Using Annotated Discourse Information of a RST Spanish-Chinese Treebank for Translation and Language Learning Tasks

Shuyuan Cao

TESI DOCTORAL UPF / 2018

DIRECTOR DE LA TESI

Dra. Iria da Cunha (Universidad Nacional de Educación a Distancia)

Dr. Mikel Iruskieta (Universidad del País Vasco) (UPV-EHU)

DEPARTAMENT DE TRADUCCIÓ I CIÈNCIES DEL LLENGUATGE

To my parents

致我最亲爱的父母

I, I did it all
我做了我能做的一切

I, I did it all
我做了我能做的一切

I owned every second
我未虚度年华

That this world could give
是上天给我的礼物

I saw so many places
我走过千山万水

The thing that I did
历遍人间

With every broken bone
也曾粉身碎骨

I swear I lived
但我发誓 我来过

< I lived --- OneRepublic>
<I lived --- 一体共和>

Acknowledgments

First, to my supervisors Iria da Cunha and Mikel Iruskieta, for their high agreement.

To Nuria Bel, for her help to get me enrolled into the PhD program, and the inspiration that she gave me during the first year of my PhD study.

To Nianwen (Bert) Xue, for his insightful ideas and fruitful collaboration during my stay at Brandeis University.

To Liu Rui and Pan Wei, for treating me like family and for the moments we shared together during these years in Barcelona. We are a family forever!

To Chen Chanwen, for your laughs and interesting talks. Without this friend, my life in Barcelona wouldn't be wonderful like now.

To Tian Mingge, for the days we spent together in the library during our PhD study, and for accompanying me on the road and all the nice moments we shared together.

To Xia Pu, for taking care of me during my stay in Boston. I will never forget the delicious foods that you showed me, the TV shows that we watched together.

To Dai Jue, for your words of motivation during my PhD study, and for sharing laughs and delicious foods with me.

To my HAP/LAP/EMLCT classmates, you guys are cute. A special thanks to Harritxu Gete for the group works we did together and the excellent work we did for LREC2018. Also, to Cristina Aceta, for all your support and encouragement.

A special thanks goes to my parents, the most important people in my life, for their unconditional love and support.

Lastly, I want to thank the following people, you guys are pretty nice:
Han Jingyi (for recommending me to my supervisor), Yang Fengrong (for the encouragement and support), Liu Shiyang (for always supporting me), Amir Zeldes (for creating the awesome annotation interface), Mireia Vargas Urpí (for the wonderful idea of developing the parallel corpus), Yao Gang (for your kind help to annotate the corpus), Vanessa Alonso, Rafael José and Elisenda Bernal (for your generous help with all of the complicated procedures and documentation), Kenny Lino (for your great corrections of my thesis), Maite Aragonés (thanks for the support of my PhD study).

Abstract

As one of the essential elements for Natural Language Processing (NLP), discourse has called much attention during recent years. Many studies explore the role of how discourse elements affect in different NLP research areas, such as parsing, sentiment analysis, machine translation evaluation, among others. Besides, along with the discourse analysis development, different treebanks annotated with discourse information for different languages form a great contribution for advancing the NLP researches.

Spanish and Chinese are two of the most spoken languages in the world; the language pair occupy an important position for NLP studies. Therefore, this study aims to make a discourse analysis between the two languages in terms of annotating discourse similarities and differences under the theoretical framework of Rhetorical Structure Theory (RST) by Mann and Thompson (1988).

Our goal, which is the main objective of this study, based on the annotation results, the study seeks to develop a protocol that includes recommendations for Spanish-Chinese translation. In addition, with a globalized context in the current society, the communication between Spanish and Chinese is more and more intensive. Therefore, another intention of our study is to develop some resources for the language learning between Spanish-Chinese.

To achieve our goals, for the development of the protocol, we firstly establish a Spanish-Chinese parallel corpus and annotate the discourse information of the entire corpus. Then we evaluate the annotation results following a qualitative method to guarantee the high quality of the annotation results. Lastly, we conclude the discourse similarities and differences to make the protocol. Regarding the language learning between the two languages, we fully use the manually annotated discourse markers (DM) to develop a question-answering module.

In recent years, there have been few contrastive works of Spanish and Chinese for discourse analysis. Therefore, this PhD study aims to partially fill a knowledge gap in the study between Spanish and Chinese.

Resumen

Como uno de los elementos esenciales para el Procesamiento del Lenguaje Natural (PLN), el discurso ha llamado mucho la atención durante los últimos años. Diversos estudios exploran el papel de cómo los elementos del discurso afectan en diferentes áreas de investigación del PLN, por ejemplo, el análisis sintáctico, el análisis de sentimientos, la evaluación de la traducción automática, entre otros. Además, junto con el desarrollo del análisis del discurso, diferentes treebanks anotados con infomación discursiva para diferentes idiomas forman una gran contribución para el avance de las investigaciones del PLN.

El español y el chino son dos de los idiomas más hablados en el mundo, ambos ocupan un lugar importante para los estudios de PNL. Por lo tanto, este estudio pretende hacer un análisis del discurso entre las dos lenguas en términos de anotar similitudes y diferencias del discurso bajo el marco teórico Teoría de la Estructura Retórica (RST) de Mann y Thompson (1988).

El objetivo principal de este estudio, basado en los resultados de la anotación, busca desarrollar un protocolo que incluya recomendaciones para la traducción entre el español y el chino. Además, en un contexto globalizado en la sociedad actual, la comunicación entre españoles y chinos es cada vez más intensa. Por lo tanto, la otra intención de nuestro estudio es desarrollar algunos recursos para el aprendizaje de idiomas entre los españoles y los chinos.

Para lograr nuestros objetivos de desarrollo del protocolo, primero establecemos un corpus paralelo español-chino y anotamos la información discursiva de todo el corpus. Luego evaluamos los resultados de la anotación siguiendo un método cualitativo para garantizar la alta calidad de los resultados de anotación. Por último, concluimos las similitudes y diferencias del discurso para hacer este protocolo. Con respecto al aprendizaje de lenguas entre el español y el chino, utilizamos completamente los marcadores discursivos (MD) anotados manualmente para desarrollar un módulo de preguntas y respuestas.

En los últimos años, han habido pocos trabajos que comparen el español y el chino. Por lo tanto, este estudio de doctorado tiene como objetivo llenar parcialmente una brecha de conocimiento entre el estudio de las lenguas española y china.

Contents

1 Introduction..... 7

1.1 Motivation..... 7
1.2 Objectives..... 9
1.3 Hypothesis..... 10
1.4 Thesis Structure..... 10
1.5 Publications..... 11

2 Theoretical Framework..... 12

2.1 Rhetorical Structure Theory (RST)..... 12
2.2 Definition of DMs..... 15
2.3 RST Annotation Tools..... 17
2.3.1 RSTTool..... 17
2.3.2 rstWeb..... 18
2.4 RST Applications..... 20
2.5 Chapter Overview..... 21

3 State of the Art..... 23

3.1 RST Annotation..... 23
3.1.1 Segmentation..... 23
3.1.2 CU Annotation..... 24
3.1.3 Discourse Structure Annotation..... 26
3.2 Corpus-based Discourse Analysis for Spanish..... 30
3.3 Corpus-based Discourse Analysis for Chinese..... 30
3.4 RST-based Comparative Studies..... 33
3.5 Language Learning Using Corpus-based Approach..... 33
3.6 Chapter Overview..... 35

4 Methodology..... 37

4.1 Corpus Compilation..... 37
4.1.1 Analysis of the Previous Spanish-Chinese Parallel Corpora..... 38
4.1.2 Development of the Corpus..... 39
4.2 Discourse Segmentation..... 42
4.2.1 Elaboration of the Discourse Segmentation Criteria..... 42
4.2.2 Evaluation of Inter-annotator Agreement..... 45
4.3 CU Annotation..... 46
4.3.1 Description of the CU Annotation Criteria..... 47
4.3.2 Evaluation of Inter-annotator Agreement..... 48

4.4 Discourse Structure Annotation..... 48
4.4.1 Description of the Discourse Structure Annotation Criteria..... 48
4.4.2 Evaluation of Inter-annotator Agreement.....51
4.5 Elaboration of a Spanish-Chinese Discourse Recommendation Protocol for Translators.....53
4.5.1 RS-tree Comparison..... 53
4.5.2 Translation Strategies.....53
4.5.3 Elaboration of Recommendations..... 54
4.6 Applications of Results for Spanish-Chinese Language Learning Tasks..... 55
4.6.1 Annotation of DMs.....55
4.6.2 General Information for Exercise Elaboration..... 55
4.6.3 Exercise Evaluation for Spanish and Chinese.....55
4.7 Chapter Overview.....55

5 Evaluation and Analysis.....57

5.1 Discourse Segmentation.....57
5.1.1 Segmentation.....57
5.1.2 Discussion of Segmentation Results..... 59
5.2 CU Annotation..... 64
5.2.1 CU Annotation Results.....64
5.2.2 Discussion of the CU Annotation Results.....65
5.3 Discourse Structure Annotation..... 74
5.3.1 Discourse Structure Annotation Results.....74
5.3.2 Discussion of Discourse Structure Annotation Results.....79
5.4 Chapter Overview.....82

6 Elaboration of a Spanish-Chinese Discourse Recommendation Protocol for Spanish-Chinese Translation.....84

6.1 Results regarding Discourse Differences and Similarities in the Corpus Annotation.....84
6.2 Discussion of the Detected Discourse Differences and Similarities..... 84
6.3 Final Recommendation Protocol.....89

7 A Spanish-Chinese Language Learning Task by Using Technological Corpus-based Resources..... 116

7.1 Level Requirement for the Spanish-Chinese Language Learning with the Corpus..... 116
7.2 A Task for Spanish-Chinese Language Learning of DMs..... 117
7.2.1 Annotation of DMs.....117
7.2.2 Exercise Elaboration.....120
7.2.3 Analysis and Discussion.....121

7.3 Chapter Overview....123

8 Conclusions and Future Work....124

8.1 Conclusions....124
8.2 Contributions....127
8.3 Limitations....127
8.4 Future Work....128

References....129

Appendix A Corpus Text Links....142

Appendix B Examples of Criteria for Annotation Steps....146

B.1 Discourse Segmentation Criteria Examples....146
B.2 CU Annotation Examples with Representative Words....153
B.3 Discourse Structure Examples....165

Appendix C Special Discourse Comparison Cases....178

Appendix D List of Annotated Texts Links....183

Appendix E Encoding of the Task for Spanish-Chinese Language Learning....187

List of Tables

Table 1. Original classification of RST relations..... 12
Table 2. Relation classification by subject matter and presentational basis..... 13
Table 3. Detailed information of each RST Treebank..... 29
Table 4. Summary of each discourse treebank for Chinese..... 32
Table 5. Genre information of the corpus..... 40
Table 6. Domain information of the corpus..... 41
Table 7. Selected discourse relations for discourse annotation.....48
Table 8. Segmentation cross tabulation of the Spanish subcorpus.....57
Table 9. Segmentation cross tabulation of the Chinese subcorpus..... 58
Table 10. K results regarding each part of the corpus (Spanish subcorpus)..... 58
Table 11. K results regarding each part of the corpus (Chinese subcorpus)..... 58
Table 12. Final discourse segmentation criteria.....64
Table 13. Statistical segmentation information of the corpus.....64
Table 14. CU annotation evaluation result of the Spanish subcorpus.....65
Table 15. CU annotation evaluation result of the Chinese subcorpus.....65
Table 16. Evaluation results of CUs annotation (Spanish subcorpus)..... 65
Table 17. Evaluation results of CUs annotation (Chinese subcorpus)..... 69
Table 18. The indications of CU in the research corpus..... 73
Table 19. Qualitative evaluation for the discourse annotation of the Spanish text BMCS1.....75
Table 20. Qualitative evaluation for the discourse annotation of the Chinese text BMCS1.....76
Table 21. Qualitative evaluation results of the Spanish subcorpus.....77
Table 22. Qualitative evaluation results of the Chinese subcorpus.....78
Table 23. Qualitative evaluation of the harmonized corpus between Spanish and Chinese..... 79
Table 24. Statistics of the translation strategies found in the corpus annotation..... 84
Table 25. Information of the different discourse cases in the corpus.....85

Table 26. Statistics of cases including a DM only in the Chinese passages..... 85
Table 27. Different discourse relations for Spanish-Chinese parallel contents..... 86
Table 28. Statistics of cases including different DMs in parallel passages.....87
Table 29. The correct Spanish translations of the Chinese DMs in the corpus.....87
Table 30. The correct Chinese translations of their Spanish DMs in the corpus..... 88
Table 31. Statistical summary of other cases..... 88
Table 32. Program accuracy of the 40 tested Spanish-Chinese parallel texts.....121
Table A.1 Text links of the BMCS part..... 142
Table A.2 Text links of the CCICE part.....142
Table A.3 Text links of the EEP part..... 143
Table A.4 Text links of the FICB part.....143
Table A.5 Text links of the FCEC part..... 143
Table A.6 Text links of the ICP part..... 144
Table A.7 Text links of the ICEG part..... 144
Table A.8 Text links of the TERM part..... 145
Table D.1 Text annotation links of the BMCS part..... 183
Table D.2 Text annotation links of the CCICE part.....183
Table D.3 Text annotation links of the EEP part..... 184
Table D.4 Text annotation links of the FICB part.....184
Table D.5 Text annotation links of the FCEC part.....184
Table D.6 Text annotation links of the ICP part..... 185
Table D.7 Text annotation links of the ICEG part..... 185
Table D.8 Text annotation links of the TERM part..... 186

List of Figures

Figure 1. Text segmentation with the RSTTool.....17
Figure 2. Discourse annotation with the RSTTool.....18
Figure 3. Segmentation annotation results as XML format..... 18
Figure 4. A segmented Chinese text by using rstWeb..... 19
Figure 5. An annotated Chinese text by using rstWeb.....19
Figure 6. Saved annotation result with rstWeb..... 20
Figure 7. An auto-fitted screenshot of RST analyses by using rstWeb.....20
Figure 8. CU of the annotated Spanish text (CCICE3_ESP)..... 25
Figure 9. CU of the annotated Chinese text (CCICE3_CHN)..... 25
Figure 10. The website of the RST Spanish-Chinese Treebank.....42
Figure 11. A parallel-segmented Spanish-Chinese text using RSTTool 43
Figure 12. A parallel-segmented Spanish-Chinese text using RSTTool 43
Figure 13. A segmented text in the website (Spanish text).....45
Figure 14. A segmented text in the website (Chinese text).....45
Figure 15. CU of the annotate Spanish text (CCICE3_ESP)..... 46
Figure 16. CU of the annotate Chinese text (CCICE3_CHN)..... 47
Figure 17. The annotate Spanish text by using the RSTTool.....49
Figure 18. The annotate Chinese text by using the RSTTool..... 49
Figure 19. Corpus consultation with different ways..... 50
Figure 20. Consultation of each selected relations.....51
Figure 21. Special case of discourse annotation for the qualitative comparison (Spanish text).....82
Figure 22. Special case of discourse annotation for the qualitative comparison (Chinese text)..... 82
Figure B.1.11 Case of Same-unit in the corpus (Spanish text)..... 153
Figure B.1.12 Case of Same-unit in the corpus (Chinese text)..... 153
Figure B.3.21 Example of SUMMARY relation in the corpus (Spanish text)..... 175
Figure B.3.22 Example of SUMMARY relation in the corpus (Chinese text)..... 175
Figure C.2.1 Partly annotation of the text FICB2 (Spanish part).....179
Figure C.2.2 Partly annotation of the text FICB2 (Chinese part).....180
Figure C.3.1 Partly annotation of the text ICP5 (Spanish part)..... 181
Figure C.3.2 Partly annotation of the text ICP5 (Chinese part)..... 181

Chapter 1

Introduction

In this chapter, we will explain the motivation to carry out this study firstly (Section 1.1). Next, we will present the objectives and Hypothesis (Section 1.2 & Section 1.3). Third, we will describe thesis structure (Section 1.4). Finally, we will show the publications of this study (Section 1.5).

1.1 Motivation

Spanish and Chinese are two of the most spoken languages in this world; the language pair occupies an important position in the Natural Language Processing (NLP) research world. Recently, discourse analysis has called much attention as an unsolved problem and is crucial for many NLP tasks (Zhou et al., 2014). The great language distance causes a great number of discourse differences between Spanish and Chinese (Cao, da Cunha and Bel, 2016). Comparative or contrastive studies of discourse structures reveal information to identify properly equivalent discourse elements in a language pair (Cao and Gete, 2018). Here we give an example to show the discourse similarity and difference between the two languages[1]:

Ex 1.
(1.1) Spanish: Aunque aún no contamos con resultados, intuimos que el modelo será más amplio que el del sintagma nominal.
[Aunque aún no contamos con resultados,]$Unit_1$ [intuimos que el modelo será más amplio que el del sintagma nominal.]$Unit_2$[2]
[Although still no get results,]$Unit_1$ [we consider that the model will be more extensive than the sentence group nominal.]$Unit_2$
(1.2) Spanish: Intuimos que el modelo será más amplio que el del sintagma nominal, aunque aún no contamos con resultados.
[Intuimos que el modelo será más amplio que el del sintagma nominal.]$Unit_1$ [aunque aún no contamos con resultados.]$Unit_2$
[We consider that the model will be more extensive than the sentence group nominal,]$Unit_1$ [although still no get results.]$Unit_2$
(1.3) Chinese: 尽管还没有取得最终结果，但是我们认为该模型已囊括了语段模型涉及的内容。
[尽管还没有取得最终结果，]$Unit_1$ [但是我们认为该模型已囊括了语段模型涉及的内容。]$Unit_2$

[1] The introduced example is an real example from our corpus.

[2] In our work, for all the examples we present, we also give an English literal translation of each example to make the readers get a better understanding of the examples in Spanish and Chinese.

[Although still no get results,]Unit$_1$ [but we consider that the model contains the sentence group nominal.]Unit$_2$

In Example 1, we can see that the Spanish passage (example 1.1) has a similar discourse structure to the Chinese passage (example 1.2). Both passages start the text with a discourse marker (DM)[3] in the first unit. However, the usage of discourse markers in both languages is different. To show the same meaning, in Chinese, it is mandatory to include two DMs: one marker is "*jinguan*" (尽管), at the beginning of the first unit, and another marker is "*danshi*" (但是), at the beginning of the second unit. Each of the two Chinese DMs is equivalent to the English discourse marker 'although'. By contrast, in Spanish, just one DM, "*aunque*" is being used at the beginning of the first unit, and also equivalents to 'although'. Moreover, the order of the discourse units in the Spanish passage can be changed and makes sense syntactically (example 1.3), but the order cannot be changed in the Chinese passage, because it does not make sense neither syntactically nor grammatically.

To get the discourse information for Spanish and Chinese (for instance, position of DM in each language, the order of the discourse units, relation between the discourse units, etc.), it is necessary to annotate the discourse information from a Spanish-Chinese parallel corpus. Following indications of Wu (2014), as a large electronic library, a corpus can provide a large amount of linguistic information. In addition, Johns (2002) considers that a corpus-based research could help the language learners get large amount of language information easily.

Regarding the method to get the above mentioned discourse information, the theory that especially designed for discourse analysis by Mann and Thompson (1988), Rhetorical Structure Theory (RST) will be the theoretical framework of this study. 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. These EDUs can be Nuclei or Satellites (Satellites offer additional information about Nuclei). The relations can be Nucleus-Satellite (e.g. Cause, Result, Concession, Antithesis) or Multinuclear (e.g. List, Contrast, Sequence). Moreover, under RST, the main information of a text called Central Unit (CU) can also be detected.

Although there are many works address with the topic of discourse analysis, few works talk about the discourse analysis for Spanish and Chinese (Cao, da Cunha and Iruskieta, 2017). The aims of the already existed works (Yao, 2008; Yang, 2008; Chien, 2012; Wang, 2013; Vargas-Urpi, 2018) for the discourse analysis between Spanish and Chinese are all for the language learning between the language pair. Yet, none of these works uses RST as the theoretical framework. The different discourse elements, such as discourse relation, order of segments, and discourse structure are not analyzed in these works.

Meanwhile, there are few works explore the discourse differences and similarities between the Spanish and Chinese under RST. To our knowledge, the only work that

[3] Discourse markers are the elements to signal relation between each text part. In Chapter 2, we will explain the definition of discourse markers in detail.

uses RST for Spanish and Chinese is the work of Cao, da Cunha and Bel (2016). The work explores the Chinese translation of the Spanish DM *aunque* in the the United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009), and the authors conclude different Chinese translations of Spanish DM *aunque*. However, this only work only concentrates on the single sentences contain the Spanish DM *aunque*, not the whole discourse structure of the text. For the Spanish-Chinese discourse analysis, there is still a gap to fill with.

Based on the presented example and mentioned related works, we can see that the translation for Spanish-Chinese from discourse level is not easy (order of segments, translation of DMs, discourse relations, etc.). In addition, in a current globalized context, translation between them is crucial between individuals and in language schools, institutions and enterprises, among other organizations. Many Spanish speakers are learning Chinese and many Chinese speakers are learning Spanish. With such frequent communications between two countries, an annotated parallel corpus with discourse information can help for translation and language learning purposes.

Therefore, this work is related with the following objectives and hypothesis.

1.2 Objectives

Based on the previous mentioned reasons, the objectives of this work are the following:

a. Main objective

To analyze the discourse differences and similarities in a Spanish-Chinese parallel corpus with the aim to use this information in tasks related to translation and language learning.

b. Specific objectives

b1. To create a Spanish-Chinese parallel corpus annotated with different discourse information in the framework of RST: discourse segments, CU, discourse relations, and discourse structure.

b2. To develop an online interface to search discourse-annotated information in the corpus.

b3. To compare the Spanish subcorpus with the Chinese subcorpus to detect the discourse differences and similarities between this language pair, and relate them with translation strategies.

b4. To analyze how discourse information is formally expressed in both languages in the parallel corpus, in terms of discourse relations, type of discourse relations (N-S or N-N), order of EDUs, and DMs.

b5. To elaborate a translation protocol with discourse recommendations for Spanish-Chinese translation, based on the comparison and analysis carried out in points b3 and b4.

b6. To design a Spanish-Chinese language learning task using the developed annotated parallel corpus.

1.3 Hypothesis

The hypotheses of this research are the following:

c1. Although Spanish and Chinese come from two different language families, there are discourse similarities between both languages in a parallel corpus.

c2. Discourse differences exist between Spanish and Chinese in a parallel corpus and they can be formally modelled in the framework of RST.

c3. The different discourse elements used in the framework of RST are adequate to formalize discourse equivalences between Spanish and Chinese.

c4. The use of a discourse-annotated Spanish-Chinese parallel corpus in the framework of RST would allow to obtain useful data for translation and language learning tasks.

1.4 Thesis Structure

Regarding the objectives and hypothesis of this work, the thesis consists of the following contents:

Chapter 1 Introduction. In Chapter 1, we give general information, the objective and the hypothesis of the thesis.

Chapter 2 Theoretical Framework. In Chapter 2, we introduce the theoretical framework. Additionally, we mention the related information with the framework, such as the annotation tools, and the possible applications with the framework.

Chapter 3 State of the Art. In Chapter 3, we analyze the related works under RST in terms of different discourse aspects. Besides, we compare the related works with our work. As the last part of Chapter 3, we talk about the corpus-based study for discourse analysis.

Chapter 4 Methodology. In Chapter 4, we talk about the methodology of this study. For instance, we describe the process to develop the research corpus. Then, we explain how to carry out each research step. We also give the evaluation method of this work in Chapter 4.

Chapter 5 Evaluation and Analysis. In Chapter 5, we evaluate the reliability of each research step, and give the qualitative analysis for the research steps.

Chapter 6 Elaboration of a Spanish-Chinese Discourse Recommendation Protocol for Translators. In Chapter 6, we give the translation protocol. In the first part of this chapter, we conclude the discourse similarities and differences based on the research results. In the second part of this chapter, we present the translation protocol with recommendations for Spanish-Chinese translation.

Chapter 7 A Spanish-Chinese Language Learning Task by Using Technological Corpus-based Resources. In Chapter 7, we discuss the relation between the language learning and our study. We explain how to carry out the language learning tasks between Spanish and Chinese through our study.

Chapter 8 Conclusions and Future Work. In Chapter 8, we conclude the work and look ahead of the future work.

Appendices. In this part, we will give the links of each text of the corpus. Moreover, we will give examples of the criteria of research step. The last part included in this part are are the special comparison cases for Spanish-Chinese discourse analysis.

1.5 Publications

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

- Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2018. The RST Spanish-Chinese Treebank. In *Proceedings of Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. In press.
- Cao Shuyuan, and Gete Harritxu. 2018. Using Discourse Information for Spanish-Chinese Language Learning. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'2018)*. 2254-2261.
- 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.
- Cao Shuyuan, da Cunha Iria, Iruskieta Mikel. 2017. Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *EPiC Series of Language and Linguistics*, 2: 315-324.
- Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2016. A Corpus-based Approach for Spanish-Chinese Language Learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications* (*NLP-TEA3*), 97-106.
- Cao Shuyuan, da Cunha Iria, and 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.
- Cao Shuyuan, da Cunha Iria, and Bel Nuria. 2016. An analysis of the Concession relation based on the Spanish discourse marker *aunque* in a Spanish-Chinese parallel corpus. *Procesamiento del Lenguaje Natural*, 56: 81-88.
- Cao Shuyuan. 2015. Elaboration of a protocol to support Chinese-Spanish translation: an approach based on a parallel corpus annotated with discourse information. In *Proceedings of XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

Chapter 2
Theoretical Framework

In Chapter 2, we will introduce the theoretical framework, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In the first section (Section 2.1) of this chapter, we will give a general introduction of the RST. In the second section (Section 2.2) of this chapter, we will discuss the concept of discourse markers (DM) that related with the theoretical framework. In the third section (Section 2.3) of this chapter, we will discuss the annotation tools related with the theoretical framework, two RST annotation interfaces called RSTTool and rstWeb. In the fourth section (Section 2.4) of this chapter, we will talk about different RST applications.

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[4].The relations 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), based on a specific kind of resemblance.

Circumstance	Antithesis
Solutionhood	Concession
Elaboration	Condition
Background	Otherwise
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. Under this classification, subject matter relations intend to make the reader recognizes the relation in question

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

meanwhile presentational relations to make the reader increases the acceptance of the nucleus.

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

Along the RST studies, the number of relations is not decided. The above mentioned 23 relations in Table 2 by Mann and Thompson are the original ones for RST study. Afterwards an extended version of the list has been provided at RST website[5], totally, 30 relations are included in the RST website. Depending on the research purpose, different RST studies cite different number of relations. For instance: (i) Huong (2007) (22 relations), (ii) Pardo and Nunes (2008) (32 relations), (iii) Maziero et al. (2011) (32 relations), (iv) Carlson, Marcu and Okurowski (2002) (78 relations are divided into 16 relation groups), etc.

The cause of various versions of the RST relations is the features in RST itself. Mann and Thompson (1988) consider that the inventory of discourse relations is an open set. Moreover, Mann, Matthiessen and Thompson (1989: 48) indicate that, RST is more than a strict theory of discourse with limited attributes, relevant modifications within the RST taxonomy should be allowed:

> *Relation definitions have the status of applications of the theory rather than elements of the theory. One might want to change or replace the definitions...such changes are to be expected and do not cross the definitional boundaries of RST.*

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

Apart from the RST, there are two other methods that have been widely used. One is the discourse theory, Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), and the other one is a corpus based approach called The Penn Discourse Treebank (PDTB) (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 form 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) indicates 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, etc.

On the other hand, 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[6] shows how the discourse relation and their arguments are annotated:

(Ex.2) 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.

The PDTB is also the unique one to adopt a theory-neutral approach to the annotation, this approach guarantees make no commitments to what kinds of high-level structures could be created from the low level annotations of relations and their arguments (Prasad et al., 2008). Besides, this approach allows the corpus to be useful for studies address with different frameworks while at the same time providing a resource to validate the various existing theories of discourse structure (Mann and Thompson, 1988; Wolf and Gibson, 2005). In the extended version PDTB 2.0 (Prasad et al., 2008), the sense annotation and the attributions associated with the relation and arguments are also annotated, as the following examples show[7]:

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

[7] Example cited from Prasad et al. (2008).

(Ex.3) Annotation: *The Mountain View, Calif., company has been receiving 1,000 calls a day about the product* since **it was demonstrated at a computer publishing conference several weeks ago**.

(Ex.4) Annotation: *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand**.

(Ex.5) Annotation: *Domestic car sales have plunged 19%* since **the Big Three ended many of their programs Sept. 30.**

The above three examples represent three senses of the word 'since', 'Temporal' in Example 4, 'Casual' in Example 5 and a third both 'Casual' and 'Temporal' in Example 6. The PDTB can be used for different NLP applications, such as parsing (Prasad, Joshi and Webber, 2010; Stepanov and Riccardi, 2014), information retrieval (Hiong, Kulathuramaiyer and Labadin, 2012), machine translation (MT) (Meyer and Polakova, 2013; Li, Carpuat and Nenkove, 2014), etc.

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. In this way, translation strategies (if there are any) can be detected in different levels of an RS-tree (da Cunha and Iruskieta, 2010; Iruskieta, da Cunha and Taboada 2015).

2.2 Definition of DMs

As one of the discourse elements under RST, DM has called much attention for NLP studies (Cao and Gete, 2018). Nowadays, DM is becoming a very popular topic for the academic world of NLP. Some works have talked about the DMs under the RST for different languages. For instance, for English, by using two different corpora, Taboada (2006) discusses the relationship between the DMs and rhetorical relations. For Spanish, da Cunha (2013) talks about the disambiguation of DMs in Spanish. For Russian, Toldova et al. (2017) detect DMs to see their frequency in Russian news. Studies that address DMs between language pairs under the RST also exist. For example, Li, Langlais and Jin (2017) describe the cross-lingual and alignment of DMs in a Chinese-English parallel corpus to describe related surveys and findings for MT between the two languages. By contrasting the similarities and differences between the English DMs and French DMs, Meyer and Popescu-Belis (2012) show how the automatic discourse disambiguation can improve the statistical machine translation (SMT) between English and French.

With the development of the NLP researches, it is impossible to list all the works that address the topic of DMs. Our intention to list the DM related works here is to show that DM is an important element for RST study. However, the definition of DMs is not easy (Taboada, 2006). Various works gives different definitions of DMs from different perspectives.

For a general definition, 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. Meanwhile, Schiffrin (2001: 54) indicates: "Discourse markers (DMs) involve linguistic items that in cognitive, expressive, social and textual domains." Guo (2015: 70) gives a related definition: "DMs are deemed to be a complex phenomenon which involves textual, pragmatic and linguocognitive variables."

From the point of view of semantics, Mosegaard (1998: 236) defines DMs as:

> *[...] linguistic items which fulfil a noun-prepositional, metadiscursive (primarily connectives) function, and whose scope is inherently variable, such that it may comprise both sub-sentential and supra-sentential units.*

From a macro-semantic perspective, Fløttum (2002) considers that, DMs are textual means that they contribute to the polyphony of voices in the narrative. Zhang (2016) shows a similar definition of DMs, together with metadiscursive phenomenon, that the two elements frame the overall presentation in written texts and facilitate the reader's guidance throughout the narrative.

From the textual level, DMs, according to Fraser (1999: 938): "impose a relationship between some aspect of the discourse segment, they are part of, call it S2, and some aspect of a prior discourse segment, call it S1." The definition from Eckle-Kohler, Kluge and Gurevych (2015) is that, from textual level, DMs are used to signal discourse relations in a text segment, as "cohesive relationships between the utterances" (Müller, 2005: 1). For discourse analysis, Das (2014: 41) concludes the following conditions of DM:

> *(i) a DM should be syntactically detachable from a sentence; (ii) DMs should usually be used in the beginning of an utterance; (iii) a DM should contain a range of prosodic features; (iv) DMs should be able to operate at local as well as global levels of discourse; and (v) DMs should be able to function at different planes of discourse (exchange structure, action structure, ideational structure, etc.).*

The topic of this study falls on the discourse comparative between Spanish and Chinese, for this reason, we follow the definition of Eckle-Kohler, Kluge and Gurevych (2015), as the DMs in the corpus are used to signal discourse relations in a text segment. Specifically, as da Cunha (2013) indicates, there are three types of DM: (i) Traditional discourse markers, (ii) Markers including lexical units, specifically, nouns and verbs, and (iii) Markers including verbal structures. For our work, we use the concept of traditional markers and markers including verbal structures[8].

[8] For traditional markers, for instance, *debido a / youyu* (由于), 'due to' in English. For markers including verbal structures, for instance *para* + infinitive / *weile* (为了) + infinitive, which means 'to' + infinitive.

2.3 RST Annotation Tools

At the moment, there are two annotation approaches for RST tasks. One is the RSTTool (O'Donnell, 2000) and another one is a newly released online annotation interface named rstWeb (Zeldes, 2016). As we use both annotation approaches for our study[9], therefore, in this section, we introduce the two annotation tools in detail.

2.3.1 RSTTool

The RSTTool[10] (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 various versions for different computer systems. In our work, we use the Mac version of the RSTTool to carry out the study.

The RSTTool is the first annotation interface for discourse annotation underthe RST. The annotation steps are: (a) segmentation and (b) discourse relations annotation. Figure 1 shows a segmented Chinese text[11] with the RSTTool and Figure 2 shows a completely annotated Chinese text by using the tool.

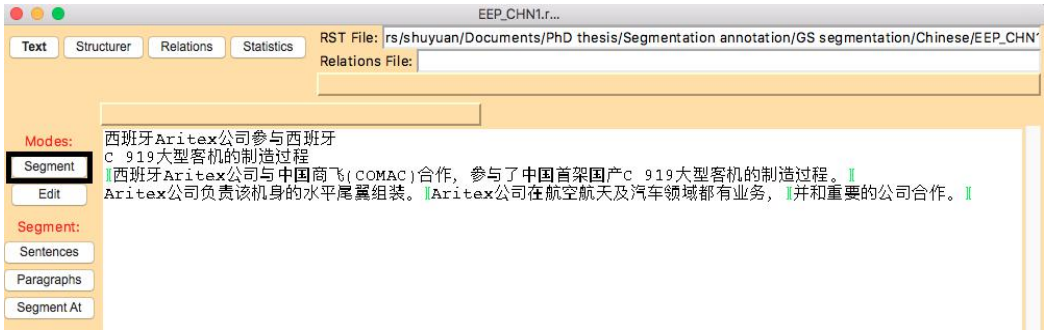

Figure 1. Text segmentation with the RSTTool

---

[9] We will give the detailed annotation information in Chapter 4.

