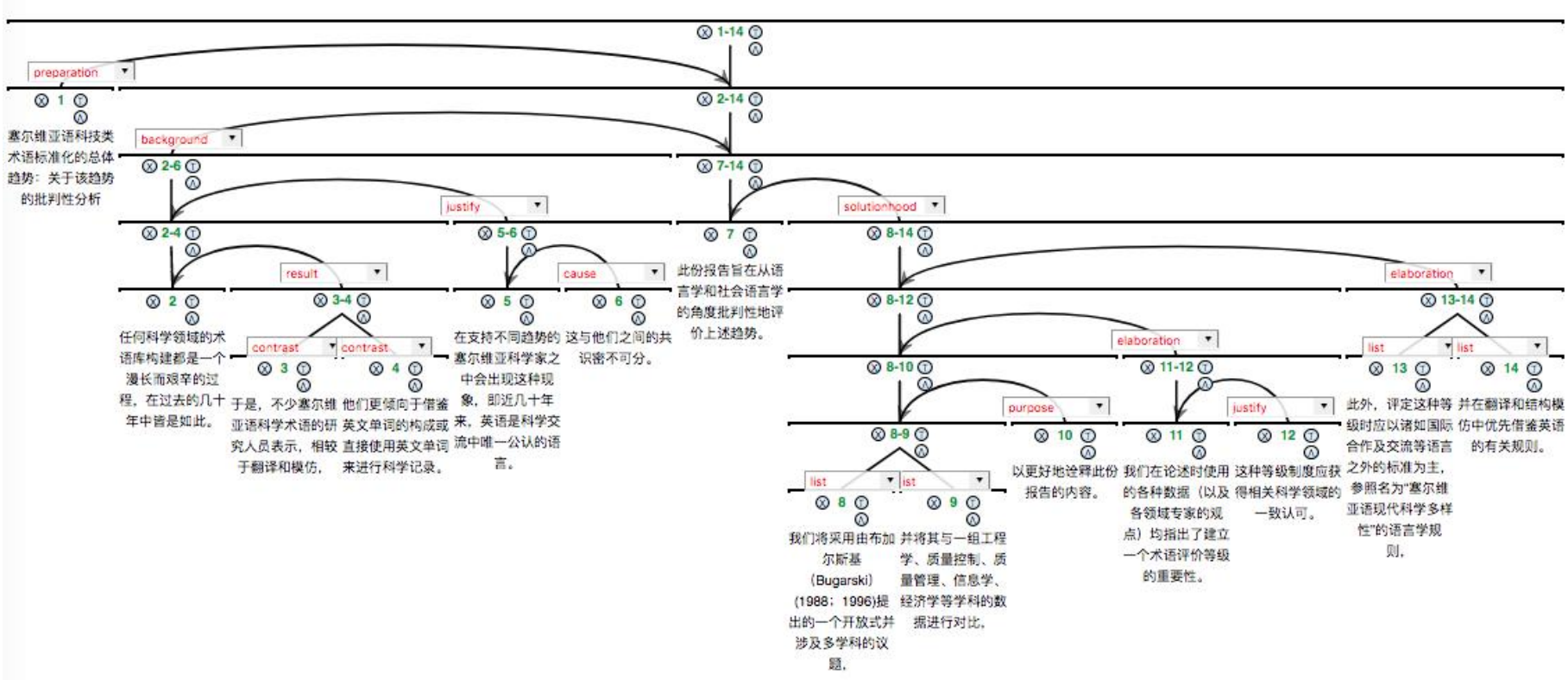

Figure 16. Discourse annotation of corpus text TERM18 by Annotator (A1)