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

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

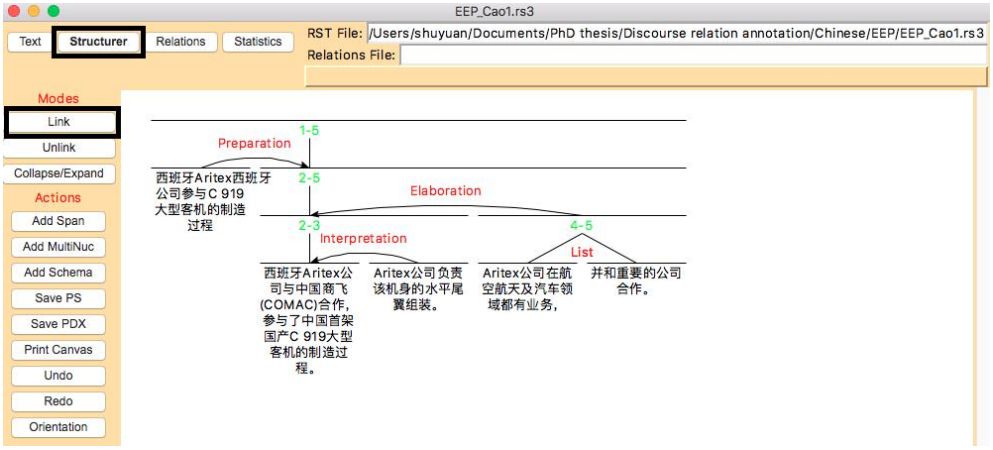

Figure 2. Discourse annotation with the RSTTool

The RSTTool saves the annotation results in 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>
```

Figure 3. Segmentation annotation results as XML format

2.3.2 rstWeb

rstWeb (Zeldes, 2016)[12] is a newly released browser based interface for RST annotations. rstWeb supports multiple annotated versions of each document,

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

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.

Figure 4. A segmented Chinese text by using rstWeb

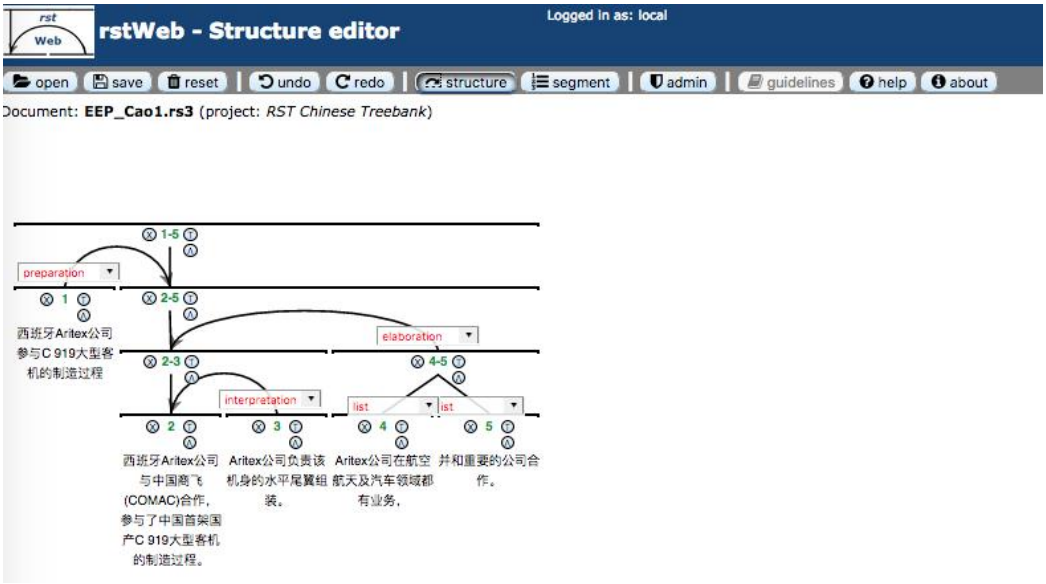

Figure 5. An annotated Chinese text by using rstWeb

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

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

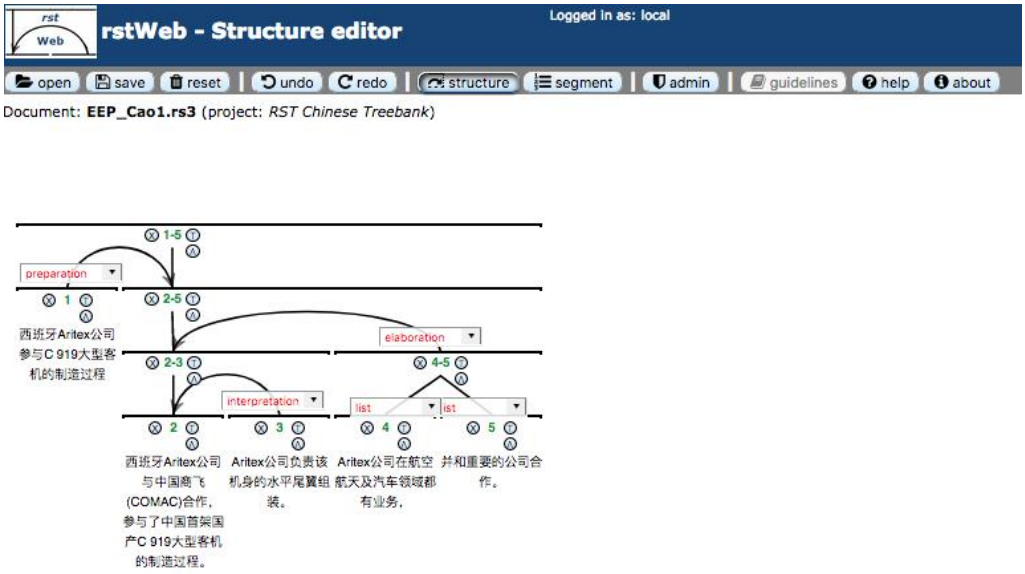

Figure 7. An auto-fitted screenshot of RST analyses by using rstWeb

2.4 RST Applications

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 analyzing 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 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).

- Information extraction (IE)

Information extraction (IE) is the task of automatically extracting 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 patent descriptions in order to extract claims in Japanese patents. Li (2010) presents a system that automatically extracts the rhetorical structure of a text to make summarizations with RST. Da Cunha (2008), da Cunha, Wanner, and Cabré (2007) work on the automatic summarization for Spanish medical texts.

- Machine translation (MT)

Machine translation (MT) explores the use of software to translate text or speech from one language to another. RST has been applied on the evaluation of MT by Fomicheva, da Cunha and Sierra (2012). 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 models, so as to maintain the semantically functional integrity and hierarchical relations of EDUs during translating.

Regarding the RST application of our work, as explained in Chapter 1, our work concentrates on the development of the language resources to help Spanish-Chinese translation from discourse level. Moreover, our another intention is to make this study serves for the language learning between Spanish and Chinese.

2.5 Chapter Overview

In this chapter, we have introduced the theoretical framework of this study, the Rhetorical Structure Theory (RST). RST is a functional theory of text organization. It describes how a text is made of smaller parts, and how the smaller parts are connected to each other in terms of certain organizational patterns. Moreover, we have also mentioned two other approaches for discourse analysis, Segmented Discourse Representation Theory (SDRT) and Penn Discourse Treebank (PDTB). Comparing to SDRT and PDTB, RST has been selected as the theoretical framework as it can

describe the whole discourse structure of a text by using inter-annotation style and intra-annotation style.

The second part of this chapter introduces the concept of DMs. Many studies have discussed about the definition of DMs, we confirm the definition of DMs for this study under the RST, which are traditional discourse markers, markers including lexical units, and markers including verbal structures. Thirdly, we have introduced two annotation interfaces under RST, which are RSTTool and rstWeb. The RSTTool is the first annotation interface for RST analysis and the rstWeb is a recently released annotation interface. Lastly, we have outlined the RST related studies for NLP applications, as an overview of the RST development. We also set forth the relation between RST and this PhD study, which aims to help Spanish-Chinese translation and language learning from discourse level.

Chapter 3
State of the Art

In this chapter, we will first talk about the related works for RST annotation: segmentation annotation, Central Unit (CU) annotation and discourse structure (Section 3.1). Secondly, we will conclude discourse annotated corpora for Spanish and Chinese (Section 3.2 and Section 3.3). Thirdly, we will analyze the comparative studies by using RST (Section 3.4). After, some corpus-based studies for language learning will be presented (Section 3.5). Lastly, a summary of this chapter will be added (Section 3.6).

3.1 RST Annotation

Linguistic information in a corpus is a crucial element for NLP studies. Therefore, to extract the linguistic information from a corpus is essential for corpus studies. As Leech (1997: 2) defines, the process of "adding such interpretative, linguistic information to an electronic corpus of spoken and/or written language data" is defined of corpus annotation. A similar opinion is confirmed by McEnery, Xiao and Tono (2006: 29-30): "corpus annotation is concerned with interpretative linguistic information", and "adds value to the corpus". Later, McEnery and Hardie (2012: 13) also confirms the definition of the corpus annotation:

> *[...] it is important to note that, setting scale aside, corpus annotation is largely the process of providing - in a systematic and accessible form - those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they worked with..*

As an theory that especially designed for discourse analysis, the RST allows us to find various discourse elements by annotating the corpus. In this study, to reflect the discourse differences between the Spanish and Chinese, we will annotate the following discourse elements: (i) segmentation, (ii) CU, and (iii) discourse structure. This section reviews the already existed works related with segmentation annotation, CU annotation and discourse structure.

3.1.1 Segmentation

With the development of the NLP community, many studies have established their own segmentation criteria for different research purposes.

- RST-based discourse segmentation

Several corpora for different languages have established their own segmentation criteria for different discourse analysis tasks under the RST: (i) for English, the RST Discourse Treebank (Carlson, Marcu and Okurowski, 2001), the Discourse Relations Reference Corpus (Taboada and Renkema, 2008), and the RST Signalling Corpus (Das and Taboada, 2017); (ii) for German, the Potsdam Commentary Corpus (Stede

and Neumann, 2014); (iii) for Spanish, the RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011); (iv) for Basque, the RST Basque Treebank (Iruskieta et al., 2013); (v) for Portuguese, the CorpusTCC (Pardo, Nunes and Rino, 2008) and Rhetalho (Pardo and Seno, 2005); (vi) for Russian, the Russian RST Treebank (Toldova et al., 2017); (vii) for Basque and Spanish, the RST Basque-Spanish DELIB Treebank (Imaz and Iruskieta, 2017); and (viii) for Spanish, Basque and English, the Multilingual RST Treebank (Iruskieta, da Cunha and Taboada, 2015).

In addition, some available discourse segmentation systems based on RST exist. For example: i) for English (Tofiloski, Brooke and Taboada, 2009), ii) for Spanish (da Cunha et al., 2012), iii) for Basque (Iruskieta and Zapirain, 2015), iv) for Brazilian Portuguese (Pardo and Nunes, 2008; Cardoso, Pardo and Taboada, 2017), and (v) for Brazilian Portuguese and Spanish (Maziero et al., 2011).

- Discourse segmentation for Chinese

Few works focus on the Chinese segmentation from the discourse level. The Penn Chinese Treebank (Xue, 2005) is especially designed for Chinese discourse analysis with the Penn Discourse TreeBank (PDTB) (Miltsakaki et al. 2004) style. In this work, segmentation criteria are based on connectives and different types of conjunctions. Moreover, Zhou and Xue (2012) describe a discourse annotation scheme for Chinese and report on the preliminary results. In this work, following the PDTB-style annotation guideline, they segment the Chinese sentences in the corpus to annotate the explicit and implicit discourse relations.

Under RST, there are three works that utilize punctuation marks to elaborate segmentation rules for Chinese (Yue, 2006; Qiu, 2010; Li, Feng and Zhou 2013).

There are three other notable works related to Chinese discourse segmentation (Xue and Yang, 2011; Yang and Xue, 2012; Xu and Li, 2013), which focus on the influence of the comma for Chinese segmentation.

- Discourse segmentation for Spanish

Several works focus on the segmentation for Spanish, as previously presented before, the RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011), the Multilingual RST Treebank (Iruskieta, da Cunha and Taboada, 2015), and the RST Basque-Spanish DELIB Treebank.

In addition, some automatic segmentation systems also establish segmentation criteria for Spanish. For instance, based on lexical and syntactic rules, by using RST, da Cunha et al. (2012) develop the first system for Spanish segmentation. The Dizer 2.0 (Maziero et al., 2011) is an adaptable on-line discourse parser that can also be used for Spanish segmentation.

3.1.2 CU Annotation

Under RST, for each segmented text, among the EDUs, there is an EDU called 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 and sentiment analysis (Iruskieta, Labaka and Desiderato, 2016). 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.

Figure 8 presents the CU of the annotated Spanish text in the corpus and Figure 9 shows the CU of its parallel annotated Chinese text.

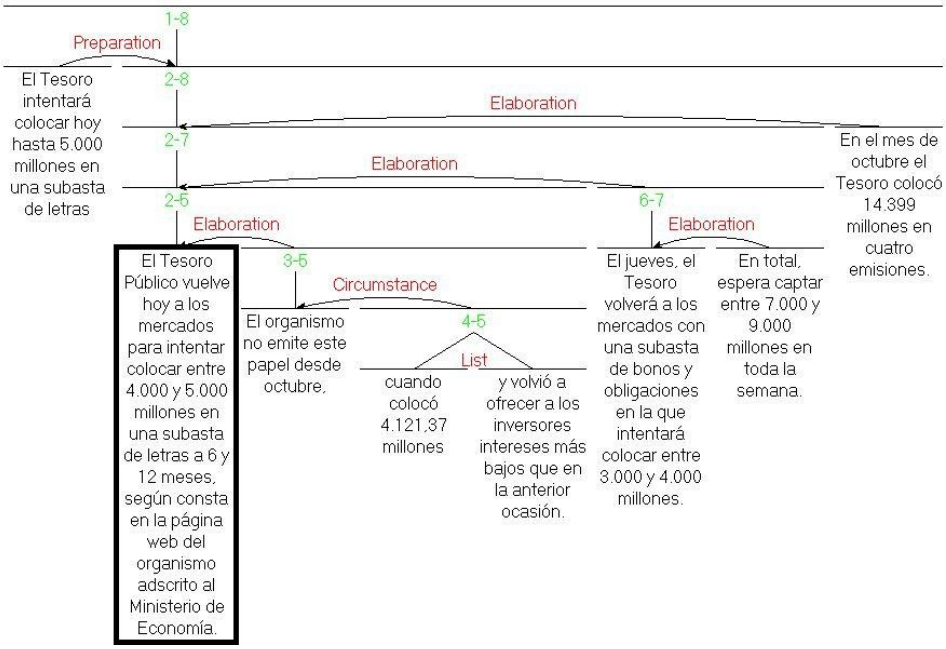

Figure 8. CU of the annotated Spanish text (CCICE3_ESP)

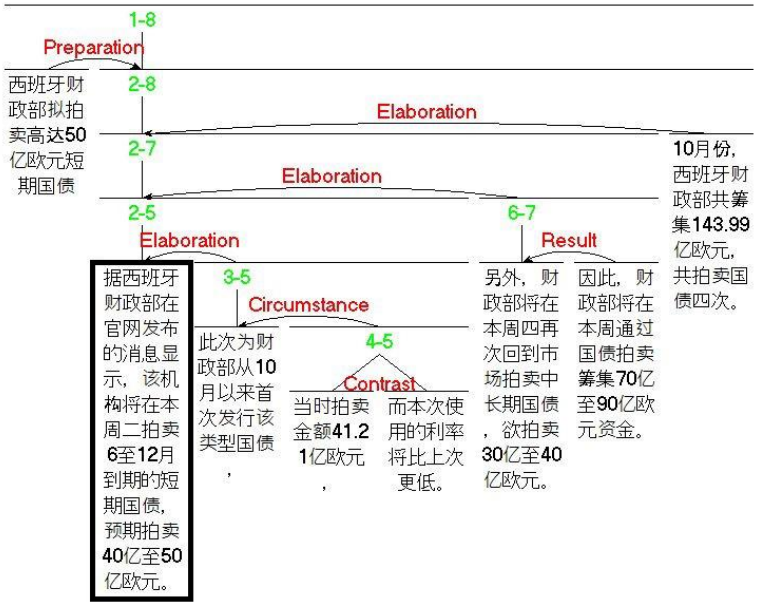

Figure 9. CU of the annotated Chinese text (CCICE3_CHN)

Figure 8 shows that, for the annotated Spanish text, all the arrows are point to EDU2, which means the content of "*El Tesoro Público vuelve hoy a los mercados para intentar colocar entre 4.000 y 5.000 millones en una subasta de letras a 6 y 12 meses, según consta en la página web del organismo adscrito al Ministerio de*

Economía.[13]" is the main information of the Spanish text. In Figure 9, for the parallel 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。)[14]".

Several studies focus on using CU with RST. For instance, the RST Basque Treebank includes the CUs for each text. For example, Iruskieta, Díaz de IIarraza and Lersundi (2014) analyze how agreement regarding the CU affects agreement when settling discourse structure.

Some automatic annotation systems are related to the CU research. A rule-based system for detecting CUs in Basque scientific abstracts is created by Iruskieta et al. (2015). Using machine learning-based techniques, Bengoetxea, Atutxa and Iruskieta (2017) develop a new system that can automatically annotate the CUs in Basque scientific abstracts[15]. Additionally, another rule-based system for annotating main ideas by Iruskieta, Labaka and Antonio (2016) has been designed for Brazilian Portuguese argumentative answer texts and Basque scientific texts[16]. The work that newly released for detecting the CU for Spanish is the work of Bengoetxea and Iruskieta (2018).

3.1.3 Discourse Structure Annotation

Several corpora for different languages have been annotated under RST, as following shows:

(i) English

The best known-annotated RST corpus for English is RST Discourse Treebank (Carlson, Marcu and Okurowski, 2001)[17]. In total, 385 journalistic texts are selected. The topics of the texts are culture, economy, editorials among others.

The Discourse Relations Reference Corpus (Taboada and Renkema, 2008)[18] 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.

---

[13] English literal translation: The Public Treasury returns today to the markets to try to place between 4,000 and 5,000 million in an auction of letters to 6 and 12 months, according to the web page of the organized ascribed to the Ministry of Economy.

[14] 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.

[15] The CU detector for Basque can be tested at `http://ixa2.si.ehu.es/CU-detector`.

[16] For Brazilian Portuguese: `http://ixa2.si.ehu.es/CU-detector/` and for Spanish: `http://ixa2.si.ehu.es/clarink/tools/ES-CU-detector/`.

[17] `https://catalog.ldc.upenn.edu/LDC2002T07` [Last consulted: 10 of January of 2018]

[18] `http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html` [Last consulted: 10 of January of 2018]

The RST Signalling Corpus (Das and Taboada, 2017)[19] is a corpus annotated for signals of coherence relations for English. The original source of the corpus comes from the RST Discourse Treebank. Based on the RST Discourse Treebank, authors add relevant signalling information.

(ii) German

The corpus for German by using RST is the The Potsdam Commentary Corpus (Stede and Neumann, 2014)[20]. The corpus includes 220 German newspaper 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 that takes RST as framework for Spanish is The RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011)[21]. The corpus contains 267 texts and 52,746 words. The texts in this corpus are 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)[22]. This corpus is a public corpus for Basque NLP tasks. It includes abstracts from three specialized domains: medicine, terminology and science. 88 documents have been selected for 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)[23]. The CorpusTCC is built for the detection of linguistic patterns and indication of rhetorical relations. This corpus contains 100 Brazilian Portuguese scientific texts with a total of 53,000 words. The topics of the corpus are related to computer science. Rhetalho is a corpus designed for parser evaluation and it consists of 40 texts, 20 from the 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 is designed for the Russian discourse analysis by Toldova et al. (2017)[24]. The corpus aims to annotate texts of four genres and domains: science, popular science, news stories, and analytic journalism. Currently, 73 annotated 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.

---

[19] `http://catalog.ldc.upenn/LDC2015T10` [Last consulted: 10 of January of 2018]

[20] `http://angcl.ling.uni-potsdam.de/resources/pcc.html` [Last consulted: 10 of January of 2018]

[21] `http://corpus.iingen.unam.mx/rst/citar.html` [Last consulted: 10 of January of 2018]

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

[23] `http://www.icmc.usp.br/~taspardo/projects.htm` [Last consulted: 10 of January of 2018]

[24] `https://github.com/nasedkinav/rst_corpus_rus` [Last consulted: 10 of January of 2018]

(vii) Basque and Spanish

The RST Basque-Spanish DELIB Treebank (Imaz and Iruskieta, 2017)[25] is an annotated bilingual RST corpus for Basque and Spanish. The corpus is an extended version of RST Basque Treebank. 100 texts in Basque and their parallel Spanish texts are included in this corpus. The corpus involves 8900 words for the Basque subcorpus and 11166 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)[26].The corpus includes 45 texts (15 texts for each language), with the English subcorpus containing 5,706 words, the Spanish subcorpus containing 6,324 words and the Basque subcorpus containing 4,800 words. The main topic of this corpus is terminology research.

Table 3 contains the information of each treebank, including the treebank information of our study.

[25] http://ixa2.si.ehu.es/diskurtsoa/rstfilo/index.php [Last consulted: 10 of January of 2018]

[26] http://ixa2.si.ehu.es/rst/ [Last consulted: 10 of January of 2018]

Corpus Name	Language	N° of texts	Topics	Genres
RST Discourse Treebank	English	385	culture, economy, editorials	journalistic texts
The Discourse Relations Reference Corpus	English	65	journal articles, advocacy letter and review texts	journal articles, advocacy letter review texts
The RST Signalling Corpus	English	385	culture, economy, editorials, etc.	journalistic texts
The Potsdam Commentary Corpus	German	220	politics	newspaper commentaries
The RST Spanish Treebank	Spanish	267	astrophysics, earthquakes, engineering, economy, linguistics, medicine, psychological and sexuality	scientific articles, conference papers, articles, reports
The RST Basque Treebank	Basque	88	medicine, terminology and science	abstracts
The CorpusTCC	Portuguese	100	computer science	scientific texts
Rhetalho	Portuguese	40	computer science	scientific paper and news
The Russian RST Treebank	Russian	73	science, popular science, news stories, and analytic journalism	news stories
The RST Basque-Spanish DELIB Treebank	Basque and Spanish	100	society and politics	exercises
The Multilingual RST Treebank	English, Spanish and Basque	45	terminology	scientific abstracts
The RST Spanish-Chinese Treebank[27]	Spanish and Chinese	100	terminology, language, culture, education, art, etc.	abstract, news, advertisements, announcements

Table 3. Detailed information of each RST Treebank

[27] The RST Spanish-Chinese Treebank is the name of our corpus. For the comparison between each RST treebank, we give a shallow information of our corpus here. Also the shallow information will be given in the following section (Section 3.4). The detailed information of the corpus can be consulted in the next chapter (Chapter 4).

3.2 Corpus-based Discourse Analysis for Spanish

The amount of discourse analysis research for Spanish under RST is still few. As we presented in the previous section (Section 3.2), there are currently only three works that build treebanks for Spanish using RST; one is The RST Spanish Treebank, the others are The Multilingual Treebank and The RST Basque-Spanish DELIB Treebank.

As mentioned before, the three corpora are accessible and free to the public. Users can consult the texts, EDUs, and discourse relations of the corpora. Although each study contains its own research purpose, the RST treebanks for Spanish are great contributions for Spanish NLP researches.

3.3 Corpus-based Discourse Analysis for Chinese

One of the earlier Chinese discourse analysis is the Penn Chinese Discourse Treebank (CDTB) (Xue et al., 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[28]. 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 only partly accessible. 3,007 text files can be consulted. The texts of this corpus are mainly taken from newswire, magazine articles and government documents. The topics of the corpus are various, such as general politics, culture, economy, travel, etc.

The Sinica Treebank is created by Huang et al. (2014)[29]. 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). The genres of the corpus are different, for instance, report, announcement, meeting record, advertisment, etc. The topics of the texts are different, for instance, politics, traveling, sports, society, etc.

The Discourse Treebank for Chinese is another project for Chinese discourse analysis and was created by Zhou et al (2014)[30]. 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 three related works so far. Yue (2006) creates the Caijingpinglun Corpus (CJPL) under RST. The CJPL

---

[28] 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]

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

[30] 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.

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. Additionally, as we mentioned in the Section 3.1, Guo (2004) annotates the discourse structure for English, Japanese and Chinese. The corpus consists of 18 editorials (six for each language) from *Los Angeles Times*, *Yomiuri Shimbun* and *Guangming Daily*. However, some limitations exist for the three works. Firstly, none of the works is available to the public[31]. Secondly, for three corpora, the source, the genre and the topic of the texts are simple. Especially the work of Guo (2004), there is no infomation about the topics of the corpus. The two 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 (Cao, da Cunha and Iruskieta, 2017). Thirdly, the authorization of the texts. Authors donot mention if they have permission to use the texts for their studies. For our work, we have asked for the permission of the usage of each text. 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 Guo (2004) contains 18 texts meanwhile the corpus of Qiu (2010) contains 10 annotated texts. Lastly, none of the works mention the evaluation of the annotation quality. Table 4 summarizes the information of all the above mentioned Chinese discourse analysis. Our study is also included in Table 4.

[31] The work of Yue (2006), we wrote to her requesting the related information, but she have not sent a response. For the work of Qiu (2010), we cannot find the contact information, neither the information of the supervisor. For the work of Guo (2004), we cannot find any contact information either.

Treebank	Framework	Nº of annotated texts	Genre	Domain	Parallel corpus	Accessible
Penn Chinese Discourse Treebank	PDTB	3,007	newswire, magazine articles and government documents	general politics, culture, economy, travel, etc.	No (Chinese only)	Yes
The Sinica Treebank	PDTB	61,087	report, announcement, meeting record, advertisement, etc.	politics, traveling, sports, society, etc.	No (Chinese only)	Yes
The Discourse Treebank for Chinese	PDTB	890	newswire, magazine articles and government documents	general politics, culture, economy, travel, etc.	No (Chinese only)	No
Caijingpinglun Corpus	RST	90	financial news commentaries	finance	No (Chinese only)	No
Qiu (2010)[32]	RST	10	news	economic, policy, society, etc.	No (Chinese only)	No
Guo (2004)[33]	RST	18	editorials	not mentioned	Yes (English-Japanese -Chinese)	No
The RST Spanish-Chinese Treebank	RST	100	abstract, news, advertisements, announcements	terminology, language, culture, education, art, etc.	Yes (Spanish-Chinese)	Yes

Table 4. Summary of each discourse treebank for Chinese

[32] The author does not give any name of the corpus.
[33] The author does not give any name of the corpus.

3.4 RST-based Comparative Studies

Thus far, there have not been many studies addressing contrastive discourse analysis with RST and less between Spanish and Chinese. Within the few that exist, there exist some comparative studies between Chinese and English that employ RST. Cui (1986) presents some aspects regarding discourse relations between Chinese and English; Kong (1998) compares Chinese and English business letters; Guy (2000, 2001) compares Chinese and English journalistic news texts.

Other studies with RST examine pairs of languages such as Japanese and Spanish (Kumpf, 1986; Marcu et al., 2000), Arabic and English (Mohamed and Omer, 1999), French and English (Delin et al., 1996; Salkie and Oates, 1999), Dutch and English (Abelen et al., 1993), Finnish and English (Sarjala, 1994), Spanish and Basque (da Cunha and Iruskieta, 2010), Spanish and Chinese (Cao, da Cunha and Bel, 2016), Spanish and Basque (Imaz and Iruskieta, 2017).

RST contrastive studies that use more than two languages are not common; those that have include work on Portuguese-French-English (Scott, Delin and Hartley, 1998). In their work, a methodology has been presented for RST contrastive analysis alongside the empirical cross-lingual results. Taking 18 editorials (6 or each language), Guo (2014) detects the discourse similarities and differences between English, Japanese and Chinese under RST, and gives some suggestions for language teaching and learning between the three languages. Iruskieta, da Cunha and Taboada (2015) use RST as theoretical framework to compare Basque, Spanish and English, so as to create a new qualitative method for the comparison of rhetorical structures in different languages and to specify why the rhetorical structure may be different in translated texts[34].

3.5 Language Learning Using Corpus-based Approach

Discourse information can benefit language learning between the language pair as, "it has been demonstrated that discourse is a crucial aspect for L2[35] learners of a language, especially at more advanced level" (Neff-van Aertselaer, 2015: 255). Corpus-based studies for different language pairs learning exist, including some works on Spanish and Chinese. For example, we highlight the following corpus-based language learning studies:

i) In order to help language learning and translation tasks between English and Chinese, Qian (2005) created an English-Chinese parallel corpus with functions of sentence search, calculation of words, search of texts and authors.

ii) To compare the similarities and differences between English and Chinese from different aspects, such as aspect marking, temporal adverbials, passive construction, among other interesting topics, Xiao and McEnery (2010) used the FLOB corpus (Albert-Ludwigs, 2007)[36] and The Lancaster Corpus of Mandarin Chinese (LCMC)

---

[34] http://ixa2.si.ehu.es/rst/ [Last consulted: 10 of January of 2018].

[35] L2 means second language.

[36] http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/ [Last consulted: 27 of July of 2016].

(McEnery and Xiao, 2004)[37], which is designed as a Chinese parallel corpus for FLOB. The study offers a great amount of language information that is useful for English-Chinese language learning, for instnce, lexical information, discourse information, grammar analysis, etc.

iii) To compare both languages via different language activities, such as exploration of language differences, comparative discourse analysis and semantic analysis, Lavid, Arús and Zamorano (2010) developed a small online English-Spanish parallel corpus. Then, based on the activity results, they give some linguistic suggestions for English-Spanish teaching, which can also help English-Spanish language learners to comprehend the language differences between both languages.

Meanwhile, corpus-based studies for Spanish-Chinese language learning are still few:

i) Yao (2008) uses film dialogues to create an annotated corpus and compares Spanish and Chinese discourse markers in order to give some suggestions for teaching and learning Spanish and Chinese.

ii) Yang (2008) compares the discourse structure of proverbs between Spanish and Chinese based on the novel Don Quijote in order to draw some conclusions for the Spanish-Chinese translation works, and language teaching and learning tasks.

iii) Taking different newspapers and books as their research corpus, Chien (2012) compares Spanish and Chinese conditional discourse markers to draw some conclusions about the conditional discourse marker for foreign language teaching between Spanish and Chinese.

iv) Wang (2013) uses Pedro Almodóvar's films *La mala educación* and *Volver* as their corpora to analyze how the subtitled Spanish discourse markers can be translated into Chinese, so as to make a guideline for human translation and audiovisual translation between the language pair.

v) Vargas-Urpi (2018) analyzes court interpreting from Chinese to Spanish based on a recording of a criminal trial. The analysis focuses on examples of errors of interpretation, speech style, and non-renditions, and can be useful for Chinese-Spanish interpreting study.

The above mentioned works are great achievements that offer different approaches for language learning. However, comparing to our work, none of them gives a friendly environment to consult Spanish-Chinese parallel corpus based on annotated discourse information, showing how foreign language learners can apply this information to improve or learn languages.

Regarding the RST-based studies for language learning by using corpus, for instance, for English and Chinese, by annotating Chinese students' and native speakers' compositions of the same topic under RST, Zhang (2010) describes and compares the rhetorical structure diagrams of these compositions from the perspective of the amount, frequency and distribution of each relation to help teachers to explore the deficiencies of Chinese students' compositions. By using news texts on *China Daily* and *The New York Times*, Fang (2008) explores the discourse features of English that were expressed by Chinese native speakers by means of RST. The study

[37] http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/ [Last consulted: 27 of July of 2016].

helps English foreign language learners acquire a better understanding of Chinese style English. In order to help Chinese students' argumentative writing in English, Li and Liao (2015) take RST as their theoretical framework to explore the different features within 60 English essays written by Chinese students. Aside from the English-Chinese language pair, there is one work focuses on the language education between Chinese and Korean and takes RST as its framework. Liang and Yang (2016) use the spoken data of Korean students and Chinese native speakers to reveal the differences in their use of causal and transitional markers, and analyse the typical errors under RST. Finally, they give some suggestions for Korean-Chinese speaking teaching. Cao, da Cunha and Bel (2016) annotate all the cases of the Spanish DM *aunque* ('although') and their corresponding Chinese translations in The United Nations Multilingual Corpus (UN). They analyze the translation strategies used in the translation process and give some suggestions for how to translate this Spanish DM into Chinese. Moreover, as we indicated in Section 3.2, the work of Guo (2014) compares the discourse structure between English, Japanese and Chinese under RST with foreign language teaching and learning purpose between the three languages.

Regarding the exercise generation aspect, some successful studies have been applied to education using different approaches. For example, in order address the challenge of automatically generating questions from reading materials for educational practise and assessment, Heilman and Smith (2010) create a statistical rule-based system to rank the output of a "wh-" question generation system. Under the situation module, Chen, Aist and Mostow (2009) test the generality of their question generation approach by extending the approach to informational text. Moreover, discourse information has also been used in their study. Another approach that can be used for question generation is the concept map. Olney, Graesser and Person (2012) erase the gap between psychological theories of question asking and computational models of question generation by computing conceptual graphs.

To our knowledge, our work is the first one to use RST for Spanish-Chinese language education, and contains the question generation function.

3.6 Chapter Overview

In this chapter, we have introduced the related works. Firstly, we review the works that are related with RST segmentation annotation, CU annotation and discourse structure annotation. The segmentation has been applied to different languages, such as English, Spanish, Basque, etc. To our knowledge, currently, the CU annotation are only been used for Basque and Brazilian Portuguese. Discourse structure annotation has been applied to develop different RST treebanks for different languages, including language pairs. For the works of segmentation annotation and discourse structure annotation, we detect the already existed works for Spanish and Chinese. Additionally, under the discourse structure category, we conclude the related works together with our works.

Then, we analyze the related discourse studies for Spanish and Chinese. Discourse analysis for Spanish are still few. In contrast, there are several discourse studies for Chinese by using different discourse theories (PDTB or RST). We have compared the discourse studies that focus on the Chinese with our study. For the previous works

who use RST for Chinese, each of them contains their own limitations, such as the limitation of corpus size, the unclear explanation of annotation process, the hidden evaluation results of annotation, etc. After, we extract the comparative studies for different language pairs under RST. For example, English-Chinese, Spanish-Chinese, Finnish -English, etc.

The works by using corpus-based approach for language learning are presented in the fifth part of this chapter. In this section, we conclude the corpus-based approach for different language pairs, English-Chinese, Spanish-Chinese and English-Spanish. Besides, we analyze the studies that use corpus-based approach for language learning in terms of RST. Our work is the first one that gives the complete discourse structure of each text in the corpus to help the translation and language learning between the Spanish and Chinese.

Chapter 4
Methodology

In Chapter 4, we will explain how we carry out the study. Firstly, we will focus on the development of the research corpus (Section 4.1). An analysis of the current existing Spanish-Chinese parallel corpora will be carried out. Based on the analysis, we will talk about what characters have been considered for the development corpus. Moreover, we will show the corpus information such as the statistical information of the corpus, the applications of the corpus. Secondly, each annotation step and the corresponded evaluation methods will be introduced in the following sections: discourse segmentation (Section 4.2), Central Unit (CU) (Section 4.3), and discourse structure (Section 4.4). Thirdly, we will talk about the considerations to develop the translation protocol that can help the Spanish-Chinese translation (Section 4.5). The following part (Section 4.6) will explain the elaboration of the exercises for the language learning between the two languages. Lastly, we will give an overview of this chapter.

4.1 Corpus Compilation

The research corpus is one of the fundamental research steps for this study. In modern linguistics, Leech defines the corpus as a body of naturally occurring language (1992: 116):

> *It should be added that computer corpora are rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) representative of some language or text type.*

Sinclair (1996) confirms Leech's definition of the corpus, as indicates, a corpus is a sorted collection of pieces of language to use be used as a sample of the language. In addition, Aston and Burnard (1998) recognize the 'linguistic criteria' as an external aspect for the selection of the texts to form a corpus. Thus, "a corpus is different from a random collection of texts or an archive whose components are unlikely to have been assembled with such goals in mind" (Aston and Burnard, 1998: 5). For computational linguistic study, Wu (2014) considers corpus as a large electronic library that provides a large amount of linguistic information.

There are many ways to define a corpus, McEnery, Xiao and Tono (2006: 5) argue the common characters for a corpus are:

> *[...]but there is an increasing consensus that a corpus is a collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety.*

As mentioned in Chapter 1, our studies is aims to create a Spanish-Chinese parallel corpus to detect the discourse information for the language pair. Currently, there are few parallel Spanish-Chinese corpora. Those corpora that already exist have their own limitations for Spanish-Chinese discourse analysis. Therefore, we decided to develop a new and more adequate Spanish-Chinese parallel corpus for our research. Our corpus contains the essential qualities as a corpus-based approach research: (i) machine readability (all the texts can be recognized by the computer), (ii) authenticity (the sources of the corpus are from real academic organizations and events), and (iii) representativeness (a corpus especially designed for discourse analysis).