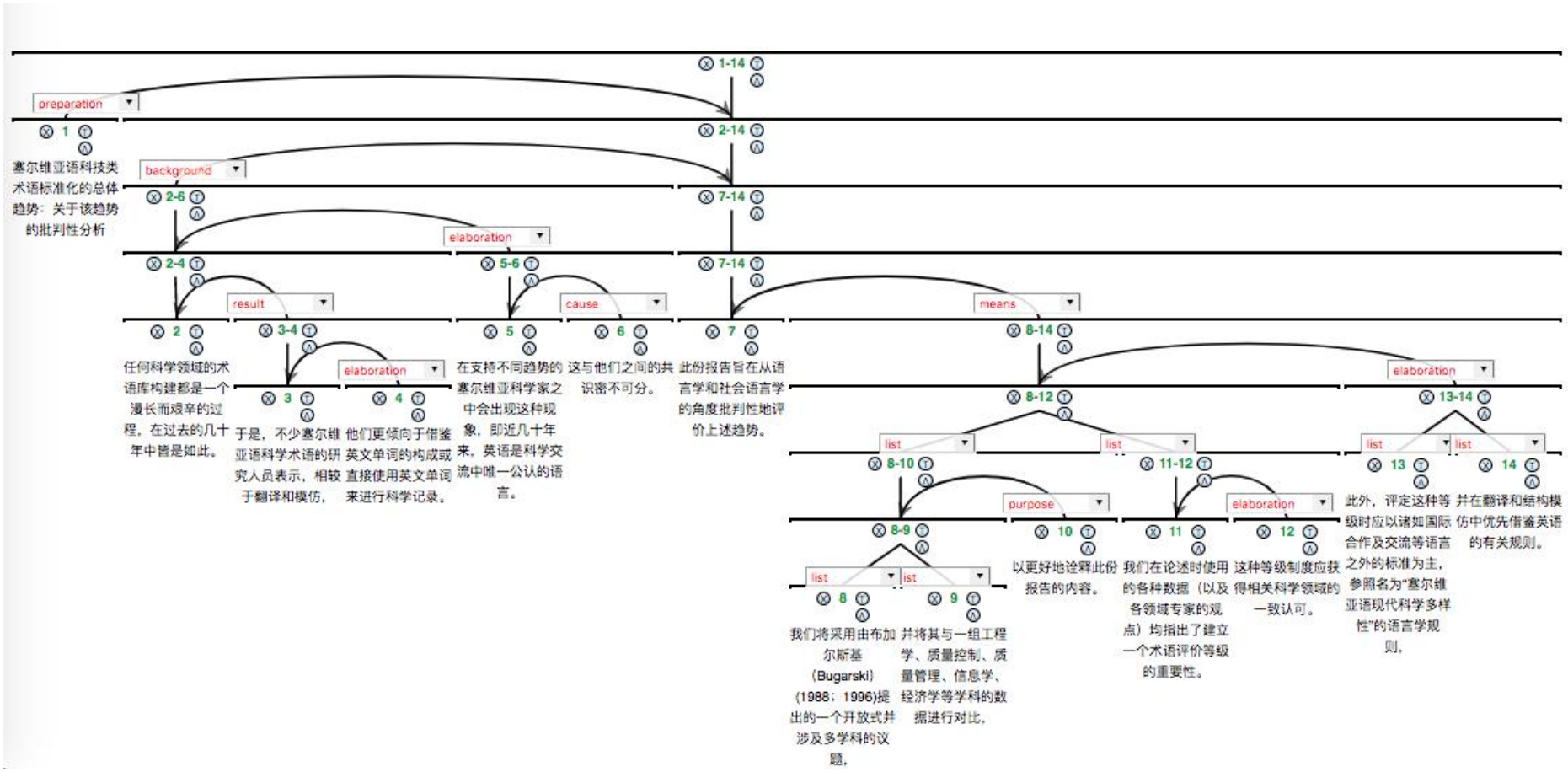

Figure 17. Discourse annotation of corpus text TERM18 by Annotator (A2)

Annotator (A1)				**Annotator (A2)**				**Qualitative Evaluation**					
CS	R	C	A	CS	R	C	A	N	R	C	A	Agree	Disagree
1	Preparation	1S	(2-14)N	1	Preparation	1S	(2-14)N	√	√	√	√	NRCA	
2-6	Background	(2-6)S	(7-14)N	2-6	Background	(2-6)S	(7-14)N	√	√	√	√	NRCA	
3-4	Result	(3-4)S	2N	3-4	Result	(3-4)S	2N	√	√	√	√	NRCA	
3|4	Contrast	3N|4N		4	Elaboration	4S	3N	√	√	√	√	NRCA	
5-6	Justify	(5-6)S	(2-4)N	5-6	Elaboration	(5-6)S	(2-4)N	√	√	√	√	NRCA	
6	Cause	6S	5N	6	Cause	6S	5N	√	√	√	√	NRCA	
8-14	Solutionhood	(8-14)S	7N	8-14	Means	(8-14)S	7N	√	√	√	√	NRCA	
8|9	List	8N|9N		8|9	List	8N|9N		√	√	√	√	NRCA	
10	Purpose	10S	(8-9)N	10	Purpose	10N	(8-0)N	√	√	√	√	NRCA	
11-12	Elaboration	(11-12)S	(8-10)N	(8-10)|(11-12)	List	(8-10N)|(11-12)N		×	×	×	×	×	NRCA
12	Justify	12S	11N	12	Elaboration	12S	11N	√	×	√	√	NCA	R
13-14	Elaboration	(13-14)S	(8-12)N	13-14	Elaboration	(13-14)S	(8-12)N	√	√	√	√	NRCA	
13|14	List	13N|14N		13|14	List	13N|14N		√	√	√	√	NRCA	

Table 12. Qualitative analysis of the corpus text TERM18

Table 12 concludes the annotation comparison between A1 and A2 by using the qualitative method. From Table 11 we can see that, besides of the full match (NRCA), two cases have different sources of disagreements:

(i) Difference choice in nuclearity entailed a N/N-N/S mix-up. From A1's annotation, we can see that the discourse relation between the EDU (11-12) and the EDU (8-10) hold a ELABORATION relation, EDU(11-12) is satellite and EDU (8-10) is the nucleus. However, the annotation of A2 is different. The annotator A2 considers the relation between the EDUs (8-10) and (11-12) is LIST, a multinuclear relation. Therefore, none of the evaluated discourse elements is match.

(ii) A relation has the same constituent and attachment point, but not the same relation label (≠R). In this case, the annotation of N and S is the same for annotator A1 and annotator A2. The only difference between the two annotators' annotation is the definition of the discourse relation. A1 gives a JUSTIFY relation for EDUs 11 and 12 meanwhile A2 defines an ELABORATION relation for the same EDUs.

Other sources of disagreement indicated by Iruskieta, da Cunha and Taboada (2015: 276) are:

- *Different choice in nuclearity entailed discrepancy in N/S relations (N/S).*
- *Relations chosen are similar in nature (Similar R).*
- *Relations with mismatched RST trees (Mismatch R).*
- *A relation is more specific than the other (Specificity).*
- *Different choice in attachment entailed a different relation (Attachment).*

The above-mentioned sources of disagreement also appear among other annotated texts. Table 13 shows the statistical result of the agreement.

Nuclearity		**Relation**		**Composition**		**Attachment**	
Match	F	Match	F	Match	F	Match	F
12 of 13	0.923	11 of 13	0.846	12 of 13	0.923	12 of 13	0.923

Table 13. F result of the annotation agreement under the qualitative method

From Table 13 we can see the annotation agreement between the two annotators is almost perfect for each discourse element.

Under the qualitative analysis, Iruskieta, da Cunha and Taboada (2015) emphasize a special case: a multinuclear relation is included in a constituent of another relation, as Figure 18 shows.