In this section, we will analyze the disadvantages of the already existing Spanish-Chinese parallel corpora and set forth the detailed information of the new constructed corpus.

4.1.1 Analysis of the Previous Spanish-Chinese Parallel Corpora

To our knowledge, the already existing parallel corpora are: (a) The Holy Bible (Resnik, Olsen & Diab, 1999), (b) The United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009) and (c) Sina Weibo Parallel Corpus (Wang et al., 2013). As mentioned before, each corpus has its own limitations for Spanish-Chinese discourse analysis. This subsection explains why they are not adequate for translation and language learning purposes between Spanish and Chinese.

- The Holy Bible (Resnik, Olsen and Diab, 1999)

The Holy Bible contains 28,000 parallel sentences and around 800,000 tokens per language (Costa-jussà, Henríquez and Banchs, 2012). The Holy Bible is not appropriate for our purposes, due to the following constraints. First of all, the genre and domain in The Holy Bible is only one, so any study based on that, and only that, will be far from being general. Secondly, one author's translation determines the same discourse style in Bible and this fact could introduce bias in comparative discourse analysis. Lastly, the texts in the Bible are very old and cannot represent the modern language style.

- The United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009)

The texts of this corpus have been extracted from official documents of the UN. It is available for the six official languages of the UN (English, Chinese, Spanish, Russian, French, Arabic and German) and consists of around 300 million words for each language. Compiled in March of 2010, this corpus consists of 463,406 documents and 80,931,645 sentences in total.

The original language of the official documents in the UN corpus is English. The other texts are all translated from English, so the Spanish-Chinese parallel corpus is actually made up of two parts. One is the translation between English and Spanish, and the other is the translation between English and Chinese. These translated Spanish and Chinese documents make up the UN Spanish-Chinese parallel corpus. Due to the linguistic realizations (translation strategies), the rhetorical structure of the target language could be modified, and would affect the coherence relations between the clauses or sentences (Iruskieta, da Cunha & Taboada, 2015). In contrast, what we want to show in our study is the discourse structure of each language and the relations between discourse segments. Therefore, because it is not a direct translation corpus,

we consider the UN Spanish-Chinese subcorpus inadequate to carry out a Spanish-Chinese discourse comparative study.

- Sina Weibo Parallel Corpus (Wang et al., 2013)

The Sina Weibo Parallel Corpus is a multilingual corpus (Wang et al., 2013), which is readily available. In this corpus, 2000 selected Chinese texts have been translated into 9 languages (English, Spanish, French, Russia, Korean, German, Arabic, Portuguese and Czech). The texts of this corpus are independent sentences and are extracted from Weibo, which is similar to Twitter.

The main limitation of this corpus regarding discourse research is that the texts it contains are only tweets. Thus, they are very short texts, and, so far, they do not usually include complex discourse structures (such as, inter-sentential discourse relations). Moreover, their discourse structures are not always expressed formally, that is, by means of discourse markers. Regarding language learning, this corpus could be useful for Spanish-Chinese speech learning (because it shows a non-formal variety expression that can be useful for high skilled language learners); however, it is not adequate for analyzing the formal variety of language, either for translation or for second language learning purposes, since, in these contexts, discourse structure can be much more complex, and discourse segments usually contain discourse markers or signals.

4.1.2 Development of the Corpus

Since none of the already existing Spanish-Chinese corpora can be used either for a discourse comparative study or for the analysis of the translation realization in coherence relations, we have elaborated a new Spanish-Chinese parallel corpus. In this section we will explain the main stages of corpus compilation.

Firstly, in order to build the corpus and avoid the limitations of the existing corpora, we determined the main characteristics that the texts should include. These characteristics are the following: (a) Texts with an equal translation process. This means texts originally written in Spanish and translated into Chinese by natives or vice versa. (b) Texts with different sizes: texts between 90 and 1,500 words. This means that they are texts with a complex discourse structure. (c) Specialized texts. This also means that they can have a complex discourse structure. (d) Texts from different domains (to obtain a heterogeneous corpus). (e) Texts from different genres (to obtain a heterogeneous corpus). (f) Texts from different sources (to obtain a heterogeneous corpus). (d) Texts from different authors (to avoid bias).

Secondly, we searched for texts with these characteristics in different sources. To obtain high translation quality and various rhetorical structures (that is, coherence structure) in our corpus, we decided to use Spanish texts and their translations into Chinese, done by Chinese translators.

In order to confirm that all the texts fulfilled this translation process, it was necessary to contact the people in charge of the organizations that had published the source documents and their translations.

Due to the limitation of the available sources and the specific characteristics that we have determined, the amounts of texts that correspond with the required

translation process are few. In total, 50 Spanish texts and their parallel Chinese texts have been selected for our study.

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

Moreover, in order to guarantee the representativeness of our corpus, we have selected different types of texts from several domains. We chose the following four genres: (a) abstracts of research papers, (b) news, (c) advertisements and (d) announcements. Table 5 shows the genre statistical information of the corpus.

Genre	**Texts**	**Source**	**Source > Target**
Abstract of research paper	30	The International Conference about Terminology (1997)	Spanish > Chinese
News	30	The Shanghai Miguel Cervantes Library, The Chamber of Commerce and Investment of China in Spain, The Spanish Embassy in Beijing, The Confucius Institute Foundation in Barcelona	
Advertisement	26	The Shanghai Miguel Cervantes Library, The Spain-China Council Foundation, The Beijing Cervantes Institute, The Granada Confucius Institute	
Announcement	14	The Spain Embassy in Beijing, Confucius Institute Foundation in Barcelona, The Beijing Cervantes Institute	
Total	100		

Table 5. Genre information of the corpus

Furthermore, the texts have been divided into the following seven domains: (a) terminology, (b) culture, (c) language, (d) economy, (e) education, (f) art and (g) international affairs. Table 6 shows the domain statistical information of the corpus.

Domain	N° of texts per language	Original source
Terminology	30	The International Conference about Terminology (1997)
Culture	12	The Shanghai Miguel Cervantes Library, The Confucius Institute Foundation in Barcelona, The Beijing Cervantes Institute, The Granada Confucius Institute
Language	16	The Shanghai Miguel Cervantes Library, The Confucius Institute Foundation in Barcelona, The Beijing Cervantes Institute, The Granada Confucius Institute
Economy	14	The Chamber of Commerce and Investment of Chinese in Spain, The Spain-China Council Foundation
Education	8	The Confucius Institute Foundation in Barcelona, The Beijing Cervantes Institute
Art	10	The Spain Embassy in Beijing, The Beijing Cervantes Institute
International affairs	10	The Spain Embassy in Beijing, The Confucius Institute Foundation in Barcelona
Total	100	

Table 6. Domain information of the corpus

Thirdly, we have enriched the corpus with POS information for the Spanish subcorpus by using Freeling (Carreras et al., 2004) and the Chinese subcorpus automatically by using the Stanford parser (Levy & Manning, 2003).

Finally, we make our corpus available to the public (see Figure 10). The corpus can be downloaded through: `http://ixa2.si.ehu.es/rst/zh/index.php`. Our corpus is the first discourse based Spanish-Chinese parallel corpus whose resources are available to the public.

Figure 10. The website of the RST Spanish-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. In addition, discourse segmentation can be useful for different NLP tasks such as the evaluation of automatic segmentation systems, and the development of discourse parsers and automatic summarizers. In this section, we will explain an overview of the related segmentation works based on discourse analysis. Then, we will explain the segmentation criteria of this work; each segmentation criterion will be presented with a Spanish-Chinese parallel example.

4.2.1 Elaboration of the Discourse Segmentation Criteria

In this work, we use the RSTTool (O'Donnell, 2000) to carry out the segmentation work. By using the RSTTool, an entire text can be divided into various independent EDUs. Figure 11 and Figure 12 include an example of a parallel-segmented Spanish-Chinese text from the corpus.

A Spanish-Chinese bilingual expert and two Spanish experts are in charge of the segmentation for the Spanish subcorpus. Meanwhile, the bilingual expert and a Chinese expert carry out the segmentation task for the Chinese subcorpus. The bilingual expert annotates all 100 texts; each of the Spanish experts annotates 25 Spanish texts. The Chinese expert annotates all the 50 Chinese texts.

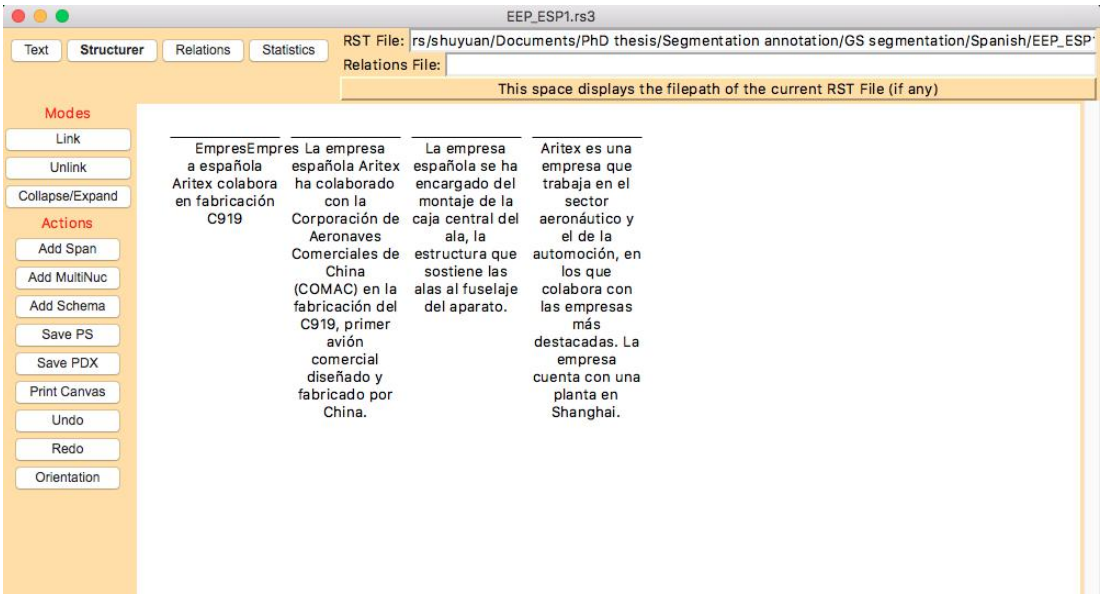

Figure 11. A parallel-segmented Spanish-Chinese text using RSTTool (Spanish text)

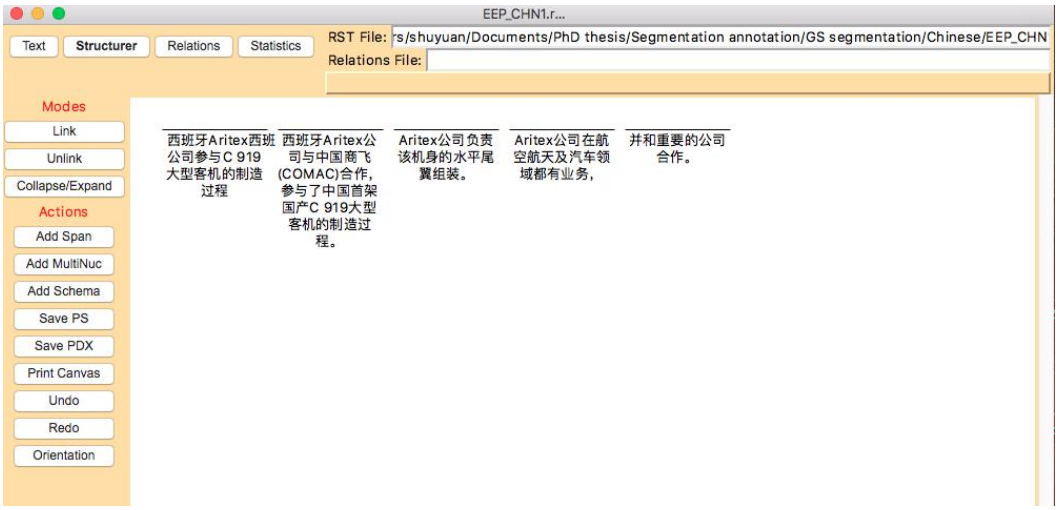

Figure 12. A parallel-segmented Spanish-Chinese text using RSTTool (Chinese text)

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. From the function and form perspective, we adopt the segmentation criteria from Iruskieta, da Cunha and Taboada (2015).

The following segmentation criteria are used in out work[38]:

[38] For all the annotation criteria, we give an example of each annotation criterion in the Appendix part. Detailed information of each annotation criterion can be consulted there.

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

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

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

Other EDUs should have a main verb or an adjunct verb phrase[39]. This is a basic segmentation criterion and segmentation criteria bellow should follow this rule. Titles are considered as the exceptions, whether they contain a verb or not, titles are always EDUs.

Discourse Marker (DM), 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.

Semicolon plus adjunct verb phrase.

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[40].

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

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

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

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

The segmentation information is available in the website. All the segmented texts can be consulted from there, as Figure 13 and Figure 14 present.

---

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

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

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

ICP_ESP1-GS.rs3 (13)

EDU	Segment	Tagger	Central Unit
1	Presentación institucional	GS	
2	El Instituto Cervantes es la institución creada por España en 1991	GS	
3	para promover, enseñar español y difundir la cultura de España y de los países hispanohablantes.	GS	
4	La sede central de la institución se encuentra en Madrid y en Alcalá de Henares (Madrid), ciudad de nacimiento del escritor Miguel de Cervantes.	GS	
5	El Instituto Cervantes está presente en cinco continentes con 77 centros, entre los cuales destaca el Instituto Cervantes de Pekín, el primero en China.	GS	
6	Nuestra institución también se encarga de:	GS	
7	• Organizar los exámenes para el Diploma de Español como Lengua Extranjera (DELE), así como de expedir certificados y diplomas oficiales para los participantes en nuestros cursos	GS	
8	• Organizar cursos de español	GS	
9	• Organizar cursos de formación para profesores de español	GS	
10	• Apoyar a hispanistas en sus actividades	GS	
11	• Estimular actividades culturales en colaboración con otras organizaciones	GS	
12	El trabajo del Instituto Cervantes está dirigido por representantes del mundo académico, cultural y literario del ámbito español e hispanoamericano.	GS	
13	En Pekín colabora con museos, galerías, teatros, editoriales y otras instituciones culturales chinas, así como españolas y latinoamericanas.	GS	

Figure 13. A segmented text in the website (Spanish text)

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 14. A segmented text in the website (Chinese text)

Figure 13 and Figure 14 contain the example of the Spanish-Chinese texts. In Figure 13, we can see that, the Spanish text has been segmented into 13 EDUs. As its parallel text, the Chinese in Figure 14 also contains 13 EDUs.

4.2.2 Evaluation of Inter-annotator Agreement

For the segmentation annotation (both Spanish and Chinese parts), we use the Kappa score to measure the agreement between the annotators in RST discourse segmentation[41]. Previous work has proved that Kappa can be used for segmentation evaluation (Iruskieta, Diaz de Ilarraza and Lersundi 2015). Kappa calculates the agreement between annotators as:

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

[41] We assign the Spanish-Chinese bilingual expert as the Annotator 1 (A1), the two Spanish speaking experts as the Annotator 2 (A2) and the Chinese native speaking linguist as the Annotator 3 (A3).

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.

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

4.3 CU Annotation

Under RST, for each segmented text, among the EDUs, there is an EDU called Central Unit (CU) that contains the key information of the text (Cao, da Cunha, and Iruskieta, 2016). CU can be applied to different NLP studies, for example, automatic summarization, development of intelligent systems and sentiment analysis (Iruskieta, Labaka and Desiderato, 2016). 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 (Cao, da Cunha and Iruskieta, 2016).

Figure 15 presents the CU of the annotate Spanish text in the corpus and Figure 16 shows the CU of its parallel annotate Chinese text.

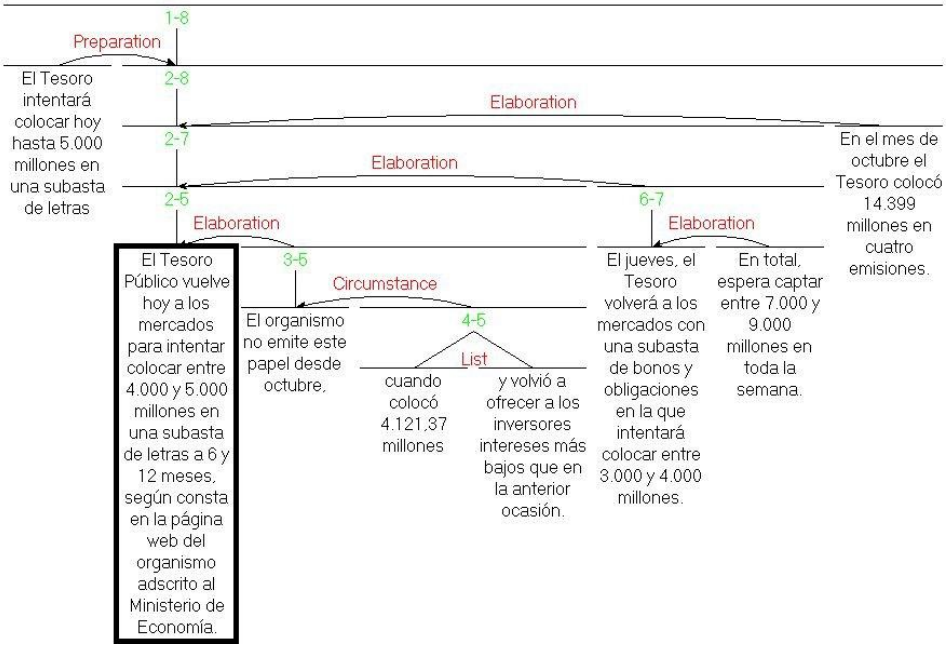

Figure 15. CU of the annotate Spanish text (CCICE3_ESP)

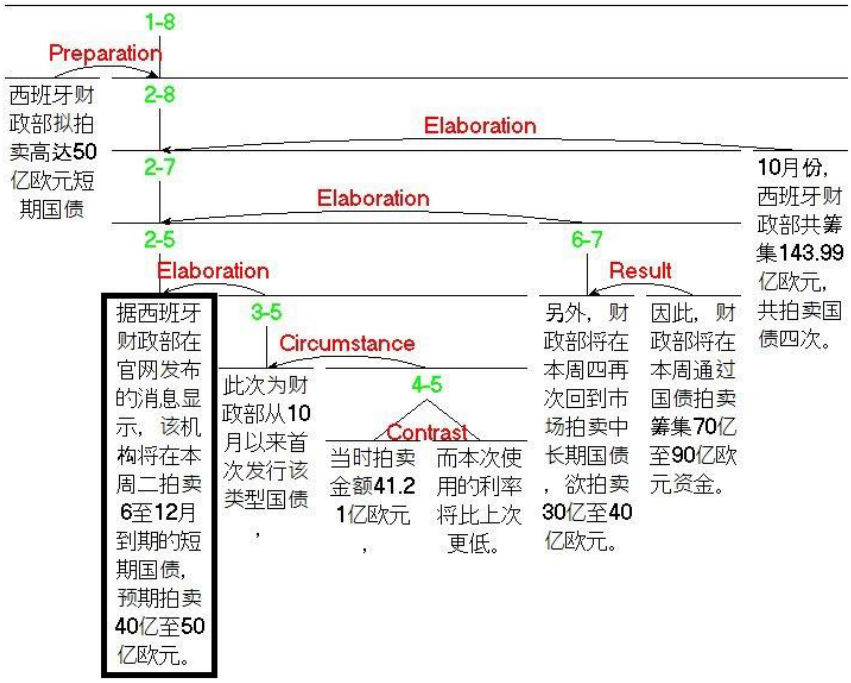

Figure 16. CU of the annotate Chinese text (CCICE3_CHN)

Figure 15 shows that, for the annotate Spanish text, all the arrows are point to EDU2, which means the content of "*El Tesoro Público vuelve hoy a los mercados para intentar colocar entre 4.000 y 5.000 millones en una subasta de letras a 6 y 12 meses, según consta en la página web del organismo adscrito al Ministerio de Economía*.[42]" is the main information of the Spanish text. In Figure 16, for the parallel 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。)[43]".

4.3.1 Description of the CU Annotation Criteria

According to van Dijk (1980), language users are able to summarize discourses, expressing the main topics of the summarized discourse. In this study, for each segmented text, the annotators decide which EDUs represent the main idea of the text.

A Spanish-Chinese bilingual linguist and a Spanish linguist annotate the CUs for all the Spanish texts. The bilingual linguist and a Chinese linguist selected the CUs for all the Chinese texts. All the words (noun, verb, proper noun, preparation, pronoun,

[42] English literal translation: The Public Treasury returns today to the markets to try to place between 4,000 and 5,000 million in an auction of letters to 6 and 12 months, according to the web page of the organized ascribed to the Ministry of Economy.

[43] 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.

and conjunction) that can represent the CUs of a text have been annotated for both Spanish and Chinese texts.

4.3.2 Evaluation of Inter-annotator Agreement

The evaluation of CUs using the Kappa score under RTS has been presented in two works; one is Burstein et al. (2001) and another one is Iruskieta, Díaz de Ilarraza and Lersundi (2014). Both works prove that the Kappa score can measure the agreement of CU annotations under RST. In this work, we also use Kappa to evaluate the CU annotation results.

4.4 Discourse Structure Annotation

Discourse structure annotation is one of the most difficult challenges for annotation works (Hovy and Lavid, 2010). As Das, Taboada and Stede (2017: 11) indicate:

> *In rhetorical analysis, as in many other pragmatic annotation tasks, a certain amount of disagreement is to be expected, and it is important to distinguish true mistakes from legitimate disagreement due to different possible interpretations of the structure and intention of a text.*

In this study, we annotate the discourse structure of all the texts in the corpus.

4.4.1 Description of the Discourse Structure Annotation Criteria

Firstly, we select the discourse relations for this study. The discourse relations that we use are in the following table (Table 7). In total, 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[44].

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 7. Selected discourse relations for discourse annotation

Secondly, we annotate the discourse structure with the selected discourse relations. A Spanish-Chinese bilingual linguist and a Spanish linguist annotate the discourse structures for all the Spanish texts. The bilingual linguist and a Chinese linguist

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

annotate the discourse structures for all the Chinese texts. For the text annotation, we follow the annotation guidelines proposed by Pardo (2005). First, we annotate the relations within the segmented sentences (intra-sentence style); then, we identify the relations between the sentences within a paragraph (inter-sentence style). Lastly, we find the relations between paragraphs.

We use the RSTTool to finish the discourse annotation task. Figure 17 shows an annotate Spanish text from the corpus with the RSTTool; meanwhile, Figure 18 presents its parallel annotate Chinese text.

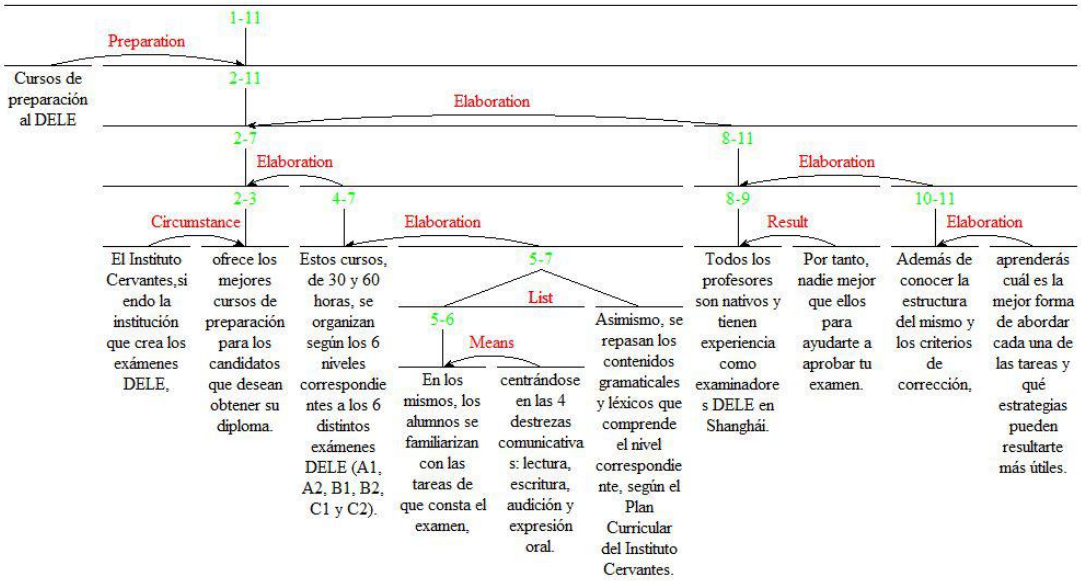

Figure 17. The annotate Spanish text by using the RSTTool

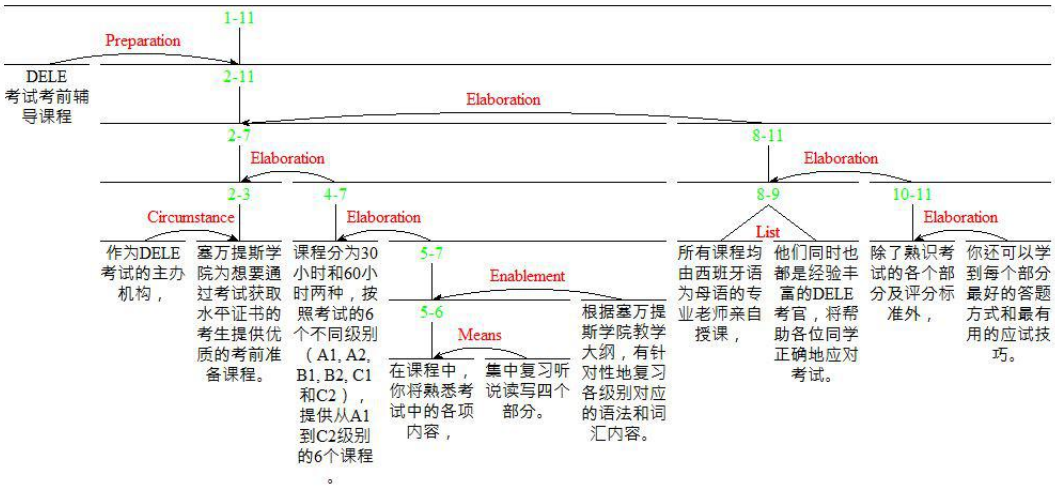

Figure 18. The annotate Chinese text by using the RSTTool

From Figure 17, we can see that the Spanish texts has been annotated with intra-sentence and inter-sentence styles. For example, we can see the intra-sentence annotation style with EDU2 and EDU3. EDU2 and EDU3 are two parts of a complete

sentence, and the relation between the EDUs are CIRCUMSTANCE. An example of the inter-sentence annotation style can be seen with EDU8 and EDU9 where EDU8 and EDU9 are two sentences that hold a LIST relation between them. As presented in Figure 18, the parallel Chinese text is also being annotate with intra-sentence and inter-sentence styles. Same as the Spanish annotate text, EDU2 and EDU3 in the Chinese text are also two parts of a sentence and the discourse relation between the two EDUs is CIRCUMSTANCE. EDU8 and EDU9 are two sentences that contain a LIST relation within them. In the Appendix part, for each selected relation, we will give an example from the corpus and its English literal translation[45].

All the annotated texts can be consulted in the website. We give the annotation results as 3 forms: rs3, text and image. Figure 19 shows the how to consult the annotated texts from the corpus[46].

Figure 19. Corpus consultation with different ways

From Figure 19, 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 20).

[45] 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.

[46] Due to the space limitation, Figure 21 shows parts of the website.

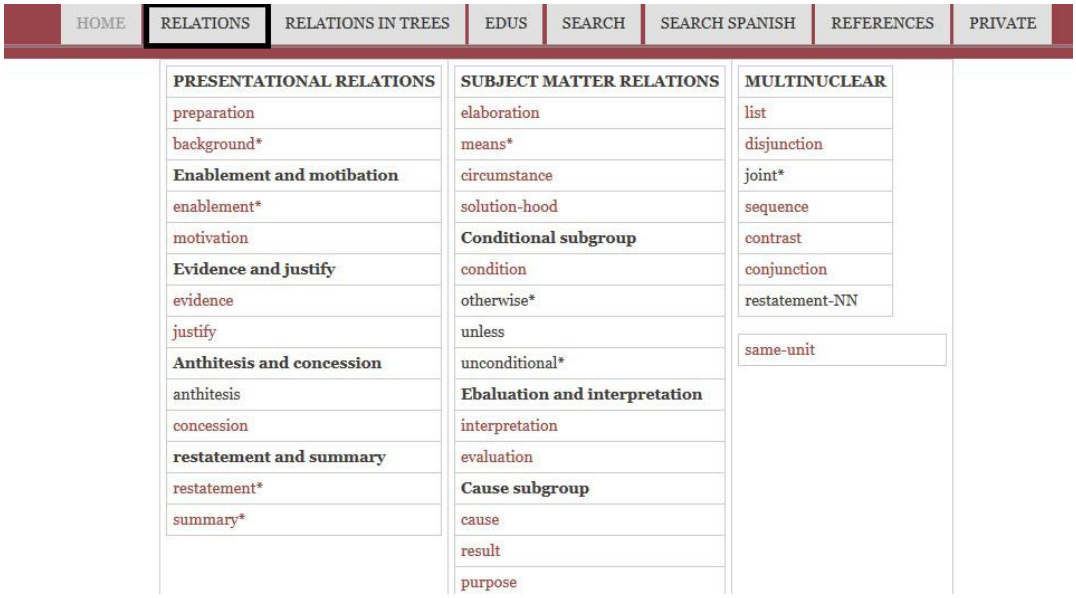

Figure 20. Consultation of each selected relations

In our website, as Figure 20 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.

4.4.2 Evaluation of Inter-annotator Agreement

Currently, there are two evaluation methods to evaluate the RST discourse structure annotation, one is a quantitative method created by Marcu (2000); and another method is a qualitative method proposed by Iruskieta, da Cunha and Taboada (2015).

- Quantitative method

Although the quantitative method is the first method for RST annotation evaluation, several works (van der Vliet, 2010; da Cunha and Iruskieta, 2010; Iruskieta, da Cunha and Taboada, 2015) specify its limitations. The main limitations are:

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

ii) Deficiencies in the descriptions. The description of comparison and weight used for agreement in certain discourse relations still need to be improved. When the annotators assign a relation that has an attachment point at different levels in a tree structure, the relations that have a change of constituents cannot be compared by the quantitative method (da Cunha and Iruskieta, 2010; Iruskieta, da Cunha, and Taboada, 2015).

Following da Cunha and Iruskieta (2010), Iruskieta, Díaz de Ilarraza and Lersundi (2013), and Iruskieta, da Cunha and Taboada (2015), for measuring annotation agreement, six factors must be considered: EDU and Span (segmentation), N-S function (Nuclearity), attachment point, constituent and discourse relation (Relation). Especially when comparing a parallel text, the quantitative cannot measure the annotation agreement adequately.

- Qualitative method

The qualitative evaluation method is a method created by Iruskieta, da Cunha and Taboada (2015). This evaluation method quantifies linguistic data for rhetorical structure and also shows linguistic features that affect rhetorical structure. This is the first study that provides a rigorous qualitative methodology for comparing 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 the types of agreement and sources of disagreements. The types of agreements under this method are (Iruskieta, da Cunha and Taboada, 2015: 276):

i) Agreement in relation, constituent, and attachment point (RCA).
ii) Agreement in relation and constituent (RC).
iii) Agreement in relation and attachment point (RA).
iv) Agreement only in relation (R).

The sources of disagreements are the following:

i) Disagreements of annotators (type A). The text doesnot contain the significant linguistic differences; instead the annotators define the distinct relations. Seven sources of such disagreements are included in their discussion (Iruskieta, da Cunha and Taboada, 2015: 277):

✓ *Different choice of in nuclearity entailed a N/N-N/S mix-up (N/N-N/S).*
✓ *Different choice in nuclearity entailed discrepancy in N/S relations (N/S).*
✓ *Relation has the same constituent and attachment point, but not the same relation label (≠R).*
✓ *Relations chosen are similar in nature (Similar R).*
✓ *Relations with mismatched RST trees (Mismatch R).*
✓ *Relation is more specific than the other (Specificity).*
✓ *Different choice in attachment entailed a different relation (Attachment).*

ii) Disagreement of language (type of L). Because of the difference in the linguistic form, the annotators assign distinct relations. Three sources of disagreements are found under this case (Iruskieta, da Cunha and Taboada, 2015: 278):

✓ *A relation is signed with a different discourse marker (Marker Change or MC).*
✓ *A different organization of constituent phrases is used, mostly from non-finite verb phrase to finite verb structure (Clause Structure Change or CSC).*

✓ *A change in unit level (phrase-clause-sentence) is done (Unit Shift or US).*

Comparing to the quantitative evaluation method, the qualitative evaluation method describes the annotation agreement in a more detailed way. Additionally, the qualitative method can explain the causes of discourse differences in texts written in different languages; therefore, we adopt this method for the annotation evaluation.

4.5 Elaboration of a Spanish-Chinese Discourse Recommendation Protocol for Translators

Once the annotation part is finished, we start to elaborate the translation protocol with recommendations that can help the translation between Spanish and Chinese. The recommendations in this protocol can be useful for Spanish and Chinese translation.

4.5.1 RS-tree Comparison

By using the qualitative method, we analyze the discourse elements of nuclearity (N), relation (R), constituent (C) and attachment point (A). We use the F-measure to measure the agreement of the annotated discourse elements. Then, we conclude the similarities and differences by counting the number of appearances.

4.5.2 Translation Strategies

The recommendations of the protocol are elaborated based on the comparison of the annotation results. The translation strategies detected by using the qualitative method are the principal aspects included in the translation protocol. The translation strategies are mentioned in Section 4.4.2[47]:

(i) Marker change. Different DMs are assigned for the same relation.

(Ex.1).
Spanish: [Es más, desde cualquier lugar los términos son recopilados, comentados y ponderados;]$_{9\text{-}N}$ [de ahí, por ejemplo, los apartados que encontrados en muchos Webs en que se difunden glosarios de términos sobre Internet o en que se exponen propuestas denominativas que los usuarios pueden incluso votar.]$_{10S\text{-}EVIDENCE}$
English: [Furthermore, terms can be compiled, discussed and assessed anywhere:]$_{9N}$ [many Web sites can be found which give glossaries of Internet terms or propose names and even invite users to vote on them.]$_{10S\text{-}ELABORATION}$

In Example 1, we can see that the Spanish passage contains a DM "*de ahí*" ('hence') does not have its English translation in the English passage. This why there is a EVIDENCE relation in the Spanish passage meanwhile the relation in English is ELABORATION.

[47] In this section, for each translation strategy, we will give a Spanish-English parallel example cited from the work of Irusksieta, da Cunha and Taboda (2015). The cases that contain these mentioned translation strategies in our corpus will be presented in the protocol part.

(ii) Clause Structure Change. A non-finite verb phrase is changed to finite verb structure.

(Ex.2).
Spanish: [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología,}$_{6N}$ {que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos (...)]$_{7\text{-}11S\text{-}ELABORATION}$
English: [All these factors lead to an increase in the number of specialist terms which enrich terminology]$_{6N\text{-}CONTRAST}$ [but also call into question some of its basic concepts (...)]$_{7N\text{-}CONTRAST}$

From the above example we can see that, the discourse relations in Spanish and English are different. This is because there is a verb in the coordination clause ('but') in English, but not in Spanish. This is why there is a ELABORATION relation in Spanish meanwhile the relation in English is CONTRAST.

(iii) Unit shift. There is a change in the unit level.

(Ex.3).
Spanish: [En esta communicación, apartir de la experiencia en trabajos de normalización de terminología catalana, se plantearía la necesidad social de la normalización terminológica,]$_{N12\text{-}LIST}$ [se comentaría algunas de las dificultades con que se enfrenta y se apuntará ideas para su enfoque dentro de la sociedad actual.]$_{N13\text{-}14\text{-}LIST}$
English: [This paper looks, on the basis of experience in the standardisation of terminology in Catalan, at the social need for standardisation of terminology,]$_{N12}$ [Some of the difficulties faced will be discussed, and ideas will be given for approaching this field in present day society.]$_{S13\text{-}14\text{-}ELABORATION}$

In the original text in Spanish, we can see that there is a comma between two EDUs, whereas there is a period between two EDUs in English. The different punctuations cause the different relations.

4.5.3 Elaboration of Recommendations

Regarding the order of the recommendations in the protocol, recommendations are grouped in three parts, based on the translation strategies mentioned in Section 4.4.2. Besides, in each of these three groups, the recommendations related to the same discourse relation (for instance, LIST, CONJUNCTION, or RESULT) are grouped together. Finally, the list of recommendations related to the same discourse relation starts with the recommendation that contains discourse relation change between Spanish and Chinese.

All the recommendations include the following four discourse aspects: (i) DMs, (ii) Discourse relations, (iii) Relation types, and (iv) EDUs order. In addition, some recommendations included in the groups Unit shift and Others also contain other aspect called Punctuation marks.

4.6 Applications of Results for Spanish-Chinese Language Learning Tasks

Another objective of this study is to create the tasks that can serve for the Spanish-Chinese language learning. In this section, we will describe how we design the tasks (a Spanish-Chinese generation system) based on the previous annotation results.

4.6.1 Annotation of DMs

The main idea of the creation of the task is to create a question-answering exercise generation with multichoice. The discourse element that we select is the discourse markers. Based on the discourse structure annotation, we annotate the DMs manually for both Spanish subcorpora and Chinese subcorpora.

4.6.2 General Information for Exercise Elaboration

For the Spanish language exercise, we use encoding to generate the texts automatically by removing the annotated DMs. The system erases the annotated Spanish DMs, and for each erased DM, the system gives the multi choices. When the student finishes the exercise, the system can grade the exercises.

Same as the Spanish part, we make a small program to take our all the discourse markers. However, the exercise design is different from Spanish language exercise. The system erases all the annotated DMs, and gives the unordered DMs to let the student to choose the corresponded one for each blank.

4.6.3 Exercise Evaluation for Spanish and Chinese

For the automatic generation program we make, we use Kappa to evaluate the correctness of our programming. Kappa gives the agreement of annotation as follows:

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

where P(A) represents the actual observed agreement, and P(E) represents chance agreement.

4.7 Chapter Overview

In this chapter, firstly, we explain the methodology of the study. We elaborate the four steps to carry out the study: (i) corpus construction, (ii) segmentation annotation, (iii) CU annotation, (iv) discourse structure annotation, (v) development of the translation protocol, and (vi) creation of language tasks for Spanish-Chinese language learning.

At the current stage, there is no Spanish-Chinese parallel corpus that is adequate for discourse analysis; our corpus is the first one that especially designed for discourse analysis. Moreover, this corpus is available to the scientific community. Based on the framework, we establish some criteria for the discourse segmentation, which is the crucial step for the rest annotation step, especially for the discourse structure quality. The central unit (CU) is the third step of the methodology. We have extracted the key information of all the texts in the corpus. Lastly, with the selected

relations, we annotate all the corpus with intra-sentence style and inter-sentence style.

In considering of the corpus, we explain the limitations of the already existed Spanish-Chinese parallel corpora and develop a new parallel corpus. We present the detailed information of the sources, genre and topics of our corpus. For each annotation step, we list the annotation criteria and their corresponded evaluate methods for the inter-annotation agreement.

Regarding the two objectives of this study, creation of the translation protocol and language tasks for language learning, we outline the process of how to get the goal of the each objective. The translation protocol contains recommendations are grouped by the translation strategies, which are produced by the qualitative comparison of the annotation results. The recommendations are related with DMs, discourse relations, type of relations and order of EDUs. Moreover, the recommendations of some cases also contains the information of the change of punctuation. Each recommendation has been given a real case for the corpus. Our protocol is the first one that supports the Spanish-Chinese translation with RST.

To achieve the goal for the Spanish-Chinese language learning, we annotate all the DMs for both Spanish subcorpus and Chinese subcorpus. By using encoding, we erase all the annotated DMs and offer different choices for students. At the moment, the system can also grade for the Spanish language exercise. To our knowledge, our system is the first one to help the Spanish-Chinese language learning with RST.

Chapter 5

Evaluation and Analysis

In this chapter, we explain the annotation evaluation for each annotation part. In addition, we provide an analysis for each evaluation results. In section 5.1, we evaluate the segmentation annotation by calculating the accuracy using the Kappa measure. We also explore the causes of our annotation disagreement. In section 5.2, we evaluate the CU annotation agreement, again using the Kappa measure. We also provide a qualitative analysis of the annotation disagreement in this section. In section 5.3, we evaluate the reliability of the discourse relation annotation by using F-measurement following a newly created qualitative analysis by Iruskieta, da Cunha and Taboada (2015). We evaluate the agreement of the following aspects: Nuclearity (N), Relation (R), Composition (C), and Attachment (A). Lastly, we summarize the chapter information (Section 5.4).

5.1 Discourse Segmentation

In this section, we explain the evaluation results of the segmentation annotation agreement between the annotators for both the Spanish subcorpus and the Chinese subcorpus. An analysis of the segmentation disagreement between the annotators will also be provided.

5.1.1 Segmentation

As mentioned previously in the methodology chapter, for the segmentation annotation of the Spanish subcorpus, we invited the two Spanish native speaking experts (25 texts for each) and a Spanish-Chinese bilingual expert (all 50 texts) to segment the Spanish texts. For the Chinese subcorpus, the Spanish-Chinese bilingual expert and another Chinese native speaking linguist (50 texts for each) are in charge of the segmentation of the Chinese texts[48]. For both the Spanish and Chinese parts, we use the Kappa score to measure the agreement between two annotators in RST discourse segmentation. Table 8 includes the statistics used to measure the agreement between the annotators for the Spanish part while Table 9 contains the evaluation results for Chinese part.

Annotator		**A2**		**Total**
		Yes	No	
A1	Yes	715	31	746
	No	142	3833	3975
Total		857	3864	4721

Table 8. Segmentation cross tabulation of the Spanish subcorpus

[48] We assign the Spanish-Chinese bilingual expert as the Annotator 1 (A1), the two Spanish speaking experts as the Annotator 2 (A2) and the Chinese native speaking linguist as the Annotator 3 (A3).

Annotator		A3		Total
		Yes	No	
A1	Yes	765	101	866
	No	204	1888	2092
Total		969	1989	2958

Table 9. Segmentation cross tabulation of the Chinese subcorpus

Table 10 and Table 11 include the Kappa agreement results regarding each part of both Spanish and Chinese parts.

Corpus Source	Kappa Agreement
ICT	0.895
SMCL	0.945
CCICS	0.855
SEB	0.786
SCCF	0.828
CIFB	0.716
BCI	0.863
GCI	0.873
Total	0.87

Table 10. K results regarding each part of the corpus (Spanish subcorpus)

Corpus Source	Kappa Agreement
ICT	0.815
SMCL	0.719
CCICS	0.744
SEB	0.711
SCCF	0.711
CIFB	0.616
BCI	0.759
GCI	0.705
Total	0.76

Table 11. K results regarding each part of the corpus (Chinese subcorpus)

Table 10 shows the Kappa agreement results of each part in the Spanish subcorpus. The highest agreement result between the annotators is 0.945; the lowest result is 0.716. The final K result of the Spanish subcorpus is 0.87, which means the preliminary segmentation between the annotators for the Spanish texts is almost perfect. Table 11 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 is reliable for Chinese.

5.1.2 Discussion of Segmentation Results

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

- Title

(Ex.1) Text name: BMCS5

A1: [• **¡Hola, amigos!** Curso en línea para niños y jóvenes, organizado en tres niveles de 9 unidades, que cubre los niveles A1 y A2 del marco común europeo de referencia para las lenguas.] (×)[49]

English: [• Hi, friends! Course on line for children and young people, organized in three levels of 9 units, that cover the levels A1 and A2 of framework common European of reference for the languages.]

A2: [• **¡Hola, amigos!**] [Curso en línea para niños y jóvenes, organizado en tres niveles de 9 unidades, que cubre los niveles A1 y A2 del marco común europeo de referencia para las lenguas.] (√)

English: [• Hi, friends!] [Course on line for children and young people, organized in three levels of 9 units, that cover the levels A1 and A2 of framework common European of reference for the languages.]

Analysis: In the preliminary segmentation criteria, we have indicated that every title is an independent EDU regardless of the verbs. The annotation of A1 doesnot take the subtitle "•¡Hola, amigos!" as an independent EDU, therefore, the annotation of A1 is not correct.

- Wrong EDU without verbs

(Ex.2) Text name: BMCS2

A1: [Los profesores cuentan siempre con el punto de vista de sus alumnos en la toma de decisiones de la clase, fomentando la autonomía del estudiante mediante el uso de las estrategias de aprendizaje más adecuadas para cada uno.] (√)

English: [Teachers rely on always with the point of view of their students in the making of decisions of the class, promoting the autonomy of the student by the use of the strategies of learning more adequate for each one.]

A2: [Los profesores cuentan siempre con el punto de vista de sus alumnos en la toma de decisiones de la clase,] [fomentando la autonomía del estudiante mediante el uso de las estrategias de aprendizaje más adecuadas para cada uno.] (×)

English: [Teachers rely on always with the point of view of their students in the making of decisions of the class,] [promoting the autonomy of the student by the use of the strategies of learning more adequate for each one.]

Analysis: Among the preliminary segmentation criteria, one criterion is the verb in each EDU. However, the annotation of A2 doesnot follow this rule, there is no verb in the second EDU.

---

[49] In this work, we use "√" to represent the correct segmentation and "×" to represent the incorrect segmentation. A1 represents the first annotator and A2 means the second annotator.

• Coordination and ellipsis

(Ex.3) Text name: CCICE4

A1: [**Los números rojos** no dejaban margen y hubo de acelerarse un programa restrictivo que llega hoy a uno de sus puntos culminantes.] (×)

English: [The number red no left margin and had of accelerated a program restrictive that reach today to one of their points culminating.]

A2: [**Los números rojos** no dejaban margen] [y hubo de acelerarse un programa restrictivo que llega hoy a uno de sus puntos culminantes.] (√)

English: [The number red no left margin] [and had of accelerated a program restrictive that reach today to one of their points culminating.]

Analysis: In this case, after the DM "*y*", the subject "los números rojos" ('the red numbers') has been erased. The annotator A1 has overlooked the elliptical subject and left the whole sentence as an independent EDU. The annotation of A2 is correct, the two EDUs are the coordinated parts with main verbs.

• Overlook of DM

(Ex.4) Text name: TERM30

A1: [La gestión terminológica, esto es, la recopilación, el análisis, la convalidación y la distribución de términos, es una tarea crucial **para** convertir la información en conocimiento comprensible y aplicable.] (×)

English: [The management of terminological, that is, the collection, the analysis, the validation and the distribution of terms, is a task crucial for convert the information in knowledge comprehensible and applicable.]

A2: [La gestión terminológica, esto es, la recopilación, el análisis, la convalidación y la distribución de términos, es una tarea crucial] [**para** convertir la información en conocimiento comprensible y aplicable.] (√)

English: [The management of terminological, that is, the collection, the analysis, the validation and the distribution of terms, is a task crucial] [for convert the information in knowledge comprehensible and applicable.]

Analysis: There is a DM "*para*" ('for') in the Spanish passage that represents the PURPOSE relation. The annotation of A1 overlooks the DM "*para*" ('for'). The annotation of A2 follows the segmentation criterion of DM and each EDU contains a main verb, which are "*ser*" ('is') in the first EDU and "*convertir*" ('convert') in the second EDU.

• Wrong understanding of DM

(Ex.5) Text name: ICP6

A1: [Para ello, se ha considerado importante dar continuidad a una programación cultural que haga coexistir equilibradamente tradición y modernidad,] [**así como** una oferta variada de cursos y actividades cuyo objetivo es atraer y fidelizar a un público heterogéneo.] (×)

English: [To this end, it has been considered important explain continuity to a program cultural that brings coexist balanced tradition and modernity,] [as well as a offer varied of courses and activities their object is attract and retain to a public heterogeneous.]

A2: [Para ello, se ha considerado importante dar continuidad a una programación cultural que haga coexistir equilibradamente tradición y modernidad, **así como** una oferta variada de cursos y actividades cuyo objetivo es atraer y fidelizar a un público heterogéneo.] (√)

English: [To this end, it has been considered important explain continuity to a program cultural that brings coexist balanced tradition and modernity, así como una oferta variada de cursos y actividades cuyo objetivo es atraer y fidelizar a un público heterogéneo.]

Analysis: The annotator A1 confuses the meaning of "*así como*" ('as well as') with the meaning "*por eso*" ('therefore'), and considers the "*así como*" as a DM, and the discourse relation between the segmented parts is RESULT. The segmentation of A1 is not correct.

• Colon

(Ex.6) Text name: EEP7

A1: [Las películas seleccionadas son**:**

- La isla mínima (Alberto Rodriguez, 2014)[50]
- Eva (Kike Maillo, 2011)
- Los últimos dias (Alex & David Pastor, 2013)
- Zipi y Zape y el club de la cánica (Oskar Santos, 2013)
- Loreak (José Maria Goenaga, Jon Garaño, 2014)
- Presentimientos (Santiago Tabernero, 2014)
- La vida inesperada (Jorge Torregrossa, 2014)] (×)

English: [The selected films are:

- La isla mínima (Alberto Rodriguez, 2014)
- Eva (Kike Maillo, 2011)
- Los últimos dias (Alex & David Pastor, 2013)
- Zipi y Zape y el club de la cánica (Oskar Santos, 2013)
- Loreak (José Maria Goenaga, Jon Garaño, 2014)
- Presentimientos (Santiago Tabernero, 2014)
- La vida inesperada (Jorge Torregrossa, 2014)]

A2: [Las películas seleccionadas son**:**]

[- La isla mínima (Alberto Rodriguez, 2014)]
[- Eva (Kike Maillo, 2011)]
[- Los últimos dias (Alex & David Pastor, 2013)]
[- Zipi y Zape y el club de la cánica (Oskar Santos, 2013)]
[- Loreak (José Maria Goenaga, Jon Garaño, 2014)]
[- Presentimientos (Santiago Tabernero, 2014)]
[- La vida inesperada (Jorge Torregrossa, 2014)] (√)

English: [The selected films are:]

[- La isla mínima (Alberto Rodriguez, 2014)]
[- Eva (Kike Maillo, 2011)]
[- Los últimos dias (Alex & David Pastor, 2013)]

[50] In this case, besides of the segmentation criterion of colon, we consider the films' names as the independent EDUs as special cases.

[- Zipi y Zape y el club de la cánica (Oskar Santos, 2013)]
[- Loreak (José Maria Goenaga, Jon Garaño, 2014)]
[- Presentimientos (Santiago Tabernero, 2014)]
[- La vida inesperada (Jorge Torregrossa, 2014)]

Analysis: The punctuation of colon is one of the segmentation rule for our study, the parts after colon should segmented after the colon. The annotator A2 follows this segmentation rule meanwhile the annotator A1 doesnot.

The analyses of the segmentation annotation of the Chinese subcorpus are the following:

- Title

(Ex.7) Text name: TERM31

A1: [2.] [术语构建] (×)

English: [2.] [Terminology construction]

A3: [2. 术语构建] (√)

English: [2. Terminology construction]

Analysis: The annotator 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

(Ex.8) Text name: TERM19

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。] (√)

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

A3: [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。] (×)

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

Analysis: The annotator 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

(Ex.9) Text name: TERM38

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。] (√)

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

A3: [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。] (×)

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

Analysis: The annotator 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 explain name / dominate') in the first EDU and "lai" (来) ('come') in the second EDU.

- Temporal adverb clause + comma + verb clause

(Ex.10) Text name: TERM25

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"。] (√)

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

A3: [**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"。] (×)

English: [When all the previous mentioned can be de-scribed in the same passage, we can establish the "legal discourse system" for Basque.]

Analysis: The annotator 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

(Ex.11) Text name: EEP2

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。] (×)

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

A3: [**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。] (√)

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

Analysis: The annotator 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 carry out a debate between annotators and we have improved our segmentation criteria and taking our preliminary 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 both Spanish and Chinese. This gold standard will be the basis for the discourse annotation of the corpus.

Table 12 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 12. Final discourse segmentation criteria

5.2 CU Annotation

As the second step of the annotation part, we have annotated the Central Unit (CU) of all the texts in the research corpus. In the section 5.2, we will evaluate the agreement of the CU annotation between the annotators for the whole corpus. A qualitative analysis of the annotation disagreement between the annotators will also be introduced in this section.

5.2.1 CU Annotation Results

As van Dijk (1980) indicates, language users are able to identify the global meaning of the texts. Depending on the natural comprehension of the texts, among the EDUs of each segmented text, we extracted the key information manually. A Spanish-Chinese bilingual expert decides the CUs of all the texts in the corpus. Concurrently, a Spanish speaking linguist annotates the CUs for the Spanish subcorpus (50 texts) and a Chinese native linguist decides the CUs for the Chinese subcorpus (50 texts)[51]. Table 13 explains the statistical information of the segmented texts of the corpus.

Corpus	Nºof texts	Words	EDUs
Spanish	50	14,636	840
Chinese	50	24, 639	952

Table 13. Statistical segmentation information of the corpus

Table 14 displays the agreement between the two annotators for the Spanish subcorpus.

[51] For the Central Unit (CU) annotation, we assign the Spanish-Chinese bilingual expert as the annotator A1, the Spanish native speaker as the annotator A2, and the Chinese native linguist as the annotator A3.

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

Table 14. CU annotation evaluation result of the Spanish subcorpus

The K result (0.961) in Table 14 means, for the CU annotation in the Spanish subcorpus, the agreement between the two annotators (A1 and A2) is almost perfect.

Table 15 displays the annotation agreement between the two annotators for the Chinese subcorpus.

A1	A3		Total	Kappa
	Yes	No		
Yes	55	13	68	0.977
No	7	878	885	
Total	62	881	953	

Table 15. CU annotation evaluation result of the Chinese subcorpus

Table 15 explains the K result of 0.977 for the Chinese subcorpus CU annotation. The result shows the agreement between the two annotators (A1 and A3) is also almost perfect.

5.2.2 Discussion of the CU Annotation Results

After finishing the evaluation part of CU annotation, we carry out a qualitative analysis to analyze the causes of the annotation disagreement. Among the 50 Spanish texts, the disagreement between the annotators fall on the following texts: BMCS2, FICB4, ICEG1, ICEG2, ICP6, ICP7, TERM38, TERM39 and TERM51. Table 16 includes the annotation results of these texts.

Texts name	Kappa result
BMCS2	0.845
FICB4	0.688
ICEG1	0.918
ICEG2	0.516
ICP6	0.877
ICP7	0.25
TERM38	0.959
TERM39	0.818
TERM51	0.628

Table 16. Evaluation results of CUs annotation (Spanish subcorpus)

From Table 16, we can see although disagreement exist in these texts, the lower results (<0.81) are in the following texts: FICB4(0.688), ICEG2(0.516), ICP7(0.25),

and TERM51(0.628). Here we explain the CU annotations of these texts and the analysis of the annotation disagreements.

- Partly match

(Ex.1) Text name: FICB4

Annotator (A1): [Con el objetivo de difundir la cultura china,] [mostrar la belleza de los parajes tibetanos,] [y profundizar en el conocimiento que del Tíbet tiene la población catalana y china residente en Barcelona,] [se celebra la primera exposición anual de la serie "Exposición Fotográfica sobre las Minorías Étnicas Chinas" dedicada al Tíbet.] (×)[52]

English: [With the aim of spreading the culture Chinese,] [show the beauty of landscapes Tibetan,] [and deepen the knowledge that Tibet has the population Catalan and Chinese residing in Barcelona,] [it celebrates the first exhibition annual of the series "Exhibition Photographic on Ethnic Minorities Chinese" dedicated to Tibet.]

Annotator (A2): [se celebra la primera exposición anual de la serie "Exposición Fotográfica sobre las Minorías Étnicas Chinas" dedicada al Tíbet.]

English: [it celebrates the first exhibition annual of the series "Exhibition Photographic on Ethnic Minorities Chinese" dedicated to Tibet.] (√)

Analysis: From the annotations of the two annotators, we can see that although the annotation of the CU is partly match, the annotator A1 annotates more context besides the CUs. The main topic of the text FICB4 is to introduce the the celebration of photography in Barcelona, therefore, the annotation of A1 is not adequate. The annotation of A2 is correct.

- No match

(Ex.2) Text name: ICEG2

Annotator (A1): [La planificación y organización de los talleres se lleva a cabo a lo largo del año académico] [para ir cubriendo la mayoría de los aspectos fundamentales de la cultura china.] (×)

English: [The planning and organization of the workshops takes place throughout the year academic] [to cover most of the aspects fundamental of culture Chinese.]

Annotator (A2): [La difusión de la cultura china es uno de los principales objetivos del Instituto Confucio de la Universidad de Granada.] (√)

English: [The dissemination of culture Chinese is one of the main objectives of the Institute Confucius of the University of Granada.]

Analysis: In this case, the two annotators explain the different contents as the CUs. The annotation of A1 introduces the workshop of the Confucius Institute meanwhile the annotation of A2 states the main objectives of the institute. Although the text talks about the general information of the Confucius Institute among other things, we consider the information about the objectives as the key information of the text because other information is related to the objectives.

---

[52] For the CU annotation analysis, we also use "√" to represent the correct annotation and "×" to represent the incorrect annotation.

(Ex.3) Text name: ICP7

Annotator (A1): [Es cada vez más frecuente que los usuarios se valgan de sus propios teléfonos inteligentes] [para acceder a contenidos útiles en su entorno laboral, académico o de investigación,] [leer] [y responder a los últimos correos electrónicos,] [o participar en los perfiles de la biblioteca en las redes sociales.] (×)

English: [It is increasingly common for users to use their own smartphones] [to access content useful in their work, academic or research environment,] [read] [and respond to the latest emails,] [or participate in the profiles of the library in networks social.]

Annotator (A2): [La Red de Bibliotecas del Instituto Cervantes (RBIC) se incorpora a esta filosofía de servicios relacionados con la movilidad de los contenidos gracias al MOPAC, una interfaz expresamente diseñada para dispositivos móviles (iPhone, Android, tablets e iPads), de uso sencillo, rápido y ágil.] (√)

English: [The Network of Libraries of the Institute Cervantes (RBIC) incorporates this philosophy of services related to the mobility of content thanks to MOPAC, an interface specifically designed for mobile devices (iPhone, Android, tablets and iPads), to use simple, fast and agile.]

Analysis: In this case, the two annotators have different opinions to decide the CU. The content that A1 annotates, is the main problem of the text, while A2 considers the information about the RBIC system the most important. Both annotations appear in the beginning of the text and the rest are introductions of the RBIC system. The annotation of A1 is the cause of the text but does not represent the main idea. The text discusses the new system and the annotation of A2 summarizes all the information and represents the idea of presenting the system; therefore the annotation of A2 is the correct one.

(Ex.4) Text name: TERM51

Annotator (A1): [La comunicación que deseo presentar en el Congreso Internacional de Terminología se basa en el contenido de un informe solicitado por la Dirección de Estudios e Información de la Viceconsejería de Política Lingüística del Gobierno Vasco que fue presentado en la primera reunión que celebra la Comisión Especial de Toponimia del Consejo Asesor del Euskera el pasado mes de noviembre en Donostia-San Sebastián.] (×)

English: [The communication that I wish to present at the Congress International of Terminology is based on the content of a report requested by the Directorate of Studies and Information of the Vice-Ministry of Linguistic Policy of the Basque Government that was presented at the first meeting held by the Special Commission on Toponymy of the Advisory Council of the Euskera last month of November in Donostia-San Sebastián.]

Annotator (A2): [definiremos el concepto de topónimo en relación con el término geográfico.] [A continuación señalaremos algunos puntos de reflexión en torno al tema que nos ocupa:] [Tras esta presentación la comunicación giraría en torno a:] [1.- Análisis de la problemática desde el punto de vista de “usuario de la terminología geográfica”.] [2.- Importancia y necesidad de normalizar la

terminología geográfica dentro de los trabajos de normalización toponímica.] [3.- Resumen del tratamiento del que ha sido objeto el término geográfico dentro de los trabajos de normalización toponímica llevada a cabo por el Instituto DEIKER de la Universidad de Deusto.] [4- Conclusiones. Ruegos y preguntas.] (√)

English: [we will define the concept of toponymy in relation to the term geographical.] [Next we will indicate some points of reflection on the subject at hand:] [After this presentation the communication would revolve around:] [1.- Analysis of the problems from the point of view of "user of the geographic terminology".] [2.- Importance and need to standardize the terminology geographical within the works standardization toponymic.] [3.- Summary of the treatment of the geographic term has been object within the works of normalization toponymic carried out by the Institute DEIKER of the University of Deusto.] [4- Conclusions. Questions and answers.]

Analysis: Same as the previous case, the two annotators takes different contents as the main topic of the text. The text TERM51 analyzes the problems that appear during the nominalization process for geographical terminology. The annotation of A1 is the background of the text; the function of the background is to support the main topic of the text. The content of the background cannot be considered as the main topic of the text. Although the contents of the main topic have been separated as independent parts, the annotator A2 annotates all the correct contents.

- Special case

(Ex.5) Text name: ICP6

Annotator (A1): [El día 14 de julio de 2006 el Instituto Cervantes asumió el famoso compromiso, esbozado por Miguel de Cervantes, en el prólogo de la segunda parte del Quijote: la dedicatoria al Conde de Lemos.] (×)

English: [On July 14, 2006 the Cervantes Institute assumed the famous commitment, outlined by Miguel de Cervantes, in the prologue to the second part of Don Quijote: the dedication to the Count of Lemos.]

Annotator (A2): [Pero la presencia del Instituto Cervantes en China se entiende no sólo como vehículo para difundir la lengua y la cultura en español, sino también como plataforma abierta al diálogo con personas e instituciones de este país que nos acoge,] (√)

English: [But the presence of the Cervantes Institute in China is understood not only as a vehicle for spreading the language and culture in Spanish, but also as an open platform for dialogue with people and institutions of this country that welcomes us,]

Analysis: The text ICP6 is a special case as it talks about the general information about the Cervantes Institute of Beijing from different aspects; thus there is no main topic of the text. For this special case, we decided the content of the objectives is the main topic, because the other aspects are all related with the objectives. The annotation of A1 presents an activity of the institute, and the annotation of A2 explains the main objective of the institute. In consequence, we think the annotation of A2 is the correct annotation.

For the Chinese subcorpus, among the 50 Chinese texts, the disagreements between the two annotators lay in the following texts: BMCS2, CCICE3, CCICE5, EEP4, ICP6, TERM25, TERM29, TERM31 and TERM51. Table 17 contains the annotation results of these texts.

Texts name	Kappa result
BMCS2	0.743
CCICE3	0.891
CCICE5	0.818
EEP4	0.333
ICP6	0.92
TERM25	0.764
TERM29	0.764
TERM31	0.966
TERM51	0.938

Table 17. Evaluation results of CUs annotation (Chinese subcorpus)

From Table 17, we can see although disagreement exist in these texts, the lower results (<0.81) are in the following texts: FICB4 (0.688), ICEG2 (0.516), ICP7 (0.25), and TERM51 (0.628). Here we explain the CU annotations of these texts and the analysis of the annotation disagreements:

- Partial match

(Ex.1) Text name: BMCS2

Annotator (A1): [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。] [我们的教材为西班牙原版教材，] [内容新颖，] [适用于中国学生学习。] (√)

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

Annotator (A3): [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 teachers are Spanish native professors,] [trained in professional Spanish language teaching (ELE),] [and has in China rich teaching experience.]

Analysis: This text concentrates on the introduction of the teachers and teaching method. Annotator A1 considers both the content about the teachers and the content discussing the teaching materials as representing the main topic. The annotation of A3 just explains the information about the teachers. Therefore, we think the annotation of the annotator A1 is correct.

- No match

(Ex.2) Text name: EEP4

Annotator (A1): [论坛旨在重申“新丝绸之路”的倡议，] [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”。] (×)

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

Annotator (A3): [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。] (√)

English: [October 28 and 29, by the State Council Development Research Center, International Relations and Sustainable Development Center, the Chinese Embassy in Spain and Toledo International Peace Center co-sponsored the second "Silk Road International Forum 2015" Held in Madrid.]

Analysis: The two annotators select different contents as the CUs. The annotator A1 thinks the aim of the forum is the main information of the text while the annotator A3 considers the content of introducing the forum is the main information of the text. Since the text EEP4 presents the international conference between Spain and China, we confirm the correct annotation is the annotator A3.

(Ex.3) Text name: TERM25

Annotator (A1): [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）] [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，] (√)

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)] [as well as the legal system must be capable of deep content Basque mastery (Spain, France and EU law) integrated in a document,]

Annotator (A3): [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: [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.]

Analysis: The text talks about the terminology translation and how to carry out the translation task. The annotation of annotator A1 matches the main idea of the text; the annotation summarizes the whole content of the text as the key information. The annotation of annotator A3 do not represent the main idea of the text. The annotation of the annotator A1 is correct.

(Ex.4) Text name: TERM29

Annotator (A1): [[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，] [还应考虑在术语和信息学方面采用不同的方法论。] [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果。] (√)

English: [This also prompted us to conducting specific research, not only to take into account the above theoretical principles,] [also should consider in terms of terminology and informatics using different methodologies.] [At the same time, we also should face these problems, conduct research and integration of related terminology concordance results .]

Annotator (A3): [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，] (×)

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,]

Analysis: The topic of the text, TERM29, is the design and management of a database. The annotation from A1 shows how to design the database and how to manage it. The annotation of A3 is the background of the text and cannot be considered as the main idea of the text. The annotation of A1 is correct.

- Special case: high agreement result but no match

(Ex.5) Text name: ICP6

Annotator (A1): [然而，塞万提斯学院在中国的设立不仅要为西班牙和其文化的推广提供一个媒介，更是要为中国的众多文化机构和文化人士提供一个开放的平台，] (√)

English: [However, Cervantes Institute in China of establishment not only for Spain and its culture of promotion offers a medium, but also for China of various cultural institutions and people offer a more open platform]

Annotator (A3): [西班牙塞万提斯学院在遥远的东方——中国首都北京，成立了自己的分院——北京塞万提斯学院，] (×)

English: [Spain Cervantes Institute in the distant East - the China capital Beijing, set up its own branch - Beijing Cervantes Institute,]

Analysis: As we have analyzed in the Spanish corpus, the text, ICP6, is a special case. It talks about the Beijing Cervantes Institute from various aspects and each part is equally important. Same as the Spanish case, we consider that the objective part is the most important idea of the text since other aspects are all related with the objectives. The annotation of annotator A1 is the objective and the annotation of the annotator A3 is not. As a result, the annotation of the annotator A1 is correct.

(Ex.6) Text name: TERM31

Annotator (A1):[IXA 研究组试图针对巴斯克语开发一个此类工具，] (√)

English: [IXA research group trying to develop a Basque for such tools,]

Annotator (A3): [近年来，各语种都在开发科技类文章术语的自动构建工具，] (×)

English: [In recent years, the minority languages in the development of science and technology terminology of the automatic construction tools,]

Analysis: The text, TERM31, presents the automatic construction of tools for the Basque language and related applications. Although the text uses a lot of information to introduce different methods, the main idea is to detect how to use these methods. This is what annotator A1 explains, that the group is developing a tool for Basque. The annotation of annotator A3 is the background of the text, not the central information of the text.

(Ex.7) Text name: TERM51

Annotator (A1):[我们需首先确认与地理术语相关的地名的概念。] [下文我们将指出几点关于这个话题的思考：] [在陈述完上述观点后，本文将围绕下列问

题展开：] [1.- 从“地理学术语使用者”的角度来分析其中出现的问题。] [2.- 地名标准化中地理专业术语标准化的重要性和必要性。] [3.- 德乌斯特大学 DEIKER 学院进行的地名标准化工作中研究的地理术语处理项目总结。] [4.- 小结。致辞和提问。] (√)

English: [We need to first confirm with geographical term related names of concept.] [Following we will point out few thoughts about the topic thoughts:] [After presenting above points, the article will focus on following issues:] [1. - From “geography term user” of point of view to analyze the problems that arise.] [2. - The names standardization process geographical term standardization of importance and necessity.] [3. - Deusto University DEIKER Institute carries out of geographical name tasks of study geographical terms project summary.] [4. - Summary. Speech and questions.]

Annotator (A3): [近十年来，德乌斯特大学 DEIKER 研究院进行了一系列巴斯克自治区的地名标准化工作，从中发现了地名学术语标准化的必要性。] (×)

English: [Over the past decade, Deusto University the DEIKER Institute has carried out a series of Basque Autonomous Region work on the names of the standardization, among them find the standardization of terminology necessity.]

Analysis: As we have analyzed for the Spanish subcorpus, the main information of TREM51 is separated in different parts. It analyzes the problems that appear during the nominalization process for geographical terminology. The annotation of annotator A1 presents all the main ideas of the text while the annotation of annotator A3 is additional information about the progress made by the Deusto University on the project. The annotation of the annotator A1 is correct.

Based on the previous analysis, we also organize the words that appear more times in the CUs as different CU indicators, as indicated in Section 4.3.1. Table 18 presents these indications for both Spanish and Chinese parts.

Noun	Verb	Proper Noun	Preposition	Pronoun	Conjunction
difusión[53] (传播/推广) (diffusion)	ofrecer (提供) (offer)	nuestro (我们) (we/us/ourselves)	desde (自. . 以来) (from)	este (本/此) (this)	no solo, sino (que) también (不仅..同时) (not only...but also)
misión (任务) (task)	intentar (旨在) (intent to/aims to)		para (为) (for)	esta (本/此) (this)	y (并/以及/还) (and/also)
propósito[54] (旨在) (purpose)	plantear (阐述) (think about)		según (据) (based on)		
finalidad (目标) (goal)	tratar (描述) (describe)				
	pretender (旨在) (aims to)				

Table 18. The indications of CU in the research corpus

[53] The Spanish word "*difisión*" ('diffusion') is a noun, however, in the corpus, its Chinese translation "*chuanbo*" (传播) ('to broadcast') and "*tuiguang*" (推广) ('to broadcast') are verbs in Chinese.

[54] The Spanish word "*propósito*" ('purpose') is a noun, however, in the corpus, its Chinese translation "*zhizai*" (旨在) ('aims to') is the verb in Chinese.

In the Appendix part, we give parallel examples of each Spanish-Chinese word pair in the corpus and the English literal translation for each example[55]. The above listed indication words can be used for other NLP studies, for example, information extraction, question-answering, summarization, etc.

5.3 Discourse Structure Annotation

Discourse structure annotation is the last step of annotation. As we explained in Section 4.4.1, a Spanish-Chinese bilingual expert annotates all the texts (100 texts) in the corpus. The two Spanish native speaking experts annotate all the 50 Spanish texts (25 texts for each), and the Chinese native speaking expert annotates all 50 Chinese experts[56].

In this section, we will evaluate the annotation quality of the Nuclearity (N), Relation (R), Composition (C), and Attachment (A) elements by using F-measure. The evaluation method follows a qualitative method that was especially designed for RST discourse annotation by Iruskieta, da Cunha and Taboada (2015).

5.3.1 Discourse Structure Annotation Results

Under the qualitative evaluation method, the used statistics method is F-measure. The agreement of the four elements (Nuclearity, Relation, Composition , and Attachment) are checked. Table 19 shows the evaluation result of the Spanish text BMCS1and Table 20 explains the evaluation result of the parallel Chinese text as examples.

---

[55] For the given examples, there are some repeated examples. This is becasue some CUs contain more than one key words.

[56] For the discourse structure annotation, the Spanish-Chinese bilingual annotator is considered as the annotator A1, the two Spanish annotators are considered as A2, and the Chinese annotator is considered as the annotator A3.

A1				A2				Qualitative evaluation					
Central Subconstituent (CS)	Realtion (R)	Constituent (C)	Attachment point (A)	Central Subconstituent (CS)	Realtion (R)	Constituent (C)	Attachment point (A)	N	R	C	A	Agreement	Dsiagreement
1	Preparation	1S	(2-12)N	1	Preparation	1S	(2-12)N	√	√	√	√	NRCA	
3-12	Interpretation	(3-12)S	2N	3-12	Elaboration	(3-12)S	2N	√	×	√	√	NCA	R
3-5|6-12	List	(3-5)N|(6-12)N		3-5|6-12	List	(3-5)N|(6-12)N		√	√	√	√	NRCA	
4-5	Elaboration	(4-5)S	3N	4-5	Elaboration	(4-5)S	3N	√	√	√	√	NRCA	
4|5	List	4N|5N		4|5	List	4N|5N		√	√	√	√	NRCA	
6	Preparation	6S	(7-12)N	6|7-12	Same-unit	6N|(7-12)N		×	×	×	×	×	NRCA
7|8|9|10|11|12	List	7N|8N|9N|10N|11N|12N		7|8|9|10|11|12	List	7N|8N|9N|10N|11N|12N		√	√	√	√	NRCA	

Nuclearity		Relation		Composition		Attachment	
Match	F	Match	F	Match	F	Match	F
6 of 7	0.857	5 of 7	0.714	6 of 7	0.857	6 of 7	0.857

Table 19. Qualitative evaluation for the discourse annotation of the Spanish text BMCS1

A1				A3				Qualitative evaluation				