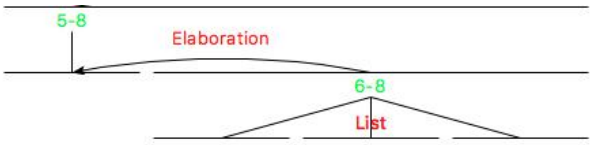

Figure 18. A multinuclear relation inside of a constituent of another relation

The example is extracted from the annotated text TERM29. EDU(6-8) and EDU(5-8) form a ELABORATION relation. However, EDU(6-8) contains EDU6, EDU7 and EDU8, and EDU10, which form a multinuclear relation known as LIST. This LIST (multinuclear) relation is inside the Elaboration relation (N-S). Two solutions are presented in this qualitative analysis: (1) Do not compare the relations and annotate as "no match"; (2) Compare the non-ambiguous relation first and leave the problematic comparisons until last. In this study, we opt for the second option. We first compare the ELABORATION relation and leave the LIST relation at last.

Table 14 gives the discourse relation annotation results of each in the corpus.

Texts name	Nuclearity		Relation		Composition		Attachment	
TERM	Match	F	Match	F	Match	F	Match	F
	313 of 357	0.877	278 of 357	0.799	313 of 357	0.877	312 of 357	0.874
BMCS	66 of 72	0.917	58 of 72	0.806	66 of 72	0.917	66 of 72	0.917
CCICE	44 of 45	0.978	38 of 45	0.844	44 of 45	0.978	44 of 45	0.978
EEP	60 of 64	0.938	54 of 64	0.844	60 of 64	0.938	60 of 64	0.938
FCEC	62 of 65	0.954	51 of 65	0.785	62 of 65	0.954	62 of 65	0.954
FICB	44 of 50	0.88	41 of 50	0.82	44 of 50	0.88	42 of 50	0.84
ICP	122 of 134	0.91	110 of 134	0.821	122 of 134	0.91	122 of 134	0.91
ICEG	19 of 22	0.864	16 of 22	0.727	19 of 22	0.864	19 of 22	0.864

Table 14. Annotation agreement of each part by using qualitative analysis

From Table 14 we can see that the annotation agreements of Nuclearity are almost perfect, all the annotation results are higher than 0.85. The annotation agreements of Relation are around 0.82, which also means the annotation results are almost perfect (>0.81). Although the annotation result in the ICEG part is 0.727, the result is substantial. Same as the annotation of Nuclearity, the annotations agreement of Composition and Attachment are also almost perfect.

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 annotation process and intra 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.

5.4 Overview of the annotation evaluation and analysis

In this chapter, we have evaluated the annotation agreement by using Kappa for segmentation annotation and CU annotation. For the segmentation annotation, the agreement for the whole corpus is 0.76 kappa, which means the preliminary segmentation criteria are reliable for Chinese. We give an error analysis for segmentation annotation and improved the segmentation criteria. The gold standard segmentation criteria have been listed in Table 6.

For CU annotation, two annotators give almost perfect agreement (all evaluated parts/clusters are >0.9). In addition, the possible words that can be considered as the signals of the CU have also been extracted.

For discourse relation annotation evaluation, we follow a qualitative analysis created by Iruskieta, da Cunha and Taboada (2015). We measure the agreement by means of Nuclearity, Relation, Composition and Attachment. This qualitative analysis has overcome the limitations of the quantitative analysis created Marcu (2000). By using the qualitative analysis, the compared EDUs are at the same discourse level, and the description of the comparison is clearer than using the quantitative evaluation method. The annotation agreements of the discourse elements (N, R, C and A) are also almost perfect.

Chapter 6

Conclusions

In Chapter 1, we have indicated the background of the research. In Chapter 2, we have introduced the theoretical framework and different discourse analysis theories and approaches. We have introduced the two annotation interfaces related with the theoretical framework. In addition, we have discussed about the applications of RST. In Chapter 3, we have explained the related works. We have presented different RST Treebanks for distinct languages. In addition, we have analyzed the related discourse studies for Chinese. In Chapter 4, we have described the methodology of this work. We have described the research corpus in detail and the different annotation steps. In Chapter 5, we have explained the evaluation results and analysis of annotation disagreements. In this chapter, we will conclude the study and look forward to the future work.