Central Subconstituent (CS)	Realtion (R)	Constituent (C)	Attachment point (A)	Central Subconstituent (CS)	Realtion (R)	Constituent (C)	Attachment point (A)	N	R	C	A	Agreement
1	Preparation	1S	(2-14)N	1	Preparation	1S	(2-14)N	√	√	√	√	NRCA
2|3	List	2N|3N		2|3	List	2N|3N		√	√	√	√	NRCA
4-14	Interpretation	(4-14)S	2N	4-14	Elaboration	(4-14)S	2N	√	×	√	√	NCA
5-6	Elaboration	(5-6)S	4N	5-6	Elaboration	(5-6)S	4N	√	√	√	√	NRCA
6	Interpretation	6S	5N	6	Interpretation	6S	5N	√	√	√	√	NRCA
7	Preparation	7S	(8-14)N	8-14	Elaboration	(8-14)S	7N	×	×	×	×	×
8|9-10|11|12-13|14	List	8N|(9-10)N|11N|(12-13)N|14N		8|9-10|11|12-13|14	List	8N|(9-10)N|11N|(12-13)N|14N		√	√	√	√	NRCA
4-6|7-14	Elaboration	(4-6)N|(7-14)N		4-6|7-14	Elaboration	(4-6)N|(7-14)N		√	√	√	√	NRCA
10	Elaboration	10S	9N	10	Elaboration	10S	9N	√	√	√	√	NRCA
12|13	List	12N|13N		12|13	List	12N|13N		√	√	√	√	NRCA

Nuclearity		Relation		Composition		Attachment	
Match	F	Match	F	Match	F	Match	F
9 of 10	0.9	8 of 10	0.8	9 of 10	0.9	9 of 10	0.9

Table 20. Qualitative evaluation for the discourse annotation of the Chinese text BMCS1

From Table 19, we can see that the annotation disagreements are: Relation (R) and no match. For the disagreement of Relation, annotator A1 considers the relation between the EDU(2) and the EDU(3-12) as ELABORATION, while annotator A2 thinks the relation between the two EDUs should be INTERPRETATION. The annotation case matches the source of disagreement: “A relation is more specific than the other”. The disagreement of NRCA reflects the source of disagreement: “Different choice in nuclearity entailed a N/N-N/S mix-up”..

Regarding the annotation case in Table 20, we can see that the annotation disagreements are also R and NRCA. However, the annotation disagreement of EDU(7) and EDU(8-14) reflects the source of disagreement: “Different choice in attachment entailed a different relation”, which is different from the case of NRCA in its parallel Spanish text.

Table 21 includes the evaluation results of each part in the Spanish subcorpus and Table 22 contains the evaluation results of each part in the Chinese subcorpus.

Source	**Nuclearity**		**Relation**		**Composition**		**Attachment**	
	Match	F	Match	F	Match	F	Match	F
ICT	290 of 315	0.921	268 of 315	0.851	290 of 315	0.921	288 of 315	0.914
SMCL	51 of 67	0.761	43 of 67	0.641	51 of 67	0.761	49 of 67	0.731
CCICS	37 of 41	0.902	30 of 41	0.732	36 of 41	0.878	37 of 41	0.902
SEB	54 of 57	0.947	50 of 57	0.877	54 of 57	0.947	53 of 57	0.930
SCCF	46 of 50	0.92	37 of 50	0.74	46 of 50	0.92	45 of 50	0.90
CIFB	39 of 44	0.886	34 of 44	0.773	39 of 44	0.886	38 of 44	0.864
BCI	96 of 108	0.889	83 of 108	0.769	96 of 108	0.889	96 of 108	0.889
GCI	15 of 15	1	15 of 15	1	14 of 15	0.933	14 of 15	0.933

Table 21. Qualitative evaluation results of the Spanish subcorpus

Source	Nuclearity		Relation		Composition		Attachment	
	Match	F	Match	F	Match	F	Match	F
ICT	313 of 357	0.877	278 of 357	0.779	313 of 357	0.877	312 of 357	0.874
SMCL	66 of 72	0.917	58 of 72	0.806	66of 72	0.917	66 of 72	0.917
CCICS	44 of 45	0.978	38 of 45	0.844	44 of 45	0.978	44 of 45	0.978
SEB	60 of 64	0.938	54 of 64	0.844	60 of 64	0.938	60 of 64	0.938
SCCF	62 of 65	0.954	51 of 65	0.785	62 of 65	0.954	62 of 65	0.954
CIFB	44 of 50	0.88	41 of 50	0.82	44 of 50	0.88	42 of 50	0.84
BCI	122 of 134	0.910	110 of 134	0.821	122 of 134	0.910	122 of 134	0.910
GCI	19 of 22	0.864	16 of 22	0.727	19 of 22	0.864	19 of 22	0.864

Table 22. Qualitative evaluation results of the Chinese subcorpus

From Table 21, we can conclude that in the Spanish subcorpus, the range of agreement for Nuclearity is from 0.761 to 1, the range of agreement for Relation is from 0.641 to 1, the range of agreement for Composition is from 0.761 to 0.947, and the range of agreement for Attachment is from 0.731 to 0.933. The annotation evaluation results of the Chinese subcorpus shows that the range of agreement of Nuclearity is from 0.864 to 0.978, the agreement of the Relation is from 0.727 to 0.844, the agreement of the Composition is from 0.864 to 0.978, and the agreement of the Attachment is from 0.84 to 0.978. The evaluation results prove that the annotation of the Spanish subcorpus and the annotation of the Chinese subcorpus are reliable.

After getting reliable annotation results for both the Spanish and Chinese subcorpora, we then harmonized the corpus to carry out the qualitative analysis. The most essential step in order to do qualitative comparison between two different languages is harmonization, where the segments must be aligned and must contain the same number of EDUs, to avoid confusing analysis disagreement and segmentation agreement (Iruskieta, da Cunha and Taboada, 2015). The following example from our corpus shows the harmonization case:

(Ex.1) Text name: CCICE3

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

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

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

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

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

From the above example, we can see the Spanish message is an interdependent EDU and its parallel Chinese message contains three EDUs. For the harmonization, Iruskieta, da Cunha and Taboada (2015) suggest the simple rule, we combine three Chinese EDUs as an independent EDU. Although the harmonization process erases some rhetorical relations, the higher level of RS-Tree structure is not affected. Table 23 contains the comparison results of the harmonized corpus.

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

Table 23. Qualitative evaluation of the harmonized corpus between Spanish and Chinese

Table 23 informs us that in the harmonized corpus, the range of agreement of the Nuclearity is from 0.855 to 1, the range of agreement of the Relation is from 0.794 to 1, the range of agreement of the Composition is from 0.761 to 0.947, and the range of agreement of the Attachment is from 0.731 to 0.933. In the harmonized Chinese subcorpus, the range of agreement of the Nuclearity is from 0.761 to 1, the range of agreement of the Relation is from 0.855 to 1, the range of agreement of the Composition is from 0.855 to 1, the range of agreement of the Attachment is from 0.855 to 0.933. The evaluation results of the harmonized corpus are better than the original corpus because of the removal of the annotation disagreements during the harmonized process for both the Spanish subcorpus and the Chinese subcorpus.

5.3.2 Discussion of Discourse Structure Annotation Results

After getting the qualitative evaluation results of the corpus, we analyze the annotation disagreements between the Spanish and Chinese. For each type of disagreements, we give an example extracted from the corpus.

- Difference of relation definition

(Ex.1) Text name: BMCS5

*Relation definition:*MEANS

Spanish: [A través de sus actividades y módulos teóricos,]$_S$ [introduce a los profesores en los principios y las estrategias básicas de la formación virtual aplicados específicamente al entorno del AVE.]$_N$

English: [Through its activities and modules theoretical,]$_S$ [introduce to the teachers to the principal and the strategies basic of the training virtual applied specially to the environment of the AVE]$_N$

Relation definition: PURPOSE

Chinese: [通过练习和理论模块，]$_S$ [指导教师掌握 AVE 课程适用的虚拟平台的基本教学原则和战略。]$_N$

English: [Through the activity and theoretical modules,]$_S$ [to guide the teachers to know the AVE courses applied of virtual platform of basic teaching principles and strategies.]$_N$

Explanation: The definition of the relation for the Spanish message is MEANS, meanwhile there is a PURPOSE between the two Chinese EDUs. In this example, the nuclear part shows an event where all the teachers need to control the required demands, and the satellite part proposes a way to make the event become true. Therefore, we think the MEANS relation is more adequate in this case.

- Difference of positions between Constituent (C) and Attachment (A)

(Ex.2) Text name: CCICE1

*Relation definition:*CAUSE

Spanish: [En 2015, por la primera vez, la región de Norte américa se convierte en el tercer feudo por primas de Mapfre,]$_N$ [desplazando en esa posición a Latam Sur.]$_S$

English: [In 2015, for the first time, the region of North America converts to the third fief by premium of Mapfre,]$_N$ [displacing in that position to Latam South.]$_S$

Relation definition: CAUSE

Chinese: [在保险方面，北美已超越南美，]$_S$ [上升为西班牙保险公司 Mapfre 第三大市场。]$_N$

English: [In insurance, North America has surpassed South America,]$_S$ [rose to the Spanish insurance company Mapfre the third-largest market.]$_N$

Explanation: From this example, we can see that the positions of nuclear and satellite in the Spanish message and Chinese message are reversed. In the Spanish message, the nuclear comes before the satellite; while in the Chinese message, the satellite comes first. Hence, for the qualitative evaluation, disagreements exist for the elements C and A.

- Difference of expression

(Ex.3) Text name: EEP3

*Relation definition:*PURPOSE

Spanish: [Se acordó establecer una hoja de ruta,]$_N$ [para identificar áreas estratégicas de interés mutuo y avanzar en proyectos concretos hasta la próxima Comisión mixta que se celebrará en 2017 en España.]$_S$

English: [It was agreed to establish a map of routine,]$_N$ [to identify areas strategic of interests mutual and to advance in projects concrete until the next Commission mixed to be held in 2017 in Spain.]$_S$

Relation definition: LIST

Chinese: [双方同时还肯定了战术领域方面的共同战略领域，]$_S$ [推进具体项目，到 2017 年在西班牙举行的下一届科技联委会。]$_N$

English: [Both sides also affirmed strategic areas of tactic area,]$_S$ [advancing specific projects, till 2017 in Spain to be held of next Science and Technology Joint Commission.]$_N$

Explanation: In the Spanish message, the word *para* ('for') in the satellite represents a PURPOSE relation. However, in the Chinese message, the sentence is a coordinated clause, and thus the subject is erased in the satellite. The relation between the nuclear and satellite is LIST. Although the expressions are different between the parallel passages, the meaning in the parallel content is the same. Both messages transmit the information about the cooperation between Spain and China. For the case of different expressions, we keep the different relation definitions.

- Special case

(Ex.4) Text name: FICB2

Spanish: [Como conclusión de la formación, los asistentes compartieron dudas y experiencias.]$_N$ [Todos los asistentes recibieron los certifciados de participación de Hanban y de la FICB.]$_S$

English: [As the conclusion to the training, the assistants shared doubts and experiences.]$_N$ [All attendees received the certificates of participation of Hanban and the FICB.]$_S$

Chinese: [之后进行了圆桌会议的讨论，全体与会教师就汉字书写等问题进行了讨论，并就海外汉语教学中的疑惑和经验展开了深入的交流。]$_N$ [培训结束后，我院为参加本次培训的每位教师颁发了汉办制作和巴塞罗那孔子学院制作的教学培训证书。]$_N$

English: [After that, roundtable meeting of discussion, all the participating teachers discussed the writing of Chinese characters and other issues, and teaching Chinese overseas conducted in-depth exchanges.]$_N$ [After the training, our institute awarded the each for participating the training of teacher the Hanban and FICB made of teaching and training certificate.]$_N$

Explanation: From Figure 21 and Figure 22 we can see that, in the Spanish message, EDU12 and EDU13 hold a ELABORATION relation, and together with EDU10 and EDU11, the four EDUs EDUs form a paragraph. Notwithstanding, in the Chinese passage, EDU12 and EDU13 are not in the same discourse level. EDU13 is an independent paragraph, meanwhile EDU10, EDU11 and EDU12 form another paragraph. For this special case, we decide not to make any modification to keep the original discourse structures for both texts.

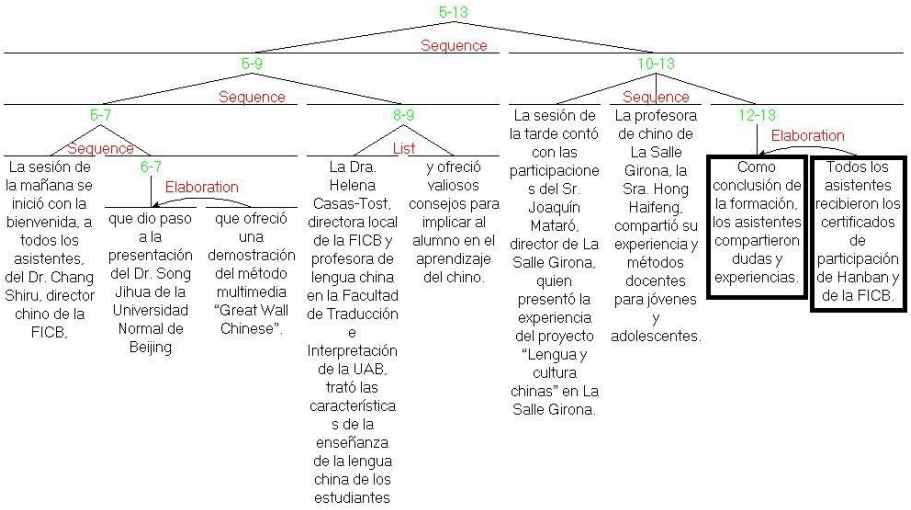

Figure 21. Special case of discourse annotation for the qualitative comparison (Spanish text)

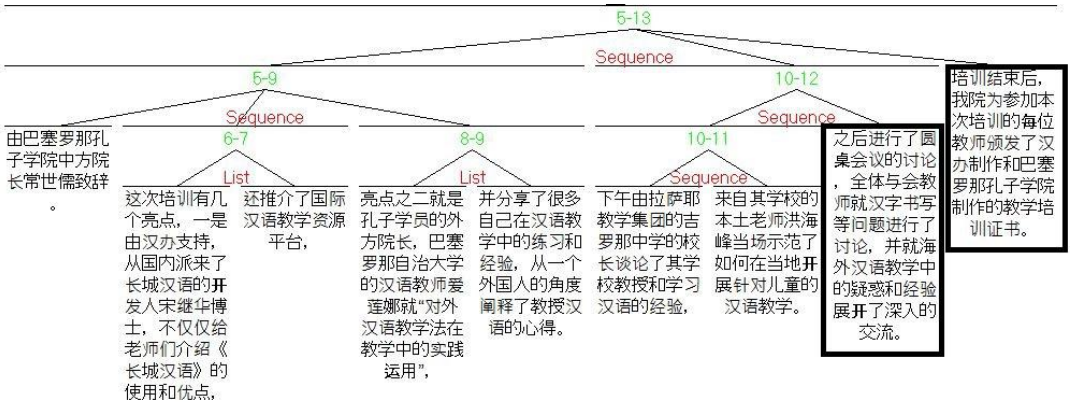

Figure 22. Special case of discourse annotation for the qualitative comparison (Chinese text)

The qualitative evaluation results of the harmonized corpus demonstrate the reliability of the annotation quality and is the Gold Standard (GS) for our study.

5.4 Chapter Overview

In this chapter, we evaluate the annotation quality for each annotation step. The evaluation measurement for segmentation and CU annotation is the Kappa measure. For the discourse structure annotation evaluation, following the qualitative method proposed by Iruskieta, da Cunha and Taboada (2015), we use F-measure to check the annotation reliability of constituent (C), attachment point (A) and the definition of relation (R).

Regarding the segmentation evaluation, for the Spanish subcorpus, the total evaluation result is 0.87 and for the Chinese subcorpus is 0.76. We carry out a qualitative analysis for the segmentation disagreement and establish the final segmentation criteria for the entire corpus. The evaluation result of the CU annotation is 0.961 for Spanish subcorpus and 0.977 for the Chinese subcorpus. Same as the

segmentation annotation, the annotators discuss the disagreements to confirm the correct CUs for each text in the corpus. The qualitative evaluation for discourse structure shows that, for the evaluated elements Nuclearity (N), Relation (R), Composition (C) and Attachment (A) in both subcorpora, all of them get the good annotation results. Additionally, we get the better evaluation results in the harmonized corpus.

Chapter 6.
Elaboration of a Spanish-Chinese Discourse Recommendation Protocol for Spanish-Chinese Translation

Based on the previous annotation evaluation results, we will talk about the following aspects in this chapter; firstly, we will present the results of discourse similarities and differences based on the annotation results (Section 6.1). Secondly, we will discuss about the discourse detected similarities and differences in terms of different translation strategies (Section 6.2). Thirdly, we will give the final recommendations protocol for Spanish and Chinese translation (Section 6.3).

6.1 Results regarding Discourse Differences and Similarities in the Corpus Annotation

Among the annotation results, we find 19 cases related with marker change. For the unit shift plus marker change cases, totally there are 2 cases and each case is different from the other 6 cases. For the other cases, in total, there are 8 cases are included. Table 24 shows the statistical conclusion of the different cases.

Translation strategy	Nº of the cases	Percentage
Marker change	19	63.3%
Unit shift	2	6.7%
Other cases	9	30%
Total	30	100%

Table 24. Statistics of the translation strategies found in the corpus annotation

6.2 Discussion of the Detected Discourse Differences and Similarities

In this section, we analyze the detected discourse differences and similarities based on the different translation strategies proposed by Iruskieta, da Cunha and Taboada (2015) that we have already mentioned in Section 4.5.2.

- **Marker change**

For the 19 cases of marker change, there are 13 cases that do not include any DM in the Spanish passage, but there is a DM in the Chinese passage. The rest 6 cases are the change of the Spanish DMs in the Chinese parallel passages, each of them is independent and different from others. Table 25 presents the appearances of the distinct cases of these 19 cases.

Spanish passage	Chinese passage	Appearances	Percentage
No DM	A new DM	13	68.4%
A DM	Another different DM	6	31.6%
Total		19	100%

Table 25. Information of the different discourse cases in the corpus

For the 13 cases whose Spanish passages do not contain any DM, but their parallel Chinese passages contain DMs, the DM "*yinci*" (因此) ('therefore' in English) is added in other 3 Chinese passages. Meanwhile, there are 3 cases that add the DM "*bing*" (并) ('and' in English) in the Chinese passages. The rest 7 cases are independent and different. The 7 cases are: (i) the added Chinese DM is "*wei*" (为) ('for/aims to' in English), (ii) the added Chinese DM is "*lingyifangmian*" (另一方面) ('on the other hand' in English), (iii) the added Chinese DM is "*yucitongshi*" (与此同时) ('at the same time' in English), (iv) the added Chinese DM is "*ruo*" (若) ('if' in English), (v) the added Chinese DM is "*ye*" (也) ('also' in English), (vi) the added Chinese DM is "*dang*" (当) ('when' in English), and (vii) the added Chinese DM is "*er*" (而) ('however' in English). Based on the number of appearances and alphabet of the names of the texts, Table 26 summarizes the information about the 13 cases.

Spanish passage	Text name	Chinese passage	Appearances	Percentage
No DM	CCICE3 FCEC1 TERM31	"*yinci*" (因此)	3	23.08%
No DM	TERM18 TERM32 TERM38	"*bing*" (并)	3	23.08%
No DM	BMCS2	"*wei*" (为)	1	7.7%
No DM	EEP2	"*lingyifangmian*" (另一方面)	1	7.7%
No DM	ICP5	"*yucitongshi*" (与此同时)	1	7.7%
No DM	TERM31	"*ruo*" (若)	1	7.7%
No DM	TERM31	"*ye*" (也)	1	7.7%
No DM	TERM50	"*dang*" (当)	1	7.7%
No DM	TERM50	"*er*" (而)	1	7.7%
Total			13	100%

Table 26. Statistics of cases including a DM only in the Chinese passages

Moreover, we give the discourse relations in the above mentioned 13 Spanish passages and their parallel Chinese passages. All the relations are included in Table 27. The presented order follows the order presented in Table 26.

Language	DM	Discourse relation
Spanish	/	Elaboration
Chinese	"*yinci*" (因此)	Result
Spanish	/	Condition
Chinese	"*bing*" (并)	List
Spanish	/	Elaboration
Chinese	"*wei*" (为)	Purpose
Spanish	/	Elaboration
Chinese	"*lingyifangmian*" (另一方面)	List
Spanish	/	Summary
Chinese	"*yucitongshi*" (与此同时)	Conjunction
Spanish	/	Evaluation
Chinese	"*ruo*" (若)	Condition
Spanish	/	Elaboration
Chinese	"*ye*" (也)	List
Spanish	/	Contrast
Chinese	"*dang*" (当)	Circumstance
Spanish	/	Contrast
Chinese	"*er*" (而)	Contrast
Spanish	y	List
Chinese	"*huozhe*" (或者)	Disjunction
Spanish	y	List
Chinese	"*zhizai*" (旨在)	Purpose
Spanish	igualmente	List
Chinese	"*tongshi*" (同时)	Conjunction
Spanish	para	Purpose
Chinese	"*ruo*" (若)	Condition
Spanish	para que	Purpose
Chinese	"*ruo*" (若)	Condition
Spanish	por lo tanto	Result
Chinese	"*dan*" (但)	Concession

Table 27. Different discourse relations for Spanish-Chinese parallel contents

For the Spanish DMs whose Chinese translations are different, the 6 cases are independent and each of them is different from others. The 6 different cases are: (i) The Spanish DM *y* is translated to "*huozhe*" (或者) ('or' in English); (ii) The Spanish DM *y* is translated to "*zhizai*" (旨在) ('aims to' in English); (iii) The Spanish DM *igualmente* is translated to "*tongshi*" (同时) ('at the same time / meanwhile' in English); (iv) The Spanish DM *para* is translated to "*ruo*" (若) ('if' in English); (v) The Spanish DM *para que* is translated to "*ruo*" (若) ('if' in English); (vi) The

Spanish DM *por lo tanto* is translated to "*dan*" (但) ('but' in English). Table 28. shows the detailed information of the different DMs between Spanish passages and Chinese passages, the presented order follows the alphabet of the names of the texts.

Text name	Spanish passage	Chinese passage	Appearances	Percentage
BMCS3	y	"*huozhe*" (或者)	1	16.7%
ICP6	y	"*zhizai*" (旨在)	1	16.7%
ICP7	igualmente	"*tongshi*" (同时)	1	16.7%
TERM31	para	"*ruo*" (若)	1	16.7%
TERM31	para que	"*ruo*" (若)	1	16.7%
TERM38	por lo tanto	"*dan*" (但)	1	16.7%
Total			6	100%

Table 28. Statistics of cases including different DMs in parallel passages

Based on the different relations for the parallel passages, we also find the similarities based on the relations that are presented in terms of DMs. To carry out the similarity comparison part, we use the following steps:

(i). For the cases that the Spanish passage do not contain any DM, but there is a new DM in its parallel Chinese text, our analysis starts from the Chinese DM. Based on the Chinese DM, we find the original Spanish translation in the corpus. Detailed information can be consulted in Table 29.

Search direction	Relation	Language	DM
Chinese → Spanish	Result	Spanish	por tanto
		Chinese	"*yinci*" (因此)
	List	Spanish	y
		Chinese	"*bing*" (并)
	Purpose	Spanish	para
		Chinese	"*wei*" (为)
	List	Spanish	por otra parte
		Chinese	"*lingyifangmian*" (另一方面)
	Conjunction	Spanish	también
		Chinese	"*yucitongshi*" (与此同时[57])

Table 29. The correct Spanish translations of the Chinese DMs in the corpus

[57] In our corpus, the Chinese DM "*yucitongshi*" (与此同时) only appears once. Therefore, we cannot give the original Spanish translation for this case. The Spanish translation of the DM "*yucitongshi*" (与此同时) is made by the author.

(ii). For the cases that the parallel contents contain their own DMs, our analysis starts from the Spanish text. Based on the DM in the Spanish passage, we find the correct Chinese translation of the corresponded Spanish DM in the corpus. Table 30 shows the adequate Chinese translation of the Spanish DMs.

Search direction	Relation	Language	DM
Spanish → Chinese	Condition	Spanish	si (BMCS3)
		Chinese	"*ruo*" (若)
	List	Spanish	y
		Chinese	"*ye*" (也)
	Circumstance	Spanish	cuando
		Chinese	"*dang*" (当)
	Contrast	Spanish	mientras
		Chinese	"*er*" (而)
	List	Spanish	y
		Chinese	"*bing*" (并)
	List	Spanish	igualmente
		Chinese	"*ye*" (也)
	Purpose	Spanish	para
		Chinese	"*yibian*" (以便)
	Purpose	Spanish	para que
		Chinese	"*yibian*" (以便)
	Result	Spanish	por lo tanto
		Chinese	"*yinci*" (因此)

Table 30. The correct Chinese translations of their Spanish DMs in the corpus

- **Unit shift**

Regarding the unit shift cases, there are 2 cases in total. Among the 2 cases, each case is different from another.

- **Other cases**

As previously indicated, for the different language forms, there are 8 cases in total. Among the 8 cases, 3 of them are the cases of unit shift plus marker change; 2 of them are the same relation type but different EDU orders. The rest 3 cases are independent and different from other cases. Table 31 summarizes the information of different language form cases.

Case	Nº of the case	Percentage
Unit shift + marker change	3	37.5%
Different order of EDU	2	25%
Special cases	3	37.5%
Total	8	100%

Table 31. Statistical summary of other cases

6.3 Final Recommendation Protocol[58]

In this section, based on the results and discussions, we will give a real case extracted from the corpus and a related discourse recommendation that can be useful for Spanish-Chinese translation. As mentioned in Section 4.5.3, the recommendations are grouped by the translation strategies: A. Marker change, B. Unit shift and C. Others. All the recommendations include the following four discourse aspects: (i) DMs, (ii) Discourse relations, (iii) Relation types, and (iv) EDUs order. In addition, some recommendations included in the groups Unit shift and Others also contain other aspect called Punctuation marks.

A. Marker Change

*** Type of relation: N-N**
Group of relation: LIST
Recommendation.A1

Text Name: TERM18

Spanish: [Esta jerarquía deberá basarse en gran medida en criterios extralingüísticos como la cooperación internacional y la comunicación,]$_{8N\_Elaboration}$ [de acuerdo a los cuales en el código lingüístico llamado "moderna variedad científica del serbio" se da prioridad a los préstamos del inglés sobre la traducción y el calco estructural.]$_{9S\_Elaboration}$

English: [This hierarchy should be based largely on criteria extralinguistic as the cooperation international and the communication,] [according to which in the code linguistic called "modern variety scientific of Serbian" it gives priority to the loans of English about the translation and the structural.]

Chinese: [此外，评定这种等级时应以诸如国际合作及交流等语言之外的标准为主，参照名为"塞尔维亚语现代科学多样性"的语言学规则，]$_{8N\_List}$ [**并**在翻译和结构模仿中优先借鉴英语的有关规则。]$_{9N\_List}$

English: [In addition, such elaboration of hierarchy should be based on as international cooperation and communication other than linguistic criteria, referring to "Serbian modern scientific diversity" of linguistic criteria,] [**and** in translation and structural prioritizing English rules.]

[58] In this protocol, for the cases that we cannot give the translation recommendation, we put them in the Appendix part.

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	*bing* (并)		Chinese	List
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* there is an ELABORATION relation in the Spanish text, *then* the DM "*bing*" (并) ('and') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into LIST relation, *meanwhile* the relation type and the order of EDUs need to be changed from N-S to N-N.					

Recommendation.A2

Text Name: EEP3

Spanish: [El programa de su visita a China incluyó visitas al Observatorio Astronómico Nacional de la Academia china de las Ciencias, a la Administración Estatal de Oceanografía y a la China Certification and Inspection Group,]$_{N\_List}$ [**y** en las que se abordaron en detalle temas en materia oceanográfica, astrofísica y de energías renovables respectivamente.]$_{N\_List}$

English: [The program of his visit to China including to Observatory Astronomical National of the Academy China of the Sciences, to the Administration State of Oceanography and to the China Certification and Inspection Group,] [**and** in which addressed with in detailed topics in oceanographic, astrophysical and energy renewable respectively.]

Chinese: [此次访华行程还包括参观中科院国家天文台，国家海洋局和中国检验认证集团，]$_{N\_List}$ [**并**分别就海洋、宇航及可再生能源材料的问题进行了讨论。]$_{N\_List}$

English: [The trip to China also included visiting to the National Astronomical Observatory of the Chinese Academy of Sciences, the State Oceanic Administration and the China Inspection and Certification Group,] [**and** addressed with marine, aerospace, and renewable problems discussed.]

Discourse markers (DM)	Spanish	y	**Relation**	Spanish	List
	Chinese	*bing* (并)		Chinese	
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*y*" ('and') is at the beginning of the second EDU in Spanish, *then* the DM "*bing*" (并) ('and') can be added at the beginning of the second EDU in Chinese, *meanwhile* the order of the EDUs is not changed.					

Recommendation.A3[59]

Text Name: TERM32

Spanish: [Seguidamente, las traducciones de las palabras clave del chino son procesadas en el acuñador]$_{12N\_Means}$ [haciendo uso de la base de datos relativa a los morfemas en chino y de las reglas para la formación de palabras en chino.]$_{13S\_Means}$

English: [Following, the translations of the keywords of Chinese are processed in the coiner] [using the usage of the base of data related to the morphemes in Chinese and the rules for the formation of words in Chinese.]

Chinese: [紧接着，中文的关键字翻译由构建器运用相关中文词素数据库进行，]$_{12N\_List}$ [**并**参考中文的单词构建规则，结果显示为一系列交替进行的翻译文件，以便专家挑选。]$_{13N\_List}$

English: [Then, Chinese keywords' translation by the builder using the relevant Chinese morpheme database,] [**and** referring to the Chinese word construction rules, the results are shown as a series of alternating translation files for expert to select.]

[59] For the three cases of the Chinese DM "*bing*" (并), two cases are the same, both cases include a MEANS relation in the Spanish passages and a LIST relation in the Chinese passages. Therefore, here we just give two examples for the case of the Chinese DM "*bing*" (并). Additionally, since the previous example has already showed the discourse similarity between the DM "*bing*" (并) and it original Spanish DM *y*, therefore, in this case we do not give the similarity example.

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Means
	Chinese	*bing* (并)		Chinese	List
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is a MEANS relation, *then* the DM "*bing*" (并) ('and') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into LIST, *meanwhile* the relation type and the order of EDUs need to be changed from N-S to N-N.					

Recommendation.A4[60]

Text Name: TERM31

Spanish: [Para ello, ya está preparado el analizador morfológico (Alegria et al., 96),]$_{N\_Elaboration}$ [el lematizador/etiquetador está a punto de finalizarse (Aduriz et al., 96) y también estamos trabajando la sintaxis del nivel superficial.]$_{S\_Elaboration}$

English: [For this, already it has prepared the analyzer morphological (Alegria et al., 96),] [the lemmatizer / tagger is about to be finalized (Aduriz et al., 96) and also we are working with the syntax of level superficial.]

Chinese: [为此，已准备了词法分析器（阿莱格里亚（Alegria）等，96），]$_{N\_List}$ [词语分析器/说明性省略语分析器**也**即将开发完成（阿杜里斯（Aduriz）等，96），我们同时也在研究和句法相关的工具。]$_{N\_List}$

English: [To this end, it has prepared lexical analyzer (Alegria et al., 96),] [the lemmatizer / tagger is **also** to be developed (Aduriz et al., 96), we are also working on syntax-based analyzer.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	*ye* (也)		Chinese	List
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, and in the Spanish text there is an ELABORATION relation, *then* the DM "*ye*" (也) ('also') can be added at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a LIST, *meanwhile* the relation type and the order of EDUs need to be changed from N-S to N-N.					

[60] The recommendation A4, A7 and A9 are similar cases, the discourse relations in the Spanish texts are all elaboration, meanwhile the discourse relations in the Chinese text are all LIST. However, the contents are different, either the DMs in the Chinese texts.

Recommendation.A5

Text Name: ICP5

Spanish: [Estudiar español en nuestro instituto no es solo aprender el idioma,]$_{N\_List}$ [sino que **también** da la oportunidad de conocer y descubrir las diferentes culturas del mundo hispánico.]$_{N\_List}$

English: [Studying Spanish in our institute is not only learning the language,] [but **also** gives the opportunity to know and discover the different cultures of the world Hispanic.]

Chinese: [在我们学院学习西班牙语，不仅仅是学习语言本身，]$_{N\_List}$ [同时**也**是学习西班牙语世界的文化。]$_{N\_List}$

English: [At our institute Studying Spanish, is not only about learning the language itself,] [but **also** about learning Spanish.]

Discourse markers (DM)	Spanish	también	**Relation**	Spanish	List
	Chinese	*ye* (也)		Chinese	
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*también*" ('also') is in the second EDU in Spanish, *then* the DM "*ye*" (也) ('therefore') can be included in the second EDU in Chinese, *meanwhile* the order of the EDUs is not changed.					

Recommendation.A6

Text Name: EEP2

Spanish: [En el acto también intervinieron Montse Aguer, comisaria de la exposición y directora del Centro de Estudios Dalinianos, Hsu Fenlan, co-comisaria, Joan Manuel Sevillano, gerente de la Fundació Dalí y Adrian Cheng fundador y Presidente de Honor de la K 11. Se estima que a la inauguración han asistido alrededor de 4.000 y hasta el 15 de febrero de 2016, fecha de su clausura se espera superar los 350.000 visitants, cifra que consechó la muestra de Jean Monet que se organizó el pasado año en esos mismos espacios.]$_{N\_Elaboration}$ [La muestra está acompañada por dos exposiciones complementarias de artistas chinos centradas en la influencia histórica del surrealismo en China en los 90 y su permanencia en la escena emergente actual.]$_{S\_Elaboration}$

English: [The act also involved Montse Agur, curator of the exhibition and director of Center of Studies Dalianian, Hsu Fenlan, co-curator, Joan Manuel Sevillano, manager of Dalí Foundation and Adrian Cheng founder and Honorary Prsident of K11. It estimated that to the inauguration have attended around 4,000 and until

the 15th of February of 2016, date of its closing it is expected exceed 350,000 visitors, a figure that was collected by Jean Monet that organized the past year in those same spaces.] [The exhibition is accompanied by two exhibitions complementary of artists Chinese focus on the influence historical of surrealism in China in the 90s and its permanence in the scene emerging current.]

Chinese: [出席开幕式的还有策展人 MontseAguer，卡拉-萨尔瓦多达利基金会执行长 Juan Manuel Sevillano，K11 艺术基金会创办人兼名誉主席郑志刚。参观总人数达四千人。展览将持续至 2016 年 2 月 15 日，预计参观人数将达到 35 万人。]$_{N\_List}$ [**另一方面**，两场中国当代艺术展览于同期举行。这些艺术家的作品受到了中国 90 年代历史影响，游走在现实与超现实的边缘。]$_{N\_List}$

English: [Attended the opening ceremony also were curator Montse Aguer, Dalí Foundation chief executive Juan Manuel Sevillano, K11 Arts Foundation founder and honorary chairman Adrian Cheng. Visiting total people number is four thousand. The exhibition will continue until February 15th of 2016, estimated visitor number will reach 350,000.] [**On another hand**, two Chinese contemporary art exhibitions are held in the same period. The artists' works are influenced by China in the 1990s and walk on the edge of reality and surrealism.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	*Lingyifengmian* (另一方面)		Chinese	List
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* there is an ELABORATION relation in the Spanish text, *then* the DM "*lingyifengmian*" (另一方面) ('on another hand') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into a LIST relation, *meanwhile* the relation type and the order of EDUs need to be changed from N-S to N-N.					

Recommendation.A7

Text Name: TERM31

Spanish: [Aunque aún no contamos con resultados, intuimos que el modelo será más amplio que el del sintagma nominal.]$_{N\_List}$ [**Por otra parte**, en la elección de términos técnicos, el caso de declinación interna puede resultar decisivo.]$_{N\_List}$

English: [Although yet no we count with results, we intuit that the model will be wider than the nominal phrase.] [**On the other hand**, in the choice of technical terms, the case of internal decline can be decisive.]

Chinese: [尽管还没有取得最终结果，我们认为该模型已囊括了语段模型涉及的内容。]$_{N\_List}$ [**另一方面**，在筛选技术术语时，单词内部变格尤为重要。]$_{N\_List}$

English: [Although yet get the final results, we consider this model has already included discourse model's contents.] [**On the other hand**, in the selecting of technical terms, word internal disqualification becomes particularly important.]

Discourse markers (DM)	Spanish	por otra parte	**Relation**	Spanish	List
	Chinese	*lingyifengmian* (另一方面)		Chinese	
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*por otra parte*" ('on the other hand') is at the beginning of the second EDU in Spanish, *then* the DM "*lingyifengmian*" (另一方面) can be included at the beginning of the second EDU in Chinese, *meanwhile* the order of the EDUs is not changed.					

Recommendation.A8[61]

Text Name: BMCS3

Spanish: [Si optas por un aprendizaje lo más parecido posible a la inmersión,]$_{List\_N}$ [**y** necesitas mejorar tu nivel de español rápidamente,]$_{List\_N}$

English: [If you choose through a learning that more similar possible to the immersion,] [**and** you need to improve your level of Spanish quickly,]

Chinese: [若您希望进行全面集中的语言学习]$_{13Disjunction\_N}$ [**或者**您希望短时间内提高您的语言水平，]$_{14Disjunction\_N}$

English: [If you wish to conduct completely of language learning,] [**or** you wish in a short time to improve your language skills,]

Discourse markers (DM)	Spanish	y	**Relation**	Spanish	List
	Chinese	*huozhe* (或者)		Chinese	Disjunction
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include a different DM to change the discourse relation, *and* in the Spanish text there is a LIST relation, *and* the DM "*y*" ('and') is at the beginning of the second EDU in Spanish, *then* the DM "*huozhe*" (或者) ('or') can be included at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a DISJUNCTION, *meanwhile* the order of the EDUs is not changed.					

[61] We have already presented the Chinese DM "*bing*" (并) and its original Spanish translation, therefore, in this example, we do not give the translation rule from the discourse similarity aspect.

Recommendation.A9

Text Name: ICP6

Spanish: [El Instituto Cervantes contribuye a la dotación de la Biblioteca "Miguel de Cervantes", sita en el Consulado General de España en Shanghai]$_{13N\_List}$ [**y** colabora activamente en un programa de actividades culturales y docentes en colaboración con instituciones culturales y educativas de otros puntos del país.]$_{14N\_List}$

English: [The Institute Cervantes contributes to the endowment of the Library "Miguel Cervantes", located in the Consulate General of Spain in Shanghai] [**and** collaborates actively in a program of activities cultural and teaching in collaboration with institutions cultural and educational of other parts of the country.]

Chinese: [此外，塞万提斯学院还资助成立了位于西班牙驻上海总领馆内的"米盖尔•德•塞万提斯"图书馆，]$_{N\_Purpose}$ [**旨在**与上海地区的众多文化和教育机构开展合作，积极组织各种文化和教学活动，扩大塞万提斯学院在北京以外地区的影响。]$_{S\_Purpose}$

English: [In addition, Cervantes Institute also sponsored the creation of the located in Spanish General Consulate in Shanghai of "Miguel de Cervantes" Library,] [**aims to** with Shanghai's various cultural and educational institutions to carry out collaborations, actively organizing different cultural and teaching programs, to expand Cervantes Institute outside Beijing of influence.]

Discourse markers (DM)	Spanish	y	**Relation**	Spanish	List
	Chinese	*zhizai* (旨在)		Chinese	Purpose
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to include a different DM to change the discourse relation, *and* in the Spanish text there is a LIST relation, *and* the DM "*y*" ('and') is at the beginning of the second EDU in Spanish, *then* the DM "*huozhe*" (旨在) ('aims to') can be included at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a PURPOSE, *meanwhile* the relation type and the order of EDUs need to be changed from N-N to N-S.					

Recommendation.A10

Text Name: ICP7

Spanish: [La versatilidad del MOPAC permite enriquecer sus prestaciones con la incorporación de otros servicios y recursos electrónicos de valor añadido para socios y que ofrecen sitios web para móviles, como los servicios de gestión de referencias bibliográficas, de préstamo interbibliotecario y suministro de

documentos, o de consulta de recursos electrónicos en red.]$_{N\_List}$ [**Igualmente**, se puede acceder a la descarga de una selección de aplicaciones específicas orientadas para el aprendizaje del español como lengua extranjera.]$_{N\_List}$

English: [The versatility of MOPAC allows to enrich its benefits with the incorporation of other services and resources electronic of value added for partners and that offer websites for smart phones, as the services of management of references bibliographic, of borrowing interlibrary and supply of documents, or of consultation of resources electronic on the network.] [**Likewise**, you can access to the download of a selection of applications specific aimed at learning of Spanish as a foreign language.]

Chinese: [MOPAC 的多功能性使许多增值服务得以实现，比如图书参考管理服务、馆际互借服务、文献传递服务以及网上电子资源查询服务。]$_{N\_Conjunction}$ [**同时**，它还能为西班牙语学习者提供相关领域的应用程序下载服务。]$_{N\_Conjunction}$

English: [MOPAC's versatility enables many value-added services, such as book reference management services, interlibrary borrowing services, documentary delivery service and online e-resource checking services.] [**At the same time**, it can also for Spanish learners offer related fields' application download services.]

Discourse markers (DM)	Spanish	igualmente	**Relation**	Spanish	List
	Chinese	*tongshi* (同时)		Chinese	Conjunction
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include a different DM to change the discourse relation, *and* in the Spanish text there is a LIST relation, *and* the DM "*igualmente*" ('likewise') is at the beginning of the second EDU in Spanish, *then* the DM "*tongshi*" (同时) ('at the same time') can be included at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a CONJUNCTION, *meanwhile* the order of EDUs is not changed.					

Recommendation.A11

Text Name: FCEC1

Spanish: [También lo es el acercar una imagen real de nuestro país a la sociedad china, país en el que España debe lograr posicionarse de una forma más robusta y nítida, alejada de tópicos. Se trata de una cuestión cuya importancia conocemos todos los que trabajamos en el exterior. Las oportunidades para España pasan porque China asocie a nuestra imagen el capital que representa nuestro idioma, el excelente trabajo de nuestras empresas en muchos ámbitos, la profesionalidad de nuestras PYMES, el liderazgo global de nuestras escuelas de negocio y Universidades, y nuestra creatividad, innovación, y avances tecnológicos, por mencionar algunos aspectos. Esta tarea nos corresponde a todos y es el mejor

substrato para nuestra actividad exterior.]N_List [**Igualmente** importante es conocer en profundidad la que hoy ya es la segunda economía, y quizá la más dinámica del mundo, situada además en el nuevo centro de gravedad que representa el eje Asia- Pacífico. Los ojos del mundo miran a China, y España también debe ser capaz de analizar en profundidad los constantes cambios que atraviesa este país y detectar las oportunidades a tiempo para, en definitiva, estar más presentes y visibles en China.]N_List

English: [It is also important to bring a real image of our country to Chinese society, a country in which Spain must achieve a more robust and clear position, away from cliches. It is a question whose importance we all know that we work abroad. The opportunities for Spain happen because China associates in our image the capital that our language represents, the excellent work of our companies in many areas, the professionalism of our SMEs, the global leadership of our business schools and Universities, and our creativity, innovation , and technological advances, to mention some aspects. This task belongs to all of us and is the best substrate for our external activity.] [**Likewise** important is to know in depth what is now the second economy, and perhaps the most dynamic in the world, also located in the new center of gravity that represents the Asia-Pacific axis. The eyes of the world look to China, and Spain must also be able to analyze in depth the constant changes that this country is going through and detect opportunities in time to, in short, be more present and visible in China.]

Chinese: [向中国传达西班牙真实的形象也是我们的重要责任，西班牙应该寻求在中国取得更加强有力的地位，而不仅仅停留在传统的形象上。这是我们长期在其他国家工作的人共同的看法。西班牙在中国有很多发展机会，因为中国人认为西班牙的形象代表的是语言的优势，各个领域公司良好的业绩，中小企业的职业性，商学院和大学的国际领先地位，国民的创造性， 创新性和高科技成就等。这是我们的任务，是我们对外交往工作的精髓。]N_List [当然，加强对中国的了解**也**同样的重要。目前中国已经成为世界第二大经济强国，或许是世界上经济最具活力的国家。它处在新的世界重心亚太地区，全世界的目光都在目视中国，因此西班牙应该具有深度分析的能力，关注这个国家正在经历的发展和变化并从中找到自己的机会。]N_List

English: [It is also our important responsibility to convey to China the true image of Spain. Spain should seek a stronger position in China, not just the traditional image. This is a common view of people who have long been working in other countries. Spain has many development opportunities in China, because Chinese people believe that the image of Spain represents the advantages of language, the good performance of companies in various fields, the professionalism of SMEs, the international leading position of business schools and universities, the creativity of the people, and the innovation And high-tech achievements. This is our task and the essence of our foreign exchanges.] [Of course, it is **also** equally important to strengthen our understanding of China. At present, China has become the world's second-largest economic power, perhaps the most dynamic economy in the world. It is in the new world center of gravity in the Asia-Pacific region, and the eyes of the whole world are on the sight of China. Therefore,

Spain should have the ability to conduct in-depth analysis, pay attention to the developments and changes that the country is experiencing and find its own opportunities.]

Discourse markers (DM)	Spanish	igualmente	**Relation**	Spanish	List
	Chinese	*ye* (也)		Chinese	
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*igualmente*" ('likewise') is at the beginning of the second EDU in Spanish, *then* the DM "*ye*" (因此) ('also') can be added in the second EDU in Chinese, *meanwhile* the order of EDU is not changed.					

Group of relation: CONJUNCTION
Recommendation.A12

Text Name: ICP5

Spanish: [El Instituto Cervantes, además, dispone de la Biblioteca Antonio Machado, que ofrece una amplia selección de literatura española y latinoamericana, películas, música, revistas y periódicos del mundo hispanohablante. Nuestros estudiantes tienen acceso gratuito a todos los fondos y disponen de un servicio de préstamo de libros, CDs y DVDs.]$_{N\_Result}$ [Asimismo, nuestros estudiantes tienen la posibilidad de participar cada semana en actividades culturales relacionadas con la cultura hispanohablante: proyección de películas, exposiciones, conciertos...]$_{S\_Result}$

English: [The Institute Cervantes, in addition, available of the Library Antonio Machado, that offers a wide selection of literature Spanish and Latin America, films, music, magazines and newspapers of the world Spanish-speaking. Our students have access free to all funds and have the service of borrowing of books, CDs and DVDs.] [Likewise, our students have the possibility of participating each week in activities cultural related with culture Spanish-speaking: projection of films, exhibitions, concerts...]

Chinese: [此外，塞万提斯学院拥有一个名为安东尼奥马查多的图书馆，该馆提供大量的西班牙和拉丁美洲文学作品，电影，音乐，杂志和西班牙语世界的报刊。塞万提斯学院的学生可以免费浏览该图书馆馆藏，可以借阅书籍，CD 和 DVD。]$_{N\_Conjunction}$ [**与此同时**，我们的学生还可以参加每周与西班牙语世界有关的文化活动：观看影片，展览，音乐会.....]$_{N\_Conjunction}$

English: [In addition, Cervantes Institute owns a named Antonio Machado of library, this library offers large amount of Spanish and Latin American literature works, films, music, magazines and Spanish-speaking world's newspapers. Cervantes Institute's students can freely access to the library collection, can borrow

readings, CDs and DVDs.] [**At the same time**, our students also can participate wach week with Spanish-speaking related activities: watch films, exhibitions, concerts...]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Result
	Chinese	*yucitongshi* (与此同时)		Chinese	Conjunction
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is a RESULT relation, *then* the DM "*yucitongshi*" (与此同时) ('therefore') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into a CONJUNCTION, *meanwhile* the relation type and the order of EDUs need to be changed from N-S to N-N.					

Group of relation: CONTRAST
Recommendation.A13

Text Name: TERM50

Spanish: [En primer lugar, dispone del prefijo prestado des-, que dispone de las dos posibilidades que también tiene en las lenguas románicas,]$_{N\_Elaboration}$ [en el derivado desegin, actúa como modificador del núcleo-base egin (antónimo de hacer),]$_{S\_Elaboration}$

English: [In the first place, it has the prefix borrowed des-, which has the two possibilities that it also in the language Romance,] [in the derivative desegin, it acts as a modifier of nucleus-base egin (antonym doing),]

Chinese: [首先，前缀 des-在罗马语系中有两种形式，]$_{N\_Contrast}$ [**而**在巴斯克语中，派生词 desegin，可以看做核心基础 egin（"做"的反义词）的修饰词，]$_{N\_Contrast}$

English: [First of all, the prefix des- has two forms in the Roman system,] [**while** in Basque the derivative desegin can be taken as the core basis egin (antonym doing) the modifier,]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	"*er*" (而)		Chinese	Contrast
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is an ELABORATION relation, *then* the DM "*er*" (而) ('while') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into CONTRAST, *meanwhile* the order of EDUs needs to be changed from N-S to N-N.					

Recommendation.A14

Text Name: TERM50

Spanish: [de manera que en euskera, también en derivación, el núcleo de la estructura se ubica a la derecha,]$_{N\_Contrast}$ [**mientras** que en las lenguas románicas va a la izquierda.]$_{N\_Contrast}$

English: [so that in Basque, also in derivation, the nucleus of the structure is located in the right,] [**while** in the languages Romance it goes to the left.]

Chinese: [因此在巴斯克语中，同样是衍生词，其结构的核心在于靠右的位置，]$_{N\_Contrast}$ [**而**罗马语系中结构重心则在左边。]$_{N\_Contrast}$

English: [Therefore, in Basque, it is also a derivative word, its structure of core lies in the right position,] [**while** the structure center of gravity in the Romanic language structure core is on the left.]

Discourse markers (DM)	Spanish	mientras	**Relation**	Spanish	Contrast
	Chinese	"*er*" (而)		Chinese	
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*mientras*" ('while') is at the beginning of the second EDU in Spanish, *then* the DM "*er*" (而) ('while') can be included at the beginning of the second EDU in Chinese, *meanwhile* the order of the EDUs is not changed.					

*** Type of relation: N-S**
Group of relation: RESULT
Recommendation. A13

Text Name: CCICE3

Spanish: [El jueves, el Tesoro volverá a los mercados con una subasta de bonos y obligaciones en la que intentará colocar entre 3.000 y 4.000 millones.]$_{N\_Elaboration}$ [En total, espera captar entre 7.000 y 9.000 millones en toda la semana.]$_{S\_Elaboration}$

English: [On Thursday, the Treasury will back to the markets with an auction of bonds and obligations in which it will try to place between 3,000 and 4,000 million.] [In total, it expects to capture between 7,000 and 9,000 million in whole week.]

Chinese: [另外，财政部将在本周四再次回到市场拍卖中长期国债，欲拍卖 30 亿至 40 亿欧元。]$_{N\_Result}$ [**因此**，财政部将在本周通过国债拍卖筹集 70 亿至 90 亿欧元资金。]$_{S\_Result}$

English: [In addition, the Ministry will in this Thursday again back to the markets auction of medium-and long-term treasury bonds, to auction 3 billion to 4 billion euros.] [**Therefore**, the Ministry of Finance in this week by treasury bonds auction will raise 7 billion to 9 billion euros.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	*yinci* (因此)		Chinese	Result
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is an ELABORATION relation, *then* the DM "*yinci*" (因此) ('therefore') can be added at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a RESULT, *meanwhile* the order of EDU in the Chinese text is not changed.					

Recommendation. A14

Text Name: TERM32

Spanish: [Las palabras clave se extraen del análisis de estas definiciones]$_{N\_Result}$ [**de manera que** pueda llegarse a una traducción literal de las palabras clave del inglés al chino.]$_{S\_Result}$

English: [The keywords extracted of analysis of these definitions] [**so that** could achieve to a translation literal of the keywords of English to Chinese.]

Chinese: [关键词已从上述定义分析中提取出，]$_{N\_Result}$ [**因此**可以将英语的字面意思翻译成中文。]$_{S\_Result}$

English: [Keywords have been from the above mentioned definitions,] [**so that** can make the English literal meaning translates to Chinese.]

Discourse markers (DM)	Spanish	de manera que	**Relation**	Result	
	Chinese	*yinci* (因此)			
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*de manera que*" ('therefore') is at the beginning of the second EDU in Spanish, *then* the DM "*yinci*" (因此) ('therefore') can be added at the beginning of the second EDU in Chinese, meanwhile the order of EDUs is not changed.					

Recommendation.A15[62]

Text Name: TERM31

Spanish: [En ese equilibrio se le otorga preferencia a la cobertura,]$_{N\_Condition}$ [siempre que haya una persona que lleve a cabo la reducción terminológica.]$_{S\_Condition}$

English: [In this balance it gives preference to coverage,] [as long as there is a person that carries out the reduction terminological.]

Chinese: [在该平衡中，覆盖度为优先考虑的部分，]$_{N\_Result}$ [**因此**必须一直有人负责精简术语词汇。]$_{S\_Result}$

English: [In this balance, coverage is the priority part,] [**therefore**, there must be someone to charge of the reduction of terminologies.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Condition
	Chinese	*yinci* (因此)		Chinese	Result
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is a CONDITION relation, *then* the DM "*yinci*" (因此) ('therefore') can be added at the beginning of the second EDU in the Chinese text, *and* the relation in the Chinese text can be changed into a RESULT, *meanwhile* the order of EDUs in the Chinese text is not changed.					

[62] For the three cases of the Chinese DM "*yinci*" (因此), two cases are the same, both cases include a CONDITION relation in the Spanish passages and a RESULT relation in the Chinese passages. Therefore, here we just give two examples for the case of the Chinese DM "*yinci*" (因此) for the explanation of discourse differences. Moreover, in Rule 1 we have already presented the original Spanish translation of DM "*yinci*" (因此), therefore, in Rule 2, we do not give the translation rule from the discourse similarity aspect.

Recommendation.A16

Text Name: TERM38

Spanish: [Si bien este aspecto es común al progreso científico y técnico]$_{N\_Result}$ [y, **por lo tanto**, característico de la neología terminológica,]$_{S\_Result}$

English: [Although this aspect is common to progress scientific and technical,] [and **therefore**, characteristic of the neology terminological,]

Chinese: [对于科技进步来说，这种现象的产生并不稀奇，]$_{S\_Concession}$ [**但**需要注意的是，介于术语新词的特点，...]$_{N\_Concession}$

English: [For the progress of science and technology, this phenomenon is not strange,] [**but** it should be noted that, due to the characteristics of new terminology, ...]

Discourse markers (DM)	Spanish	por lo tanto	**Relation**	Spanish	Result
	Chinese	*dan* (但)		Chinese	Concession
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to include a different DM to change the discourse relation, *and* in the Spanish text there is a RESULT relation, *and* the DM "*por lo tanto*" ('therefore') is at the beginning of the second EDU in Spanish, *then* the DM "*dan*" (但) ('but') can be included at the beginning of the second EDU in Chinese, *and* the relation in the Chinese text can be changed into a CONCESSION, *meanwhile* the order of the EDUs needs to be changed from N-S to S-N.					

Recommendation.A17

Text Name: CCICE4

Spanish: [La reducción de costes continuaría aplicándose a rajatabla en BBVA Portugal, que anuncia el despido de 187 trabajadores y el cierre de 26 oficinas más en todo el país.]$_{N\_Result}$ [**Por tanto**, su presenciaen el país vecino se va a mermar de forma considerable, hasta el punto de que únicamente permanecerán abiertas 12 sedes.]$_{S\_Result}$

English: [The reduction of costs would continue applied to be strictly in BBVA Portugal, which announces the dismissal of 187 workers and the closure of 26 offices more in all the country.] [**Therefore**, their presence in country neighboring is going to decrease considerably, till the point that only will remain open 12 headquarters.]

Chinese: [西班牙 BBVA 银行将关闭葡萄牙分行 26 个营业点，并解雇 187 名员工，继续降该国的营运成本。]$_{N\_Result}$ [**因此**，BBVA 将不断缩小葡萄牙的营业点，直到最后剩下 12 个。]$_{S\_Result}$

English: [Spain BBVA Bank will close Portugal branch banks 26 offices, and dismissal 187 employees, continue reducing this country costs .] [**Therefore**, BBVA will continue decrease Portugal offices, until last remain 12 headquarters.]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>por tanto</td><td rowspan="2">Relation</td><td>Spanish</td><td rowspan="2">Result</td></tr>
<tr><td>Chinese</td><td>yinci (因此)</td><td>Chinese</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-S</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>N-S</td></tr>
<tr><td>Chinese</td><td>N-S</td><td>Chinese</td><td>N-S</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, and the DM “por tanto” (‘therefore’) is at the beginning of the second EDU in Spanish, then the DM “yinci” (因此) (‘therefore’) can be included at the beginning of the second EDU in Chinese, meanwhile the order of the EDUs is not changed.</td></tr>
</table>

Group of relation: PURPOSE
Recommendation.A18

Text Name: BMCS2
Spanish: [Los profesores cuentan siempre con el punto de vista de sus alumnos en la toma de decisiones de la clase,]$_{N\_Means}$ [fomentando la autonomia del estudiante mediante el uso de las estrategias de aprendizaje más adecuadas para cada uno.]$_{S\_Means}$
English: [Teachers have always with the point of view of their students in the making decision of the class,] [forcing the autonomy of students by using the strategies learning more appropriate for each one.]
Chinese: [在教学的进程中塞万提斯的老师总会站在学生的立场和视角，]$_{N\_Purpose}$ [**为**每个学生制定最适合他的教学计划，促进学生的自主学习意识。]$_{S\_Purpose}$
English: [In teaching of process Cervantes teachers always stand on the students’ position and perspective,] [**for** each student to develop the most suitable for his teaching plan, to promote students’ of autonomous learning.]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>/</td><td rowspan="2">Relation</td><td>Spanish</td><td>Means</td></tr>
<tr><td>Chinese</td><td>“wei” (为)</td><td>Chinese</td><td>Purpose</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-S</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>N-S</td></tr>
<tr><td>Chinese</td><td>N-S</td><td>Chinese</td><td>N-S</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to add a DM in Chinese, and in the Spanish text there is a MEANS relation, then the DM “wei” (为) (‘for’) can be added at the beginning of the second EDU in the Chinese text, and the relation in the Chinese text can be changed into a PURPOSE, meanwhile the order of EDUs is not changed.</td></tr>
</table>

Recommendation.A19

Text Name: BMCS2

Spanish: [**Para** ilustrarlo]$_{S\_Purpose}$ [recordaremos el ejemplo de Levi: musical voice, musical criticism, musical comedy.]$_{N\_Purpose}$

English: [**In order to** illustrate,] [we will remember the example of Levi: musical voice, musical criticism, musical comedy.]

Chinese: [**为**了更好地说明，]$_{S\_Purpose}$ [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）。]$_{N\_Purpose}$

English: [**For** better illustration,] [we will use Levi's example: musical voice, musical criticism, musical comedy.]

Discourse markers (DM)	Spanish	para	**Relation**	Spanish	Purpose
	Chinese	"*wei*" (为)		Chinese	
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*para*" ('in order to') is at the beginning of the first EDU in Spanish, *then* the DM "*wei*" (为) ('for') can be included at the beginning of the first EDU in Chinese, *meanwhile* the order of EDUs is not changed.					

Recommendation.A20[63]

Text Name: TERM31

Spanish: [**Para** obtener una cobertura del 95%]$_{S\_Purpose}$ [se suele reducir la precisión al 50%,]$_{N\_Purpose}$

English: [**In order to** obtain 95% coverage,] [accuracy is usually reduced to 50%.]

Chinese: [**若**覆盖率需达到 95%，]$_{S\_Condition}$ [通常精确度会降低至 50%，]$_{N\_Condition}$

English: [**If** the coverage reaches to 95%,] [usually the accuracy will be reduced to 50%,]

[63] Since the meaning of Spanish DMs *para* and *para que* are the same, and in the corpus the case of these two DMs are the same, both Chinese translations are "*ruo*" (若), therefore, we only give the example of the DM *para*, the translation rule of *para que* is the same as well. Regarding the discourse similarity, the Chinese translation of the two Spanish DMs is the "*yibian*" (以便). Thus, we also only give the example of the Spanish DM *para.*

Discourse markers (DM)	Spanish	para	**Relation**	Spanish	Purpose
	Chinese	*ruo* (若)		Chinese	Condition
Relation type	Spanish	N-S	**EDUs order**	Spanish	S-N
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to include a different DM to change the discourse relation, *and* in the Spanish text there is a PURPOSE relation, *and* the DM "*para*" ('in order to') is at the beginning of the first EDU in Spanish, *then* the DM "*ruo*" (若) ('if') can be included at the beginning of the first EDU in Chinese, *and* the relation in the Chinese text can be changed into CONDITION, *meanwhile* the order of EDUs is not changed.					

Recommendation.A21

Text Name: ICEG2

Spanish: [Visite la Memoria de los Talleres]$_{N\_Purpose}$ [**para** conocer todas los talleres que hemos realizado en el Instituto Confucio de la Universidad de Granada.]$_{S\_Purpose}$

English: [Visit the Workshop Report] [**in order to** know all the workshops that we have realized in the Institute Confucius of the University of Granada.]

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，]$_{N\_Purpose}$ [**以便**您更好地参与其中。]$_{S\_Condition}$

English: [Welcome to visit Granada University Confucius Institute organized workshops records,] [**so that** you can better participate.]

Discourse markers (DM)	Spanish	para	**Relation**	Spanish	Purpose
	Chinese	*yibian* (以便)		Chinese	
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*para*" ('in order to') is at the beginning of the second EDU in Spanish, *then* the DM "*yibian*" (以便) ('so that') can be included at the beginning of the second EDU in Chinese, *meanwhile* the order of EDUs is not changed.					

Group of relation: CONDITION

Recommendation.A22

Text Name: TERM31

Spanish: [Además, en las áreas en las que la terminología evoluciona de modo dinámico, como por ejemplo la informática,]$_{7N\_Evaluation}$ [sin ese tipo de

instrumental resulta prácticamente imposible llevar a cabo un trabajo terminológico efectivo.]$_{8S\_Evaluation}$

English: [In addition, in the fields in that the terminology evolves module dynamically, such as computer science,] [without this type of instrumentation it is practically impossible to carry out a work terminological effective.]

Chinese: [此外，在诸如计算机等术语呈现动态发展的领域，**若**未使用上述工具，]$_{7S\_Condition}$ [在实际操作中则不可能进行有效的术语整理工作。]$_{8N\_Condition}$

English: [In addition, in such as computer science presents dynamic development area, **if** without using above mentioned tool,] [in practical operation is impossible to carry out effective terminology organization work.

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Evaluation
	Chinese	*ruo* (若)		Chinese	Condition
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is an EVALUATION relation, *then* the DM "*ruo*" (若) ('if') can be added in the first EDU in the Chinese text, *and* the relation in the Chinese text can be changed into a CONDITION, *meanwhile* the order of EDUs needs to be changed from N-S to S-N.					

Recommendation.A23

Text Name: BMCS3

Spanish: [**Si** optas por un aprendizaje lo más parecido posible a la inmersión, y necesitas mejorar tu nivel de español rápidamente,]$_{S\_Condition}$ [los cursos intensivos son una buena opción.]$_{N\_Condition}$

English: [**If** you opt for a learning that more close possible to the immersion, and you need to improve your Spanish level quickly,] [the intensive courses are a good option.]

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，]$_{S\_Condition}$ [紧凑课程是一个很好的选择。]$_{N\_Condition}$

English: [**If** you wish to conduct a comprehensive and focused language study or you wish in short time to improve your language level,] [a compact course is a good choice.]

Discourse markers (DM)	Spanish	si	**Relation**	Spanish	Condition
	Chinese	"*ruo*" (若)		Chinese	
Relation type	Spanish	N-S	**EDUs order**	Spanish	S-N
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*si*" ('if') is at the beginning of the first EDU in Spanish, *then* the DM "*ruo*" (若) ('therefore') can be added at the beginning of the first EDU in Chinese.					

Group of relation: CIRCUMSTANCE
Recommendation.A24

Text Name: TERM50
Spanish: [en el segundo caso, mientras el núcleo está compuesto por el mismo prefijo,]$_{N\_Contrast}$ [el complemento es la base de derivación;]$_{N\_Contrast}$
English: [in the second case, while the nucleus is composed of the same prefix,] [the complement is the base of derivation;]
Chinese: [在第二种情况下，当核心部分由同一个前缀构成，]$_{S\_Circumstance}$ [补充成分即为派生的基础，]$_{N\_Circumstance}$
English: [In the second case, **when** the core part is composed of the same prefix,] [the supplementary component is the basis of the derivative,]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Contrast
	Chinese	"*dang*" (当)		Chinese	Circumstance
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to add a DM in Chinese, *and* in the Spanish text there is a CONTRAST relation, *then* the DM "*dang*" (当) ('when') can be added in the first EDU in the Chinese text, *and* the relation in the Chinese text can be changed into a CIRCUMSTANCE, *meanwhile* the order of EDUs needs to be changed from N-N to S-N.					

Recommendation.A25

Text Name: TERM50
Spanish: [pero **cuando** queremos buscar un ejemplo del modelo complemento del prefijo/núcleo de la base (deshojar) desostatu,]$_{S\_Circustance}$ [nos encontramos que no está bien formado.]$_{N\_Circumstance}$

English: [but **when** we want to look for an example of the model complement of the prefix / kernel of the base (deshojar) desostatu,] [we find that it is not well formed.]

Chinese: [但是**当**我们想找一个可以补充说明前缀/基础核心关系（诸如(deshojar（使落叶）) desostatu）的补充实例时，]$_{S\_Circumstance}$ [发现这个单词的构成并不符合要求。]$_{N\_Circumstance}$

English: [But **when** we want to find a could supplement the prefix/basic core relation ship (such as (deshojar) desostau) a supplemental example,] [it was found that the word composition did not meet the requirements.]

Discourse markers (DM)	Spanish	cuando	**Relation**	Spanish	Circumstance
	Chinese	"*dang*" (当)		Chinese	
Relation type	Spanish	N-S	**EDUs order**	Spanish	S-N
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to include the same DM to maintain the same discourse relation both in Spanish and Chinese, *and* the DM "*cuando*" ('when') is at the beginning of the first EDU in Spanish, *then* the DM "*dang*" (当) ('when') can be included at the beginning of the first EDU in Chinese, *meanwhile* the order of EDUs is not changed.					

B. Unit shift

Recommendation.B1

Text Name: BMCS2

Spanish: [Metodología actual.]$_{S\_Interpretation}$ [El material de enseñanza procede de España, ...]$_{N\_Interpretation}$

English: [Methodology current.] [The material of teaching comes from Spain, ...]

Chinese: [领先的教学方法]$_{S\_Preparation}$ [我们的教材为西班牙原版教材...]$_{N\_Preparation}$

English: [Leading teaching method] [Our material is the Spanish original textbook...]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>/</td><td rowspan="2">Relation</td><td>Spanish</td><td>Interpretation</td></tr>
<tr><td>Chinese</td><td>/</td><td>Chinese</td><td>Preparation</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-S</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>S-N</td></tr>
<tr><td>Chinese</td><td>N-S</td><td>Chinese</td><td>S-N</td></tr>
<tr><td colspan="3" rowspan="2">Punctuation marks</td><td colspan="2">Spanish</td><td>.</td></tr>
<tr><td colspan="2">Chinese</td><td>/</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to do a unit shift, and there is no DM neither in Spanish nor in Chinese, and in the Spanish text there is an INTERPRETATION relation, and in the Chinese text there is an PREPARATION relation, and the two EDUs in Spanish are divided by a comma, then both EDUs in Chinese can be divided without any punctuation, meanwhile the order of EDUs is not changed.</td></tr>
</table>

Recommendation.B2

Text Name: EEP7

Spanish: [La muestra de este año ha sido un reflejo de los desafíos a los que se enfrenta el cine español en la actualidad.]$_{N\_Evidence}$ [Las tendencias globalizadas exigen a los renovados autores y talentos que incorporen las nuevas tecnologías y que desarrollen una innovadora experimentación genérica.]$_{S\_Evidence}$

English: [The show of this year has been a reflection of the challenges that facing the film Spanish today.] [The tendency globalized requires the renewed authors and talents to incorporate the new technologies and to develop a innovative experimentation generic.]

Chinese: [此次挑选的这一系列影片流派各异，]$_{N\_Evaluation}$ [体现了西班牙电影界国内市场的繁荣和与时俱进的气象，反映了西班牙电影向国外市场扩张的趋势及其国际威望。]$_{S\_Evaluation}$

English: [The selection of this series of films varies in genres,] [showing the Spanish film industry's prosperity of the domestic market and the trend of advancing with the times, reflecting the Spanish films to foreign market's trend of expanding and their international prestige.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Evidence
	Chinese	/		Chinese	Evaluation
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Punctuation marks			Spanish		.
			Chinese		，
Recommendation for translation					
If the translator wants to do a unit shift, *and* there is no DM neither in Spanish nor in Chinese, *and* in the Spanish text there is an EVIDENCE relation, *and* in the Chinese text there is an EVALUATION relation, *and* the two EDUs in Spanish are divided by a comma, *then* both EDUs in Chinese can be divided by a period, *meanwhile* the order of the EDUs is not changed.					

C. Other cases

Group of cases: Unit shift + marker change

Recommendation.C1

Text Name: ICP3

Spanish: [Los actos son en su mayoría gratuitos]$_{S\_Concession}$ [**pero** para las actividades que se realizan en nuestro auditorio se recomienda acudir unos minutos antes del comienzo, ya que el aforo de la sala es limitado a 90 personas.]$_{N\_Concession}$

English: [The acts are mostly free] [**but** for the activities that take place in our auditorium it is recommended to go a few minutes before the start, as the capacity of the room is limited to 90 people.]

Chinese: [我们的绝大部分文化活动面向公众免费开放。]$_{N\_Elaboration}$ [由于场地有限（多功能厅限 90 人），建议大家在每次活动开始前，提前几分钟入场。]$_{S\_Elaboration}$

English: [Our most cultural activities are open to public for free.] [Due to the space limited (multi-function hall limited to 90 people), we recommend that in each activity start before, you present yourself a few minutes early.]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>pero</td><td rowspan="2">Relation</td><td>Spanish</td><td>Concession</td></tr>
<tr><td>Chinese</td><td>/</td><td>Chinese</td><td>Elaboration</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-S</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>S-N</td></tr>
<tr><td>Chinese</td><td>N-S</td><td>Chinese</td><td>N-S</td></tr>
<tr><td colspan="3" rowspan="2">Punctuation marks</td><td colspan="2">Spanish</td><td>/</td></tr>
<tr><td colspan="2">Chinese</td><td>。</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to do a unit shift and a marker change, and in Spanish the DM “pero” (‘but’) is at the beginning of the second EDU, there is a CONCESSION relation, and both EDUs are divided without any punctuation, then in Chinese, the DM can be erased, a period between both EDUs can be added, and the relation can be changed into an ELABORATION relation, meanwhile the order of EDUs needs to be changed from S-N to N-S.</td></tr>
</table>

Recommendation.C2

Text Name: TERM31

Spanish: [En las lenguas de flexión compleja, el tratar solamente el aspecto formal de las palabras acarreará malos resultados]$_{N\_List}$ [**y** será necesaria la lematización.]$_{N\_List}$

English: [In the languages of bending complex, treating only the aspect formal of the words will lead to poor results] [**and** will be necessary the lemmatization.]

Chinese: [对于词尾有复杂变化的语言来说，仅看单词表面就进行分析只会造成很糟糕的局面。]$_{N\_Circumstance}$ [**此时**词根分析就变得更为不可或缺。]$_{S\_Circumstance}$

English: [For words that have complex variations of a language, only check the word at the surface to carry out the analysis can bring a bad situation.] [**At this time**, the stemming analysis becomes more essential.]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>y</td><td rowspan="2">Relation</td><td>Spanish</td><td>List</td></tr>