6.1 General conclusions

In this work, we have introduced the first open RST Chinese Discourse Treebank (RCDT). As the most spoken language in the world, the Chinese hold an important position in the NLP research community. Meanwhile, with the growing interesting on discourse analysis, it is important to discover how discourse elements are being expressed in Chinese.

As mentioned in Chapter 3, the previous existed discourse studies for Chinese analysis are based on the discourse approach PDTB (Marcus, Santorini and Marcinkiewicz, 1993; Prasad et al., 2008); few works use RST (Manna and Thompson, 1988) for Chinese discourse analysis. Although two works explore the Chinese discourse structure by means of RST, none of them is accessible. Moreover, two corpora include very few annotated texts. The simple genre and the topics of texts in the corpora cannot reflect the discourse structure diversity. With the aim to fulfill these gaps, we have compiled a new corpus for Chinese discourse analysis. The main characters of the corpus are the complexity of discourse structure and heterogeneity. The corpus consists of texts from different sources. Moreover, the genres and topics of the texts are different.

The annotation work has been divided into three steps: segmentation, Central Unit (CU) annotation and discourse relations annotation (both intra-sentence annotation and inter-sentence annotation). Two Chinese annotators have participated in this project. For each annotation part, we have evaluated the annotation agreement. Additionally, we have analyzed the disagreements for each annotation part. The results show that the annotations of our corpus are reliable. Lastly, we have made our annotation data available to the scientific community. All the data can be consulted at `http://ixa2.si.ehu.es/rst/zh/`.

6.2 Contributions

This research concentrates on the Chinese discourse analysis under RST. The research corpus has been annotated with discourse information. The main contributions of this work are the following:

- *Research corpus.* The corpus is enriched with POS information, which can be useful for different NLP tasks, for instance, machine translation. The texts in the corpus can be downloaded and can be applied to other discourse analysis.
- *Segmentation.* Segmentation is the crucial step for discourse analysis and many other NLP tasks. In this work, we have elaborated gold standard for segmentation, which can be useful for other RST analysis. Besides, the gold standard can also be used for other languages under discourse analysis.
- *Central Unit.* All the texts have been annotated with their CUs.
- *Discourse annotation.* Discourse relations show the coherence of a language and can be useful for several NLP tasks, such as, discourse parsing, information extraction, automatic summarization and evaluation of MT (Taboada and Mann, 2006).

6.3 Future work

The results of this dissertation set various new lines of research. Currently, the corpus contains the annotated Chinese subcorpus and its annotated Spanish subcorpus. Since this study focuses on the discourse analysis of Chinese, therefore, we donot give the information about the Spanish subcorpus here. Currently, we are comparing the

annotation results between the Chinese subcorpus and the Spanish subcorpus. We will also publish the annotation data of the Spanish subcorpus.

An annotated Chinese-Spanish parallel corpus with discourse information can be useful for both human linguistics tasks and automatic linguistics tasks. For human linguistics tasks, the parallel corpus will help the human translation between Chinese and Spanish. For automatic linguistics tasks, this parallel corpus will help the MT between the two languages.

Reference

Asher Nicholas, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Bengoetxea Kepa, Atutxa Aitziber, and Iruskieta Mikel. 2017. A Machine Learning based Central Unit Detector for Basque Scientific Texts. *Procesamiento del Lenguaje Natural*, 58: 37-44.

Braud Cholé, Plank Barbara, and Søgaard Anders. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING' 2016)*, 1903-1913.

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2016. A Spanish-Chinese Parallel Corpus for Natural Language Processing Purposes. In *Proceedings of Parallel Corpora: Creation and Application International Symposium PaCor2016*. 12.