<tr><td>Chinese</td><td>cishi (此时)</td><td>Chinese</td><td>Circumstance</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-N</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>N-N</td></tr>
<tr><td>Chinese</td><td>N-S</td><td>Chinese</td><td>N-S</td></tr>
<tr><td colspan="3" rowspan="2">Punctuation marks</td><td colspan="2">Spanish</td><td>/</td></tr>
<tr><td colspan="2">Chinese</td><td>。</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to do a unit shift and a marker change, and in Spanish the DM “y” (‘and’) is at the beginning of the second EDU, there is a LIST relation, and both EDUs are divided without any punctuation, then in Chinese, the DM can be changed into “cishi” (此时) (‘at this time’), a period between both EDUs can be added, and the relation can be changed into a CIRCUMSTANCE relation, meanwhile the type of relation and the order of EDUs needs to be changed from N-N to N-S.</td></tr>
</table>

Recommendation.C3

Text Name: EEP3

Spanish: [Se acordó establecer una hoja de ruta]$_{N\_Purpose}$ [**para** identificar áreas estratégicas de interés mutuo y avanzar en proyectos concretos hasta la próxima Comisión mixta que se celebrará en 2017 en España.]$_{S\_Purpose}$

English: [It was agreed to establish a roadmap] [**to identify** areas strategic of interest mutual and to advance in projects concrete until the next Commission Mixed to be held in 2017 in Spain.]

Chinese: [双方同时还肯定了战术领域方面的共同战略领域，]$_{N\_List}$ [推进具体项目，到 2017 年在西班牙举行的下一届科技联委会。]$_{N\_List}$

English: [Both sides also affirmed their tactic area of the common strategic areas,] [advancing concrete projects, until in 2017 in Spain to hold the next Mixed Commission.]

<table>
<tr><td rowspan="2">Discourse markers (DM)</td><td>Spanish</td><td>para</td><td rowspan="2">Relation</td><td>Spanish</td><td>Purpose</td></tr>
<tr><td>Chinese</td><td>/</td><td>Chinese</td><td>List</td></tr>
<tr><td rowspan="2">Relation type</td><td>Spanish</td><td>N-S</td><td rowspan="2">EDUs order</td><td>Spanish</td><td>N-S</td></tr>
<tr><td>Chinese</td><td>N-N</td><td>Chinese</td><td>N-N</td></tr>
<tr><td colspan="3" rowspan="2">Punctuation marks</td><td colspan="2">Spanish</td><td>/</td></tr>
<tr><td colspan="2">Chinese</td><td>，</td></tr>
<tr><td colspan="6">Recommendation for translation</td></tr>
<tr><td colspan="6">If the translator wants to do a unit shift and a marker change, and in Spanish the DM “para” (‘for’) is at the beginning of the second EDU, there is a PURPOSE relation, and both EDUs are divided without any punctuation, then in Chinese, the DM can be erased, a comma between both EDUs can be added, and the relation can be changed into a LIST relation, meanwhile the relation type and the order of EDUs need to be changed from N-S to N-N.</td></tr>
</table>

Group of cases: Different oder of EDUs

Recommendation.C4

Text Name: CCICE1

Spanish: [En 2015, por la primera vez, la región de Norteamérica se convierte en el tercer feudo por primas de Mapfre,]$_{N\_Cause}$ [desplazando en esa posición a Latam Sur.]$_{S\_Cause}$

English: [In 2015, for the first time, the region of North America becomes the third premium fief of Mapfre,] [displacing in this position to Latam Sur.]

Chinese: [在保险方面，北美已超越南美，]$_{S\_Cause}$ [上升为西班牙保险公司 Mapfre 第三大市场。]$_{N\_Cause}$

English: [In insurance, North America has surpassed South America,] [rose to the Spanish insurance company Mapfre third largest market.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Cause
	Chinese	/		Chinese	Cause
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to change the EDUs order, *and* there is no DM neither in Spanish nor in Chinese, *and* there is a CAUSE relation in both Spanish and Chinese, *then* the order of the EDUs needs to be changed from N-S to S-N.					

Recommendation.C5[64]

Text Name: ICP3

Spanish: [pero para las actividades que se realizan en nuestro auditorio se recomienda acudir unos minutos antes del comienzo,]$_{N\_Cause}$ [**ya que** el aforo de la sala es limitado a 90 personas.]$_{S\_Cause}$

English: [but for the activities that take place in our auditorium it is recommended to go a few minutes before the start,] [becasue the capacity of the room is limited to 90 people.]

Chinese: [**由于**场地有限（多功能厅限 90 人），]$_{S\_Cause}$ [建议大家在每次活动开始前，提前几分钟入场。]$_{N\_Cause}$

English: [**Due to** the space limited (multi-function hall limited to 90 people),] [we recommend that in each activity start before, you present yourself a few minutes early.]

Discourse markers (DM)	Spanish	ya que	**Relation**	Spanish	Cause
	Chinese	*youyu* (由于)		Chinese	Cause
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	S-N
Recommendation for translation					
If the translator wants to change the EDUs order, *and* there is a DM "*ya que*" ('because') in Spanish and a DM "*youyu*" (由于) ('dueto') in Chinese, *and* there is a CAUSE relation in both Spanish and Chinese, *then* the order of the EDUs needs to be changed from N-S to S-N.					

[64] There is another same case of *ya que* and "*youyu*" (由于), therefore, we do not give another case here.

Chapter 7
A Spanish-Chinese Language Learning Task by Using Technological Corpus-based Resources

In this chapter, we will explain the usage of the corpus for Spanish-Chinese language learning. Firstly, we will indicate the learning level for using the corpus (Section 7.1). Secondly, we will describe the process to elaborate the tasks for Spanish-Chinese language learning (Section 7.2). Concretely, we will list the annotated DMs and discuss about the considerations during the elaboration process. Lastly, we will review the chapter information (Section 7.3).

7.1 Level Requirement for the Spanish-Chinese Language Learning with the Corpus

Following the information that we have showed in the Chapter 1, due to the huge population that speak both languages, the Spanish and Chinese language pair occupies an important position for linguistic/NLP studies. In the current globalized context, communication between the two languages is crucial for individuals and language schools, institutions and enterprises, among other organizations. This can be witnessed by the many Spanish speakers are learning Chinese and many Chinese speakers are learning Spanish (Cao, da Cunha and Iruskieta, 2017).

Regarding the language aspect, from the examples we have presented in the first chapter, we can see that the extensive grammatical rules, syntactic structure and discourse elements between them are different. Thus, it is not easy to carry out any linguistic tasks (such as, language teaching tasks, translation tasks, text edition tasks, etc.) between the two languages. Moreover, Spanish is a language that has gender, so the distinction of grammatical gender is crucial between masculine and feminine (among irregular constructions). There are some adjectives with a particular ending for feminine ('JJ+a'[65], such as *pública* 'feminine_public', *extranjera* 'feminine_foreigner', *china* 'feminine_chinese') and for masculine ('JJ+o', such as *moderno* 'masculine_modern', *chino* 'masculine_chinese', *rico* 'masculine_rich'). In Chinese, for example, the masculine *chino* and feminine *china* are translated as "*zhongguo*" (中国) ('China'). For Chinese students who take Spanish as their second language, they must be aware of this language phenomenon.

Based on the above mentioned reasons, we decided to make our study useful for Spanish-Chinese language learning with the corpus. In this study, for the Chinese users who learn Spanish, we adopt the language level standardizations of *Instituto Cervantes*, the official Spanish organization to check the language level for L2

[65] In the Stanford parser JJ means adjective.

Spanish learning[66]; for the Spanish users who learn Chinese, we adopt the language level standardizations of *Hanban* (汉办), the official organization of the Chinese government for L2 Chinese learning[67].

As we have mentioned, the corpus includes specialized texts from different sources, which include terminology from several domains. Therefore, the users of our annotated-corpus and search tool should have an intermediate or advanced level of the language. As the webpage of *Instituto Cervantes* indicates, in the Spanish initial level only some basic expressions and vocabulary are learned. The webpage of *Hanban* (汉办) offers similar information about the Chinese initial level.

In order to use our annotated-corpus and search tool, the appropriate levels for Spanish foreign language learners are level B2 (intermediate level) and level C (including C1 and C2) (advanced level). Level B2 requires language users to understand complex texts with different topics. Level C1 requires understanding a wide variety of long and demanding texts, and also writing and expressing well-structured texts in Spanish. Level C2 is a more advanced level and requires Spanish learners have enough linguistic competence to prove a spontaneous capacity of adaptation to any context, with a great deal of semantic and grammatical precision.

The appropriate level for Chinese foreign language learners is level 4 (intermediate level) and advanced level (level 5 and level 6). Level 4 requires language learners to know a certain amount of words and produce texts related to a wide range of topics, in order to maintain a fluent communication with native speakers. Level 5 requires learners to read magazines, newspapers, and films and give a full-length speech. Level 6 language learners should easily comprehend written and spoken information in Chinese.

7.2 A Task for Spanish-Chinese Language Learning of DMs

In this section, we will explain the process to elaborate the task for Spanish-Chinese language learning. Moreover, we will analyze the elaboration result and discuss the limitations of the designed tasks.

7.2.1 Annotation of DMs

In this study, we have categorized different types of DMs as following show[68]:

---

[66] A detailed explanation of the Spanish level for L2 learners can be consulted: http://dele.cervantes.es/en/information/levels/spanish_levels.html [Last consulted: 17 of September of 2016]

[67] The detailed explanation of the Chinese level for L2 learners can be consulted: http://english.hanban.org/node_8002.htm [Last consulted: 17 of September of 2016]

[68] For the annotation of DMs in the corpus, we use the unharmonized corpus because in the harmonized corpus, some relations are erased, including some relations that contain the DMs, as we indicated in the Chapter 5. Moreover, based on the language learning purpose, in some annotation cases, we annotate some DMs of the sentences that do not follow the segmentation criteria.

- N-S type:

(Ex.1) Antithesis
Nuclear: The author favors the idea.
Satellite: The author disfavors the idea.
Spanish DM(s): aunque; por el contrario; sino
Chinese DM(s): 但是
(Ex.2) Cause
Nuclear: A situation.
Satellite: Another situation that causes that one.
Spanish DM(s): como; debido a; ya que
Chinese DM(s): 因为；由于
(Ex.3) Circumstance
Nuclear: Text shows the ideas or the events that occur in the interpretive text.
Satellite: An interpretive context of situation or time.
Spanish DM(s): cuando
Chinese DM(s): 作为；如同
(Ex.4) Concession
Nuclear: A situation confirmed by the author.
Satellite: Another situation inconsistent but also affirmed.
Spanish DM(s): pero; sino que; si bien
Chinese DM(s): 尽管...但是；然而；无论如何；尽管
(Ex.5) Condition
Nuclear: Action or situation whose occurrence results from the occurrence of the conditioning situation.
Satellite: A condition situation.
Spanish DM(s): si
English: 若；如果
Chinese DM(s):
(Ex.6) Elaboration
Nuclear: The basic information.
Satellite: Additional information of the basic information.
Spanish DM(s): además; además de;
Chinese DM(s): 此外；另外
(Ex.7) Evidence
Nuclear: A claim.
Satellite: Information that increases the reader's belief in the claim.
Spanish DM(s): de acuerdo a; de acuerdo con; de ahí; tal y como
Chinese DM(s): 比如
(Ex.8) Interpretation[69]
Nuclear: A situation.
Satellite: An interpretation of the situation.
Spanish DM(s): en concreto

[69] Due to the translation strategy, not all the relations include both Spanish DMs and Chinese DMs.

Chinese DM(s): /
(Ex.9) Purpose
Nuclear: An intended situation.
Satellite: The intent behind the situation.
Spanish DM(s): a fin de; con afán de; con la movilidad; con el objetivo de; con este fin; con tal fin; de manera que; para; para ello
Chinese DM(s): 以便；旨在；为了
(Ex.10) Restatement
Nuclear: A situation.
Satellite: A reexpression of the situation.
Spanish DM(s): es decir
Chinese DM(s): 即
(Ex.11) Result
Nuclear: A situation.
Satellite: Another situation which is caused by that one.
Spanish DM(s): en consecuencia; de manera que; por consiguiente
Chinese DM(s): 于是；因此
(Ex.12) Summary
Nuclear: A text.
Satellite: Summary of the text.
Spanish DM(s): en resumen
Chinese DM(s): 总之；总而言之；

- N-N type:

(Ex.13) Conjunction
Nuclear: A situation or an action.
Satellite: Another situation or another action that happens at the same time.
Spanish DM(s): al mismo tiempo
Chinese DM(s): 同时；与此同时
(Ex.14) Contrast
Nuclear: One alternate.
Satellite: The other alternate.
Spanish DM(s): por el contrario
Chinese DM(s): 而；相反
(Ex.15) Disjunction
Nuclear: An alternative.
Satellite: Another alternative.
Spanish DM(s): o
Chinese DM(s): 或；或是；或者；亦或
(Ex.16) List
Nuclear: An item.
Satellite: A next item.
Spanish DM(s): e; ni; y; no solo; por un lado; por otro lado; sino también; tanto como

Chinese DM(s): 并; 并且; 和; 一方面..另一方面; 及; 以及; 还; 不仅(仅); 也; 既不…也不; 同样也

(Ex.17) Sequence

Nuclear: An item.

Satellite: A next item.

Spanish DM(s):a continuación; antes de; en primer lugar; en tercer lugar; por último; seguidamente; tras

Chinese DM(s): 首先; 接下来; 紧接着

7.2.2 Exercise Elaboration

- Exercises for L2 Spanish learner

For the Spanish language exercise, we use Python script to generate the texts automatically by removing the annotated DMs. Moreover, for each text, there are multiple-choice answers for the users.

First of all, we have manually annotated all the Spanish DMs in the research corpus, as indicated in Section 7.2.1. Secondly, we write a program to generate the Spanish texts one by one automatically. The following steps have been carried out to create the program: (a) We make a file containing the names of the documents we want to use to generate the exercises; (b) We elaborate a list of the DMs we want to remove. In the case of those composed by two parts, only the first part has to be in the file, such as *tanto…como* (as much as), only *tanto* (so much) is in the file; (c) We create a dictionary with second parts of the composed DMs, which are related to the their first part; (d) We make another dictionary that contains the information of each DM, in this program, we use "always" to represent the DMs that have to be removed always, the meaning of "BOS" is to remove the DMs that only appear at the beginning of the sentences and, the significance of "2p" is to show a composed DM; and (e) We use one more dictionary to collect 7 groups of DMs, depending on their discourse meaning. Within each group, the DMs are grouped if it is difficult to distinguish between them. In this case, one cannot be used as a distractor of the other. Thirdly, we create the final program to grade the answers.

- Exercises for L2 Chinese learner

Same as the Spanish part, we make a small program to take our all the discourse markers. However, the exercise design is different from Spanish language exercise.

Firstly, we have annotated all the DMs in each Chinese text of the entire corpus. Secondly, we aimed to erase all the DMs of a Chinese text, and the following aspects have been considered during the design process: (a) We make a file that contains the names of the documents we want to use to generate the exercise; (b) We put all the annotated DMs into a list; and (c) We collect all the texts in a folder named "Texts", meanwhile, we create another folder called "Exercise", which contains all the outputs of the generated texts.

The reason to make two different designs for Spanish and Chinese texts is because, although the texts are parallel, compared to the Spanish texts, the Chinese texts are more difficult to understand because of the different meanings but the same word

(including some annotated DMs[70]), therefore, we consider that, for a Chinese text, it is better to remove the DMs and mix the correct answers to let the users to choose so that the they can understand the text better by filling the DMs.

7.2.3 Analysis and Discussion

As explained in the previous section (Section 4.6.3), to evaluate the performance of the automatic generation program, we use Kappa to evaluate the correctness of our programming. Kappa gives the agreement of annotation as follows:

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

where P(A) represents the actual observed agreement, and P(E) represents chance agreement.

Since different DMs contain different numbers of words, we use DMs of 60 texts (30 Spanish texts and their parallel Chinese texts) to develop our program and 40 texts (20 Spanish texts and their parallel Chinese texts) to test the accuracy of the programming. Table 32 shows the K results of the 20 tested Spanish texts and their parallel Chinese texts[71].

Text Name	Spanish	Chinese
TERM18	0.963	0.885
TERM19	0.878	0.857
TERM23	0.975	0.877
TERM25	0.866	0.657
TERM30	0.950	0.746
TERM31	0.971	0.891
TERM34	0.914	0.861
BMCS3	1	0.795
BMCS5	1	0.962
CCICE1	0.797	0.538
CCICE4	0.931	0.921
EEP3	1	0.873
EEP4	0.912	0.955
FICB3	1	0.907
FICB4	0.886	0.662
FCEC2	0.905	0.866
ICP2	0.927	0.806
ICP6	0.963	0.897
ICP7	1	0.822
ICEG1	0.973	0.907

Table 32. Program accuracy of the 40 tested Spanish-Chinese parallel texts

[70] For example, the Chinese DM "*wei*" (为) means 'aim to' in English, but it also means 'as', 'to help' in Chinese.

[71] The corpus has different sources, therefore, the number of selected texts for test based on the percentage of each part in the research corpus, and the appearances of annotated DMs in each text.

From Table 32, we can see that our program works quite well for all the Spanish texts, among the 20 tested Spanish texts, 5 of them have 100% accuracy. Other texts maintain the accuracy from 0.86 to 1 except the text CCICE1 (0.797). After analyzing the outputs, we find that the common limitation for the Spanish texts is that, not all the annotated Spanish DMs have been erased. Here we give the text CCICE1 as an example. In this short text, two Spanish DMs have been removed (*y* [‘and’] ; *así* [‘thus’]), while one DM (*por el contrario* [‘in contrast’]) is not.

For the Chinese texts, we can see that, the lowest results of Chinese texts fall on TERM25(0.657), TERM30(0.746), BMCS3(0.795), CCICE1(0.538) and FICB4 (0.662). We give a qualitative analysis for these low resulted texts and we find some common limitations for these texts; here we give CCICE1 as an example. In this text, the Chinese character “*wei*” (为) appear 7 times, however, none of them can be considered as the DM (whose meaning is ‘aim for’). Based on the short text content, the character “*wei*” (为) means ‘as’. Another limitation related with this character appears in the text FICB4. The Chinese phrase “*zuowei*” (做为) (‘as’ in English) contains the annotated DM “*wei*” (为), however, in this case, together with the character “*zuo*” (做) (‘make / to’ in English), it cannot consider “*wei*” (为) as a DM.

The sequence of the phrases also brings us some limitations during the test process. Among the annotated Chinese DMs, one of them is “*zhizai*” (旨在), whose meaning is ‘to do something’ or ‘aims to’ in English. In the text TERM25, the phrase “*zhuzhi*” (主旨) (‘main purpose’) that ends with “*zhi*” (旨) is next to the phrase “*zaiyu*” (在于) (‘lie in’) who starts with “*zai*” (在)[72]. Since there is no space between Chinese characters in a text, our program considers “*zhi*” (旨) and “*zai*” (在) as a DM.

Here are the conclusions to explain the limitations in the Chinese texts:

- A DM could have other meaning depending on the text content, but our program cannot understand the text content.
- Some DMs are composed of two Chinese characters, however, our program just annotates one character, for instance, “*yiji*” (以及), the output of this DM removes the character “*ji*” (及)[73]. Other similar cases exist.
- The phrase sequences can cause the characters combination to be interpreted as an annotated DM, such as the example of “*zhuzhi zaiyu*” (主旨//在于).
- Some Chinese DMs are single characters, but they can convert to a new phrase together with another different character, under this case, we cannot consider this character as a DM anymore.

To conclude, we present the first Spanish-Chinese parallel corpus that can be used for multiple choice language learning tasks. We were able to achieve a perfect result for the Spanish subcorpus. Moreover, we can also grade the users’ answers with our program. While we found some limitations with the Chinese subcorpus, it can still generate good L2 Chinese language exercises.

---

[72] The original content in TERM25 is “*zhuzhi zaiyu*” (主旨//在于), and as we have indicated, “*zhizai*” (旨在) is a DM.

[73] “*yiji*” (以及) and “*ji*” (及) are two annotated Chinese DMs in our work, both of them represent a LIST discourse relation. The meaning of the two DMs are same, is ‘and’ in English.

For future work, we will annotate more cases for the Chinese subcorpus to get a better result, and also make our program able to grade the Chinese language exercises as well.

7.3 Chapter Overview

At the beginning of this chapter, we discussed the potential uses for Spanish-Chinese language learning with our corpus. Also we indicated the language learning level for the users who want to use the corpus.

The main part of this chapter is the introduction of language learning with the corpus. We create the first RST based Spanish-Chinese question generation with multi choice system for language learning. Based on the annotation results, with the annotated DMs, we develop the first multiple choice exercise system for L2 learning between the two languages under RST. For the Spanish part, the performance accuracy is from 0.797 to 1, for the Chinese part, the performance accuracy is from 0.538 to 0.921. We also analyze the limitations of the system, such as, the multi-meaning of the word that can be considered as a DM; the phrase sequence and the recognition of the Chinese characters are also aspects that affect the system performance accuracy, etc.

Chapter 8
Conclusions and Future Work

The previous chapters have introduced the our study step by step. In this chapter, we will provide the conclusion for this dissertation. In particular, we will summarize the study, and look ahead at future work.

This chapter includes the following sections, in Section 8.1, we will give a conclusion of this study in general. In Section 8.2, we will talk about the contributions of the thesis. In Section 8.3, we will analyze the limitations of this work. In Section 8.4, we will outline the future work.

8.1 Conclusions

As we introduced in Chapter 1, since Spanish and Chinese are two of the most spoken languages in the world, for the global communication and NLP research, it is important to carry out the linguistic study between this language pair. Moreover, with the growing interesting on discourse analysis for NLP studies, the main focus of this study is a comparative discourse analysis between Spanish and Chinese with a parallel corpus annotated with discourse information,which can help the language translation and learning between Spanish and Chinese.

Our work is the first one to analyze this language pair by building and annotating parallel texts under RST for Spanish and Chinese. The result is a discourse-annotated corpus, called the RST Spanish-Chinese Treebank. The corpus can be consulted online: `http://ixa2.si.ehu.es/rst/zh/`. The information of the corpus is also accessible.

The study has been divided into different steps: (i) corpus compilation, (ii) discourse segmentation annotation and evaluation, (iii) Central Unit (CU) annotation and evaluation, (iv) discourse structure annotation and evaluation, (v) construction of the translation protocol, and (vi) creation the language learning tasks. For each annotation step, we have invited experts with experiences in discourse analysis to participate.

Regarding the objectives and hypothesis that we proposed in Chapter 1, the objectives have been achieved and the hypothesis have been confirmed, as follows.

a. Main objectives

To analyze the discourse differences and similarities in a Spanish-Chinese parallel corpus with the aim to use this information in tasks related to translation and language learning.

Although there are three parallel corpora for Spanish and Chinese, as the explanation in Chapter 4, none of these corpora can be used for the discourse analysis between the two languages. We have elaborated a new Spanish-Chinese parallel corpus. To obtain a heterogeneous corpus, we select texts of distinct domains and genres from different sources. Meanwhile, to avoid bias, the texts are from different authors. Moreover, to guarantee the original discourse structure of each text, all the

selected texts in the corpus are the Spanish texts and their parallel Chinese texts translated by Chinese translators. The annotation allows us to explore the discourse differences and similarities of the parallel contents in the corpus.

By using the qualitative evaluation method for the discourse annotation, we have evaluated elements of Nuclearity (N), Relation (R), Constituent (C) and Attachment (A). The mentioned and evaluated elements gives us the useful information to create the translation protocol with discourse recommendations and language learning tasks.

b. Specific objectives

b1. *To create a Spanish-Chinese parallel corpus annotated with different discourse information in the framework of RST: discourse segments, CU, discourse relations, and discourse structure.*

This objective has been achieved because of the a new Spanish-Chinese parallel corpus we make for the study. Our corpus contains 100 parallel texts. We also use two parsers to enrich the POS information for both Spanish subcorpus and Chinese subcorpus.

The annotation tasks has been divided into different steps: (i) corpus compilation, (ii) discourse segmentation and evaluation, (iii) CU annotation and evaluation, (iv) discourse structure annotation and evaluation. For each annotation step, we have invited experts with experience in discourse analysis to participate.

b2. *To develop an online interface that can search the annotated discourse information in the corpus.*

This objective has been achieved because we make the corpus accessible and the corpus can be consulted here: `http://ixa2.si.ehu.es/rst/zh/`. The related information of the corpus is also accessible.

b3. *To compare the Spanish subcorpus with the Chinese subcorpus to detect the discourse differences and similarities between this language pair, and relate them with translation strategies.*

This objective has been achieved because we compare the segmentation differences and similarities between Spanish and Chinese. Moreover, following the qualitative evaluation for discourse structure annotation, we compare the discourse elements of Nuclearity (N), Relation (R), Constituent (C) and Attachment (A) between Spanish and Chinese.

b4. *To analyze how the discourse information are expressed formally in both languages, in terms of the annotated discourse elements in the parallel corpus, discourse relations, type of discourse relations (N-S or N-N), the order of EDUs, and the discourse markers.*

This objective has been achieved because we analyze the annotation results of each research step regarding the following discourse elements: discourse relations, type of discourse relations (N-S or N-N), the order of EDUs, and the discourse markers.

For the discourse segmentation, based on linguistic function and linguistic form, we elaborate a preliminary segmentation criteria that can be applied for both Spanish and Chinese. After the segmentation work, we analyze the segmentation disagreements to get the final segmentation criteria.

For the CU annotation and evaluation, we annotate the CUs of all the texts in the

corpus. We analyze the CU annotation disagreements. Additionally, we also conclude the representative words of the CUs in the corpus.

For the discourse structure annotation and evaluation, we use the qualitative evaluation proposed by Iruskieta, da Cunha and Taboada (2015), as the indication in Chapter 5. The evaluated elements are Nuclearity (N), Relation (R), Constituent (C) and Attachment (A). Besides, we analyzed the annotation disagreements to get the GS corpus.

b5. *To elaborate a translation protocol with discourse recommendations for Spanish-Chinese translation, based on the comparison and analysis produced in points 2 and 4.*

This objective has been achieved because we develop a protocol for Spanish-Chinese translation. The translation protocol that we make for Spanish-Chinese translation is the first one that gives the recommendations from discourse level; especially, to detect the translation strategy used under RST.

The recommendations in the protocol have been inserted with the real cases from the corpus and each recommendation contains the descriptions of discourse relations, type of discourse relation, discourse markers, and order of EDUs.

b6. *To design a Spanish-Chinese language learning task in terms of the annotation of the created parallel corpus.*

This objective has been achieved because based on the annotation results, we extract manually all the DMs from the corpus. Then, we develop the first multiple choice exercise module for L2 learning between the two languages under RST.

Hypothesis

c1. *Although Spanish and Chinese are from two different language systems, there will be certain discourse similarities between the languages in a parallel corpus.*

This hypothesis has been confirmed because from the annotation for both Spanish and Chinese subcorpora we get the discourse similarities regarding segmentation, CU, discourse relation, and order of EDUs.

c2. *The differences between the discourse produced in Spanish and Chinese could be modelled by using discourse information given in the framework of RST.*

This hypothesis has been confirmed because for the discourse segmentation differences, differences of CUs, differences of discourse relations, and differences order of EDUs, all these differences have been modelled.

c3. *The different discourse elements used in the framework of RST are adequate to formalize discourse equivalences between Spanish and Chinese.*

This hypothesis has been confirmed because the discourse elements that we detect in the corpus can be compared between Spanish and Chinese.

c4. *The use of a discourse-annotated Spanish-Chinese parallel corpus in the framework of RST would allow to obtain useful data for translation and language learning tasks.*

This hypothesis has been confirmed because we can use the comparison results to detect the translation strategy for the elaboration of the translation protocol and developing a task for Spanish and Chinese language learning.

8.2 Contributions

The main contributions of this dissertation address with the comparative discourse analysis between Spanish and Chinese under RST and the development of the language learning sources between them. The research are published in 6 peer-reviewed publications as listed in the Chapter 1, in which we describe the research process, as well as the evaluation and analysis. The following are the contributions that resulted from the work presented in this thesis:

- *The RST Spanish-Chinese parallel corpus*. As indicated, there is no Spanish-Chinese parallel corpus whose theoretical framework is RST. Our corpus is the first one with the annotated discourse information. All the texts can be downloaded online and the corpus can be used for different NLP studies, although the research purposes are diverse.
- *Segmentation*. Segmentation is a crucial step for corpus annotation from discourse level. Our work is the first that create the segmentation criteria for both Spanish and Chinese. All the segmented texts are open to the academic community. The segmentation can be used for other NLP tasks, for instance, discourse parsing, text mining, information retrieval, etc. The gold standard can also be useful for other languages under discourse analysis.
- *Central Unit (CU)*. Central Unit is the main idea of the text. This work is the first one annotates the CUs for both Spanish and Chinese. We also annotate the possible words that can represent the CUs. This part can be used for text summarization, word embedding, text generation, among others.
- *Discourse structure*. Discourse relations show the coherence of a language. Each discourse element of a language are detected through discourse structure. In this work, we list all the discourse similarities and differences between Spanish and Chinese by annotating the corpus. In addition, we have analyzed the discourse differences and similarities, for both discourse differences and similarities, we elaborate a protocol to include the related information.
- *Translation protocol with recommendations*. As the first translation protocol with recommendations for Spanish-Chinese translation from discourse level, the human translators who work with the language pair can adopt the recommendations for Spanish-Chinese translation tasks.
- *Question generation system*. As the first question-answering module for Spanish-Chinese language learning with discourse information, this part can help the Spanish-Chinese language teaching activities in the class. It can also text the students' language level through their answers. Moreover, language language learners can take our system as an implement resource to improve their language levels by doing the exercises within the system.

8.3 Limitations

Although we have developed the first RST Spanish-Chinese Treebank that can help the language translation and language learning, this study still contains the following limitation:

- *Corpus size*. As we explained before, due to the limited adequate language

resources for Spanish-Chinese discourse analysis, we decide to use only 50 Spanish texts and their translated Chinese texts.

- *Evaluation of the translation protocol*. Due to the limited time, we do not carry out the evaluation work of our translation protocol. Still, we do not know the comments either suggestions from other Spanish-Chinese translators.
- *Question-answering module*. As we analyzed in Chapter 7, this system still contains limitations on the recognition of DMs for both Spanish and Chinese subcorpora.

8.4 Future Work

We create the RST Spanish-Chinese Treebank with the purpose of helping the translation between Spanish and Chinese, and also Spanish-Chinese language learning. Although we have realized different tasks, still we can set new lines of research based on this dissertation, which are the following:

- *Discourse parsing*. Since we have annotated the discourse structure for both Spanish and Chinese, therefore, in the future, we have the intention to develop the parser that can parse the Spanish and Chinese from discourse level automatically.