Chen Keh-Jiann, Huang Chu-Ren, Chang Li-Ping, and Hsu Hui-Li. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC' 1996)*, 167-176.

da Cunha, Iria. 2008. *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. PhD thesis. Barcelona: Universitat Pompeu Fabra.

da Cunha Iria. 2016. Towards discourse parsing in Spanish. In *Proceedings of TextLink – Structuring Discourse in Multilingual Europe*, 40-44.

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, San Juan Eric, Torres-Moreno Juan-Manuel, Castellón Irene, and Lloberes Marina. 2016. Extending Automatic Discourse Segmentation for Texts in Spanish to Catalan. In *Proceedings of the 1st International Workshop on Modeling, Learning and Mining for Cross/Multilinguality*, 36-45.

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

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.

Fomicheva Marina, da Cunha Iria, and Sierra Gerardo. 2012. La estructura discursiva como criterio de evaluación de traducciones automáticas: una primera aproximación. *Empiricism and Analytical Tools for 21st Century Applied Linguistics*: 973-986.

Guzmán Francisco, Joty Shafiq, Màrquez Lluís, and Nakov, Preslav. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL' 2014)*, 687-698.

Hanneforth Thomas, Heintze Silvan, and Stede Manfred. 2011. Rhetorical Parsing with Underspecification and Forests. In *Proceedings of NNACL-HLT 2013*, 31-33.

Heilman Michael, and Sagae Kenji. 2015. Fast Rhetorical Structure Theory Discourse Parsing. *arXiv:1505.02425*.

Hiong Siaw Nyuk, Kulathuramaiyer Narayanan, and Labadin Jane. 2012 Towards Structure-Based Paraphrase Detection Using Discourse Parser. *Journal of Information Retrieval and Knowledge Management*, 2: 96-103.

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.

Huang Chu-Ren, Chen Feng-Yi, Chen Keh-Jiann, Gao Zhao-Ming, and Chen Kuang-Yu. 2000. Sinica Treebank: design criteria, annotation guidlines, and on-line interface. In *Proceedings of the second workshop on Chinese language processing*, 29-37.

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

Iruskieta Mikel. 2015. Corpus exploration of discourse relations in RST. In *Proceedings of 1st Training School: Methods and tools for the analysis of discourse relational devices*, 18-22.

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, Díaz de Ilarraza Arantza, and Lersundi Mikel. 2011. Bases para la implementación de un segmentador discursivo para el euskera. In *Proceedings of Anais do III Workshop A RST e os Estudos do Texto*, 18-29.

Iruskieta Mikel, and Zapirain Benat. 2015. EusEduSeg: A Dependency-Based EDU Segmenttion for Basque. *Procesamiento del Lenguaje Natural*, 55: 41-48.

Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, and Alicia Pérez. 2015. Using relational discourse structure information in Basque sentiment analysis. In *Proceedings of 5th Workshop RST and Discourse Studies*.

Joty Shafiq, Carenini Giuseppe, and Ng Raymond T. 2015. CODRA: A Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3): 385-435.

Kepa Bengoetxea, Mikel Iruskieta. 2018. A Supervised Central Unit Detector for Spanish. *Procesamiento del Lenguaje Natural*, 60: To appear.

Levy Roger, and Manning Christopher. 2003. In *Proceedings of 41st Annual Conference of the Association for Computational Linguistics (ACL' 2003)*, 439-446.

Li Chengcheng. 2010. Automatic Text Summarization based on Rhetorical Structure Theory. In *Proceedings of 2010 International Conference on Computer Application and System Modeling (ICCASM' 2010)*, 595-598.

Li Junyi Jessy, Carpuat Marine, and Nenkova Ani. 2014. Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short papers) (ACL' 2014)*, 283-288.

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

Marcu Daniel. 1997. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL' 97/EACL' 97)*, 96-103.

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

Marcus Mitch, Santorini Beatrice, and Marcinkiewicz Mary Ann. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

Mathkour Hassan I., Touir Ameur A., and Al-Sanea Waleed A. 2008. Parsing Arabic Texts Using Rhetorical Structure Theory. *Journal of Computer Science*, 4(9): 713-720.

Mayor Aingeru, Alegria Iñaki, Díaz de Ilarraza Arantza, Labaka Gorka, Lersundi Mikel, and Sarasola Kepa. 2009. Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural*, 43: 197-205.

Meyer Thomas, and Polakova Lucie. 2013. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the Workshop on Discourse in Machine Translation (DiscoMT)*, 43-50.

Moens Marie-Francine, and de Busser Rik. 2002. First steps in building a model for the retrieval of court decisions. *International Journal of Human-Computer Studies*, 57(5): 429-446.

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

Otegi Arantxa, Imaz Oier, Díaz de Ilarraza Arantza, Iruskieta Mikel, and Uria Larraitz. 2017. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural*, 58: 77-84.

Pardo Thiago Alexandre Salgueiro, and Nunes Maria das Graças Volpe. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, 15(2): 43-64.

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 A. S. and Seno Eloize R. M. 2005. Rhetalho: um corpus dereferência anotado retori-camente. *Anais do V Encontro de Corpora*. São Car-los-SP, Brasil.

Pórtoles José. 2001. *Marcadores del discursivo*. 4th edition. Barcelona: Ariel.

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.

Prasad Rashmi, Joshi Aravind, and Webber Bonnie. 2010. Exploiting Scope for Shallow Discourse Parsing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC' 2010)*, 2076-2083.

Qiu Wusong. 2010. *Jiyu XiuciJiegou Lilun de Hanyu Xinwenpinglun Yupian Jiegou Yanjiu* (基于修辞结构理论的汉语新闻评论语篇结构研究, *[*The Discourse Structure Analysis of Chinese News Comments: Based on the Rhetorical Structure Theory*]*). Master thesis. Nanjing: Nanjing Normal University.

Schiffrin Deborah. 2001. Discourse markers: language, meaning, and context. *The handbook of discourse analysis*, 1: 54-75.

Shinmori Akihiro, Okumura Manabu, Marukawa Yuzo, and Iwayama, Makoto. 2002. Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases. In *Proceedings of the 3rd NTCIR Workshop*.

Surdeanu Mihai, Hicks Thomas, and Valenzuela-Escárcega Marco A. 2015. Two Practical Rhetorical Structure Theory Parsers. In *Proceedings of NAACL-HLT 2015*, 1-5.

Stepanov Evgeny A., and Riccardi Giuseppe. 2014. Towards Cross-Domain PDTB-Style Discourse Parsing. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, 30-37.

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

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 the 6th Workshop Recent Advances in RST and Related Formalisms*, 29-33.

Tu Mei, Zhou Yu, and Zong Chenqing. 2013. A Novel Translation Framework Based on Rhetorical Structure Theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL' 2013)*, 370-374.

van Dijak Teun A. 1980. *MACROSTRUCTURES: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* New Jersey: Lawrence Erlbaum.

van der Vliet Nynke. 2010. Inter annotator agreement in discourse analysis. http://www.let.rug.nl/%7Enerbonne/teach/rema-stats-meth-seminar/presentations/NvdV-Cohens-Kappa-2010.pdf

Wilks Yorick. 2009. *Machine Translation: Its scope and limits*. 3rd ed. New York: Springer.

Xue Nianwen, Xia Fei, Chiou Fu-Dong, and Plamer Martha. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2): 207-238.

Yue Ming. 2006. *Hanyu Caijingpinglun de Xiucijiegou Biaozhu ji Pianzhang Yanjiu* (汉语财经评论的修辞结构标注及篇章研究, *[*Annotation and Analysis of Chinese Financial News Commentaries in terms of Rhetorical Structure*]*). PhD thesis. Beijing, Communication University of China.

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

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.