- *Question-answering module*. Our question-answering module cannot grade for the Chinese part yet, for the future work, we will make the module can grade for the Chinese part also. We will also add more functions for this module, for example, the option of the POS information for two languages, the selection of the prepositional for Spanish and Chinese, the selection of discourse relation, etc.

References

Abelen Eric, Redeker Gisela, and Thompson Sandra A. 1993. The Rhetorical Structure of US-American and Dutch fund-raising letters. *Text*, 13(3): 323-350

Albert-Ludwigs Christian Mair. 2007. The FLOB Corpus (online). `http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html` [Last consulted: 27 of July of 2016].

Altenberg Bengt. 2002. Concessive connectors in English and Swedish. *Information Structure in a Cross Linguistic Perspective*, 21–43.

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

Aston Guy, and Burnard Lou. 1998. *The BNC Handbook*. . Edinburgh: Edinburgh 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.

Bengoetxea Kepa, and Iruskieta Mikel. 2018. A Supervised Central Unit Detector for Spanish. *Procesamiento del Lenguaje Natural*, 60: 29-36.

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.

Burstein Jill, Marcu Daniel, Andreyev Slava, and Chodorow Martin. 2001. Towards Automatic Classification of Discourse Elements in Essays. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL'2001)*, 98-105.

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

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

Cao Shuyua, 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.

Cao Shuyua, da Cunha Iria, and Iruskieta Mikel. 2017. Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *EPiC Series in Language and Linguistics*, 2: 12.

Cardoso Paula C. F., Pardo Thiago A. S., and Taboada Maite. 2017. Subtopic annotation and automatic segmentation for news texts in Brazilian Portuguese. *Corpora*, 12(1): 23-54.

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

Carlson Lynn, Marcu Daniel, and Okurowski Mary Ellen. 2002. RST Discourse Treebanks [Corpus]. Linguistic Data Consortium.

Carreras Xavier, Chao Isaac, Padró Lluís, and Padró Muntsa. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, 239-242.

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.

Chen Wei, Aist Gregory, and Mostow Jack. 2009. Generating Questions Automatically from Informational Text. In *The 2nd Workshop on Question Generation, volume 1 of AIED 2009 Workshops Proceedings*, 17-24.

Chien Yi-Shan. 2012. *Análisis contrastivo de los marcadores condicionales del español y del chino*. PhD thesis. Salamanca: Universidad de Salamanca.

Chief Lian-Cheng, Huang Chu-Ren, Chen Keh-Jiann, Tsai Mei-Chih, and Chang Li-li. What Can Near Synonyms Tell Us? *Computational Linguistics and Chinese Language Processing*, 5(1): 47-60.

Costa-jussà Marta R., Henríquez Carlos A., and Banchs Rafael E. 2004. Evaluation Indirect Stretesgies for Chinese-Spanish Statistical Machine Translation. *Journal of Artificial Intelligence Research*, 45: 761-780.

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

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

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

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

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

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

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.

Das Debopam. 2014. *Signalling of Coherence Relations in Discourse*. PhD thesis. Vancouver: Simon Fraser University.

Das Debopam, and Taboada Maite. 2017. RST Signalling Corpus: a corpus of signals of coherence. *Language Resources and Evaluation*, 1-36.

Das Debopam, Taboada Maite, and Stede Manfred. 2017. The Good, the Bad, and the Disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms*, 11-19.

Delin Judy, Hartley Anthony, Paris Célic, Scott Donia, and Linden, Keith Vander 1994. Expressing procedural relationships in multilingual instructions. In *Proceedings of Seventh International Workshop on Natural Language Generation*, 61-70.

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

Fabricius-Hansen Cathrine 2005. Elusive connectives: A case study on the explicitness dimension of discourse coherence. *Linguistics*, 43 (1): 17-48.

Fang Yan. 2008. *Zhongguo yingyu yupian tezheng yanjiu: zhongmei shiguxing xinwenwenben duibi fenxi* (中国英语语篇特征研究：中美事故性新闻文本对比分析 , *[*Discourse of China English: Rhetorical Relations in News Texts in China Daily and The New York Times*]*). Master thesis. Shanghai: Donghua University.

Fløttum Kjersti. 2002. La polyphonie dans une perspective macro-semantique. *Marco-syntax et marco-semantique*, 337-359.

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.

Fournier Chris. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'2013)*, 1702-1712.

Fraser Bruce. 1999. What are discourse markers? *Journal of Pragmatics*, 31: 931-952.

Guo Fang. 2015. A Review of Discourse Markers from the Functional Perspective. *Journal of Arts and Humanities*, 4 (4): 69-75.

Guo Qiang. 2014. *Jiyu xiucijiegou lilun de yingrihan shelun xiuciguanxi de duibi yanjiu* (基于修辞结构理论的英日汉社论修辞关系的对比研究, *[*A Comparative Study of the Rhetorical Relations in English, Japanese and Chinese Editorials - Based on Rhetorical Structure Theory*]*). Master thesis. Changchun: Northeast Normal University.

Guy Ramsay. 2000. Linearity in Rhetorical Organisation: A Comparative Cross-cultural Analysis of Newstext from the People's Republic of China and Australia. *International Journal of Applied Linguistics*, 10(2): 241-58.

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

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 NNAACL-HLT 2013*, 31-33.

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

Heilman Michael, and Smith Noah A. 2010. Good Question! Statistical Ranking for Question Generation. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (NAACL'2010)*, 609-617.

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 guidelines, and on-line interface. In *Proceedings of the second workshop on Chinese language processing*, 29-37.

Huong Le Thanh. 2007. An approach in automatically generating discourse structure of text. *Journal of Computer Science and Cybernetics*, 23(3): 212-230.

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

Iruskieta Mikel, Díaz de Ilarraza Arantza, Labaka Gorka, and Lersundi Mikel. 2015. Detection of Central Units in Basque Scientific Abstracts. In *Proceedings of the 5th Workshop "RST and Discourse Studies"*.

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

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

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

Iruskieta Mikel, Díaz de IIarraza Arantza, and Lersundi Mikel. 2013. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 1-32.

Johns Tim. 2002. Data-Driven learning: The perpetual challenge. *Language and Computers*, 1: 107-117.

Knott Alistair, and Sanders Ted. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2): 135-175.

Kumpf Lorrane. 1975. *Structuring Narratives in a Second Language: A description of Rhetoric and Grammar*. PhD thesis. Los Angeles: University of California.

Kong, Kenneth C. C. 1998. Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text*, 18(1): 103-141.

Lavid Julia, Arús Jorge, and Zamorano Juan Rafael. 2010. Desiging and exploiting a small online English-Spanish parallel corpus for language teaching purposes. *Corpus-Based Approach to English Language Teaching*, 138-148.

Leech Geoffrey. 1992. Corpora and theories of linguistic performance. *Directions in Corpus Linguistics*, 105-122.

Levy Roger, and Manning Christopher. 2003. Is it harder to parse Chinese or Chinese Treebank? 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 Hongkun , and Liao Zhihua. 2015. A Rhetorical Structure Theory Based Approach to Texture Coherence in Chinese EFL Learners' Argumentative Writing. *Overseas English*, 16: 201-204.

Li Hongzheng, Langlais Philippe, and Jin Yaohong. 2017. Translating Implicit Discourse Connectives Based on Crosslingual Annotation and Alignment. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, 93-98.

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.

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

Liang Shanshan., and Yang Zhenglin. 2016. *Hanguo xuesheng kouyu duochong yinguozhuanzhe yupian shiyong qingkuang fenxi* (韩国学生口语多重因果转折语篇使用情况分析, *[*A study of Multiple Causal and Transitional Discourse in Korean Students' Spoken Chinese*]*), *Chinese Teaching in the world*, 30(3): 356-367.

Mann William C., Matthiessen Christian M. I. M.,and Thompson Sandra A. 1989. *Rhetorical structure theory and text analysis*. ISI research report. California: University of California.

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

Marcu Daniel. 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.

Marcu Daniel, Carlson Lynn, Watanabe Maki. 2000. The automatic translation of discourse structures. In *Proceedings of 1st North American Chapter of the Association for Computational Linguistics conferences (NAACL'2000)*, 9-17.

Maziero Erick Galani, Pardo Thiago Alexandre Salgueiro, da Cunha Iria, Torres-Moreno Juan-Manuel, and SanJuan Eric. 2011. Dizer 2.0-An Adaptable

On-line Discourse Parser. In *Proceedings of Anais do III Workshop "A RST e os Estudos do Texto"*, 1-17.

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.

McCarthy M, and Carter R. 2001. Size isn't everything: spoken English, corpus and the classroom. *TESOL Quarterly*, 35(2): 337-340.

McEnery Tony, and Xiao Richard. 2004. The Lancaster Corpus of Mandarin Chinese [online]. `http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/` [Last consulted: 27 of July of 2016].

McEnery Tony, Xiao Richard, and Tono Yukio. 2006. *Corpus-Based Language Studies: an advanced book.* . New York: Routledge.

McEnery Tony, and Hardie Andrew. 2012. *Corpus linguistics: method, theory and practice*. New York: Cambridge University Press.

Meyer Thomas, and Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2012)*, 129-138.

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.

Miltsakaki Eleni, Prasad Rashmi, Joshi Aravind, and Webber Bonnie. 2004. The Penn Discourse Treebank. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC'2004)*, 2237-2240.

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.

Mohamed Aysha H, and Omer Majzoub R. 1999. Syntax as a Marker of Rhetorical Orgnization in Written Texts: Arabic and English. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 37(4): 291-305.

Mosegaard Hansen Maj-Britt. 1998. The semantic status of discourse markers. *Lingua*, 104: 235-260.

Mosegaard Hansen Maj-Britt. 2008. *Particles at the Semantics/Pragmatics Interface:*

Synchronic and Diachronic Issues. Amsterdam: Elsevier.

Müller Simone. 2005. *Discourse Markers in Native and Non-native English Discourse*. Amsterdam/Philadelphia: John Benjamins.

Neff-van Aertselaer JoAnne. 2015. *Learner Corpora and Discourse*. Cambridge: Cambridge University Press.

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

Olney Andrew M., and Graesser Arthur C., and Person Natalie K. 2012. Question Generation from Concept Maps. *Dialogue and Discourse*, 3(2): 75-99.

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

Pardo Thiago Alexandre Salgueiro, 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.

Qian Zhiying. 2005. *Yinghan/Hanying pingxingfanyi yuliaoku de sheji jiqi zai fanyi zhong de yingyong* (汉英/汉英平行翻译语料库的设计及其在翻译中的应用 *[*The Design of Chinese-English/English-Chinese Parallel Translation Corpus and its Application in Translation Studies*]*). Master thesis. Shanghai: East China Normal University.

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.

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

Resink Philip, Olsen Mari Broman, and Diab Mona. 1999. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". Computers and the Humanities, 33: 129-153.

Salkie, Raphael; Oates, Sarah Louis. 1999. Contrast and concession in French and English. *Languages in Contrast*, 2(1): 27-56.

Sarjala Marja. 1994. Signalling of reason and cause relations in academic discourse. *Anglicana Turkuensia*, 13: 89-98.

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

Scott Donia R., Delin Judy, and Hartley Anthony F. 1998. Identifying congruent pragmatic relations in procedural texts. *Languages in contrast*, 1(1): 45-82.

Seuren Pieter. 1998. *Western Linguistics: An Historical Introduction*. Oxford: Blackwell.

Sidarenka Uladzimir, Peldszus Andreas, and Stede Manfred. 2015. Discourse Segmentation of German Texts. *Journal for Language Technology and Computational Linguistics*, 30(1): 71-98.

Sinclair John. 1996. EAGLES preliminary recommendations on corpus typology. *Expert Advisory Group on Language Engineering Standards report*, 1-36.

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*.

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

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.

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

Taboada Maite. 2004. Rhetorical relations in dialogue: A contrastive study. *Discourse across Languages and Cultures*, 75-97.

Taboada Maite. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38: 567-592.

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.

Taboada Maite, and Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, 6: 17-41.

Tofiloski Milan, Brooke Julian, and Taboada Maite. 2009. A Syntactic and Lexical-Based Discourse Segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL'2009)*, 77–80.

Tognini-Bonelli Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamin.

Toldova Svetlana, Pisarevskaya Dina, Ananyeva Margarita, Kobozeva Maria, Nasedkin Alexander, Nikiforova Sofia, Pavlova Irina, and Shelepov Alexey. 2017. Rhetorical relation markers in Russian RST Teebank. 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 Dijk Teun A. 1980. *MACROSTRUCTURES: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. New Jersey: Lawrence Erlbaum Associations.

van der Vliet Nynke. 2010. Inter annotator agreement in discourse analysis. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.8561&rep=rep1&type=pdf` [Last consulted: 29 of Oct, 2017]

Vargas-Urpi Mireia, 2018. Judged in a Foreign Language: A Chinese-Spanish Court Interpreting Case Study. *The European Legacy. Toward New Paradigms*. DOI: 10.1080/10848770.2018.1492814.

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

Wang Yi-Chen. 2013. Los marcadores conversacionales en el subtitulado del español al chino: análisis de La mala educación y Volver de Pedro Almodóvar. PhD thesis. Barcelona: Universitat Autònoma de Barcelona.

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

Wolf Florian, and Gibson Edward. 2005 Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2): 249-290.

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

Xiao Richard, and McEnery Tony. 2010. *Corpus-Based Contrastive Studies of English and Chinese*. New York: Routledge

Xu Shengqin, and Li Peifeng. 2013. Recognizing Chinese Elementary Discourse Unit on Comma. In *Proceedings of International Conference on Asian Language Processing (IALP'2013)*, 3-6

Xue Nianwen. 2005. Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, 84-91.

Xue Nianwen, and Yang Yaqin. 2011. Chinese sentences segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'2011)*, 631-635.

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

Yang Yaqin and Xue Nianwen. 2012. Chinese Comma Disambiguation for Discourse Analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'2012)*, 786-794.

Yang Yunmei. 2008. *Hanxi yanyu duibi yanjiu---Yi Tangjikede weili* (汉西谚语对比研究---以《唐吉坷德》为例 *[*Comparative study of Spanish and Chinese proverbs --- case study of Don Quijote*]*). Master thesis. Shandong: Shandong University.

Yao Junming. 2008. *Estudio comparativo de los marcadores del discurso en español y en chino a través de diálogos cinematográficos*. PhD thesis. Valladolid: Universidad de Valladolid.

Yu Shiwen, Duan Huiming, and Zhu Xuefeng. 2005. *Ciyujianlei ji dongci xiang mingci piaoyi xianxiang de jiliang fenxi* (词语兼类暨动词向名词飘移现象的计量分析 *[*A Quantitative Analysis on Multi-class Words and Shift from Verbs to Nouns in Chinese*]*). Natural language understanding and large-scale computing content (自然语言理解与大规模内容计算), 70-76.

Yue Ming. 2006. *Hanyu caijingpinglun de xiucijiegou biaozhu ji pianzhangyanjiu* (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.

Zhang Li. 2016. *Mianxiang jiqifanyi huihua fenxi de xiucijiegou lilun yingyong yanjiu* (面向机器翻译会话分析的修辞结构理论应用研究, *[*Onthe Application of the Rhetorical Structure Theory to Conversation Analysis in the Context of Machine Translation*]*). Master thesis. Xi'an:Shananxi Normal University.

Zhang Man. 2016. A multidimensional analysis of metadiscourse markers across written register. *Discourse Studies*, 18(2): 1-19.

Zhang Yi. 2010. *Zhongguo daxuesheng he yingmei liuxuesheng tongti yingwenzuowen de xiucijiegou fenxi yu bijiao* (中国大学生和英美留学生同题英文作文的修辞结构分析与比较, *[*An Analysis of Compositions by Chinese

Students and Native Speakers Based on Rhetorical Structure Theory*]*). Master thesis. Hangzhou: Zhejiang University.

Zhou Yuping, and Xue Nianwen. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'2012)*, 69-77.

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.

Appendices

Appendix A
Corpus Text Links

This part gives the link of each text in the corpus.

Text name	Link
BMCS1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_CHN1-GS.txt
BMCS2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_CHN2-GS.txt
BMCS3	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_ESP3-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_CHN3-GS.txt
BMCS4	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_ESP4-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_CHN4-GS.txt
BMCS5	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_ESP5-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/BMCS_CHN5-GS.txt

Table A.1 Text links of the BMCS part

Text name	Link
CCICE1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_CHN1-GS.txt
CCICE2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_CHN2-GS.txt
CCICE3	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_ESP3-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_CHN3-GS.txt
CCICE4	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_ESP4-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_CHN4-GS.txt
CCICE5	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_ESP5-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/CCICE_CHN5-GS.txt

Table A.2 Text links of the CCICE part

Text name	Link
EEP1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN1-GS.txt
EEP2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN2-GS.txt
EEP3	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP3-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN3-GS.txt
EEP4	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP4-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN4-GS.txt
EEP5	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP5-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN5-GS.txt
EEP6	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP6-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN6-GS.txt
EEP7	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP7-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN7-GS.txt
EEP8	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP8-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN8-GS.txt

Table A.3 Text links of the EEP part

Text name	Link
FICB1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_CHN1-GS.txt
FICB2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_CHN2-GS.txt
FICB3	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_ESP3-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_CHN3-GS.txt
FICB4	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_ESP4-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_CHN4-GS.txt
FICB5	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_ESP5-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FICB_CHN5-GS.txt

Table A.4 Text links of the FICB part

Text name	Link
FCEC1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FCEC_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FCEC_CHN1-GS.txt
FCEC2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FCEC_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/FCEC_CHN2-GS.txt

Table A.5 Text links of the FCEC part

Text name	Link
ICP1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN1-GS.txt
ICP2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN2-GS.txt
ICP3	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP3-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN3-GS.txt
ICP4	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP4-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN4-GS.txt
ICP5	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP5-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN5-GS.txt
ICP6	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP6-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN6-GS.txt
ICP7	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_ESP7-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICP_CHN7-GS.txt
ICP8	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_ESP8-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/EEP_CHN8-GS.txt

Table A.6 Text links of the ICP part

Text name	Link
ICEG1	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICEG_ESP1-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICEG_CHN1-GS.txt
ICEG2	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICEG_ESP2-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/ICEG_CHN2-GS.txt

Table A.7 Text links of the ICEG part

Text name	Link
TERM18	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM18_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM18_CHN-GS.txt
TERM19	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM19_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM19_CHN-GS.txt
TERM23	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM23_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM23_CHN-GS.txt
TERM25	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM25_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM25_CHN-GS.txt
TERM28	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM28_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM28_CHN-GS.txt
TERM29	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM29_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM29_CHN-GS.txt
TERM30	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM30_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM30_CHN-GS.txt
TERM31	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM31_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM31_CHN-GS.txt
TERM32	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM32_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM32_CHN-GS.txt
TERM34	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM34_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM34_CHN-GS.txt
TERM38	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM38_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM38_CHN-GS.txt
TERM39	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM39_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM39_CHN-GS.txt
TERM40	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM40_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM40_CHN-GS.txt
TERM50	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM50_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM50_CHN-GS.txt
TERM51	http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM51_ESP-GS.txt http://ixa2.si.ehu.es/rst/zh/zh.TXT/txt-ak/TERM51_CHN-GS.txt

Table A.8 Text links of the TERM part

Appendix B
Examples of Criteria for Annotation Steps

This part gives the examples of the criteria for each annotation step: discourse segmentation, Central Unit (CU) and discourse structure.

B.1 Discourse Segmentation Criteria Examples

B.1.1 Paragraphs and line breaks

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

Text name: FCEC1
Text in Spanish: [Queridos amigos**,**] […]
English: [Dear friends,] [...]
Text in Chinese: [亲爱的朋友们，] [...]
English: [Dear friends,] [...]

Explanation: Both Spanish and Chinese passages start with a greeting, a comma and a line break follow after the greeting.

B.1.2 Sentences and periods

Definition: In our study, the period marks the end of an independent EDU.

Text name: ICP4
Text in Spanish: [Las convocatorias de plazas de profesor titular para los Institutos Cervantes **aparecen** publicadas en la prensa española**.**] [También aparecen todas las convocatorias en la página web del Instituto Cervantes.]
English: [The calls of position of professor titular for those Cervantes Institutes appear notices in press Spanish.] [Also appear all the calls in the webpage of Cervantes Institute.]
Text in Chinese: [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: In the Spanish passage, the first EDU ends with the verb "*aparecer*" ('appear') and a period. In the Chinese passage, after the word "*gongbu*" (公布) ('publish'), there is a period, followed by another sentence.

B.1.3 Question mark and exclamation mark

Definition: Both marks are signals of a sentence boundary.

Text name: TERM34

Text in Spanish: [**¿**Dónde se encuentra el límite**?**] [**¿**Dónde se encuentra la clave, en el carácter de predicado/no predicado, en relación argumental, de campo, posesiva o en otro aspecto**?**]

English: [Where find the limitation?] [Where find the key, in the character of predicate/no predicate, in relation argument, possessive or in other aspect?]

Text in Chinese: [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?]

Explanation: In both Spanish and Chinese passages, at the end of each sentence, there is a question mark as the sentence boundary.

B.1.4 Other EDUs should have a main verb or an adjunct verb phrase[74]

Definition: This is a basic segmentation criterion and segmentation criteria bellow should follow this rule. Titles are considered as the exceptions, whether they contain a verb or not, titles are always EDUs.

Text name: CCICE3

Text in Spanish: [En el mes de octubre el Tesoro **colocó** 14.399 millones en cuatro emisiones.]

English: [The month of October the Treasury placed 14,399 millions in four issues.]

Text in Chinese: [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 millions in four issues.]

Explanation: In the Spanish example, the verb "*colocar*" ('place') is included in the sentence meanwhile in the Chinese passage, the word "*chouji*" (筹集) is a verb and means 'raise' in English.

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

B.1.5 Discourse Marker (DM), verb and comma

Definition: 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.

Text name: TERM31

Text in Spanish: [**Como** se suelen emplear modelos morfosintácticos,] [resulta conveniente tener analizado el texto o, por lo menos, etiquetado.]

English: [As it often uses morph-syntax models,] [realizes convenient has analyzed the text, or, at least, labeled.]

Text in Chinese: [**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: In the Spanish example, the sentence begins with the DM "*como*" ('due to/since'), which holds a CAUSE relation between the two segmented parts. Two verbs are included in the first EDU, "*soler*" ('be accustomed to') and "*emplear*" ('utilize'). The verbs "*resultar*" ('turn out') and "*tener*" ('have') are included in the second EDU. In the Chinese example, 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').

Text name: TERM19

Text in Spanish: [En ese caso, la normalización **no sólo** dejaría de ser eficaz,] [**sino** que no cumpliría sus finalidades.]

English: [In this case, standardization not only ceases to be effective,] [but it would not comply its goals.]

Text in Chinese: [此时，标准**不但**会失效，] [**而且**也不能发挥作用。]

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

Explanation: In the Spanish passage, the DM "*no sólo*" ('not only') is in the first EDU and another DM "*sino*" ('but') is in the second EDU. In its parallel Chinese passage, 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.

B.1.6 Semicolon plus adjunct verb phrase

Definition*:* If there is a semicolon and the sentence is being divided into adjunct verb phrases, the separated parts are considered as EDUs.

Text name: TERM34

Text in Spanish: [Por ejemplo, el adjetivo marginal en inglés tiene un uso referencial y predicativo respectivamente en los conjuntos marginal nota y marginal case**;**] [en castellano, por el contrario, diferencian el uso referencial y el predicativo empleando nota al margen y caso marginal.]

English: [For example, the adjective marginal in English has a use referential and predicative respectively in the joint marginal note and marginal case;] [in Spanish, in contrast, differentiate referential and predicative employ note to margin and case marginal.]

Text in Chinese: [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: The verb “*tener*” (‘have’) is in the first segment in the Spanish text, another verb appears in the second segment is “*diferenciar*” (‘differentiate’). In the Chinese text, the first segment contains a verb “*yong*” (用) (‘apply’), the second segment contains a verb “*caiyong*” (采用) (‘use’). In the Spanish text, we can that a semicolon separates the sentence into two parts; the two parts form a CONTRAST relation. The Chinese text shows the same circumstance, a semicolon separates the text into two parts, the two separated parts from a CONTRAST relation.

B.1.7 Parenthetical and dash

Definition: 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[75].

Text name: TERM29

Text in Spanish: [Efectivamente, el diseño y la gestión de las bases de datos terminológicos plantean problemas diversos tanto de índole teórica y

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

metodológica] [**(**¿cómo se representa un término?,] [¿existe una representación mínima?,] [¿cómo se clasifican los términos? **)**] […]

English: [Indeed, the design and the management of the basics of data terminology pose problems various much of nature theory and methodological] [(how represents a terminology?,] [exists a representation minor?,] [how classify the terminologies?)] […]

Text in Chinese: [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？ **)**] […]

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: In the Spanish example, the parenthetical unit does not modify its previous part; it should be an independent segment. Three verbs are included in each EDU, which are "*plantear*" ('set out'), "*representar*" ('represent'), and "*clasificar*" ('clasify'). In the Chinese passage, the parenthetical unit does not modify its previous part neither. 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.

B.1.8 Coordination and ellipsis with verbs

Definition: Coordinated clauses with verbs are considered independent EDUs (even they include a null subject).

Text name: TERM25

Text in Spanish: [...] [en la Universidad de Deusto **venimos** traduciendo textos doctrinales del campo del Derecho desde 1994] [y **queremos** expresar las dificultades que hemos tenido a lo largo de estos años y, así mismo, también los logros conseguidos,] […]

English: [...] [in the University of Deusto we have been translating texts doctrinal in the campus of Law from 1994] [and we want express the difficulties that we have had over years and, likewise, also the achievements] […]

Text in Chinese: […] [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 Spanish text, the two-segmented parts have the same implicit subject "*nosotros*" ('we'), the two coordinated clauses include verbs "*venir*" ('come')

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

B.1.9 Relative, modifying and appositive clauses

Definition: Relative clauses, clauses that modifies a noun or adjective or appositive clauses are not considered independent EDUs.

Text name: BMCS5
Text in Spanish: [Las más modernas herramientas de comunicación **(chat, foro, blog, wiki y correo electrónico)**, que permiten dinámicas de grupo entre personas desde cualquier lugar del mundo.]
English: [The more modernized tools of communication (chat, forum, blog, wiki and email), that permit dynamic of group between persons from anywhere of the world.]
Text in Chinese: [现代化的交流工具**（聊天，论坛，博客，wiki 和电子邮件）**，辅助学生在任何地方都与组内同伴交流活动。]
English: [Modern communication tools (chats, forums, blogs, wiki and emails), helps students in anywhere with inside group companions interact.]

Explanation: The verb which is being included in the Spanish text is "*permitir*" ('permit'), and the Chinese text includes the verb "*fuzhu*" (辅助) ('help'). The names of the communication tools in the parenthetical part are appositive of the "*las más modernas herramientas de comunicación*" ('those modernized communication tools') in the Spanish text, and also appositive of "*xiandaihua de jiaoliugongju*" (现代化的交流工具) ('modernized communication tools') in the Chinese text. These names cannot compose an independent segment.

B.1.10 Reported speech

Definition: In this study, we do not consider reported speech as an independent EDU.

Text name: CCICE3
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, we donot segment the attributed parts as indepedent EDUs, we only segment within the attributed parts.

B.1.11 Truncated EDUs

Definition: For the cases of truncated EDUs, we use the non-relation label of Same-unit (Carlson, Marcu and Okurowski, 2003). Figure 11 and Figure 12 show the case of Same-unit from the research corpus (a Spanish-Chinese parallel case).

Text name: TERM29

Text in Spanish: [Efectivamente, el diseño y la gestión de las bases de datos terminológicos plantean problemas diversos tanto de índole teórica y metodológica] [(¿cómo se representa un término?,] [¿existe una representación mínima?,] [¿cómo se clasifican los términos?)] [como de índole informática] [(¿cuál debería ser la estructura de una base de datos terminológicos?,] [¿qué relaciones debe contemplar?,] [¿cuál es la unidad de un diccionario?).]

English: [Indeed, the design and the management of the basics of data terminology pose problems various much of nature theory and methodological] [(how represents a terminology?,] [exists a representation minor?,] [how classify the terminologies?)] [as of nature information technology,] [(which should be the structure of a base of data terminology?] [what relations should take into account?,] [which is the unit of a dictionary?).]

Text in Chinese: [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: In the Spanish text, we can see that the EDU(5-8) and the EDU(9-12) hold a Same-unit relation. Both the EDU(6-8) and the EDU(10-12) are the inserted parts of the sentence. The EDU(6-8) is an additional information of the EDU(5), and the EDU(10-12) is the additional information of the EDU(9). The Chinese text shows the same case, 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.

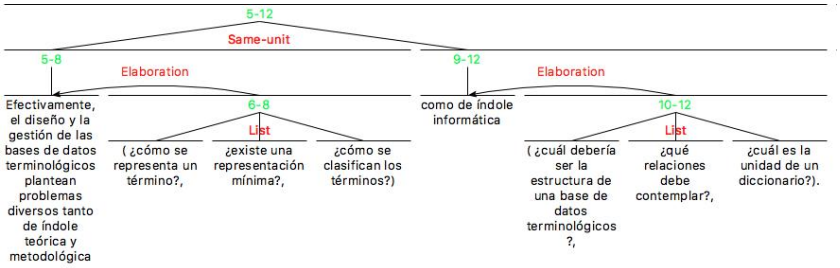

Figure B.1.11 Case of Same-unit in the corpus (Spanish text)

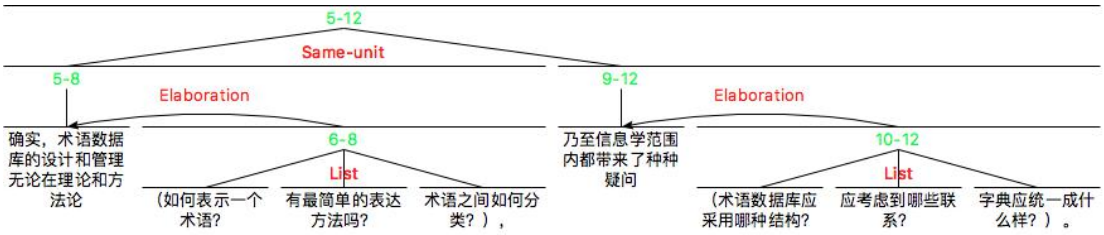

Figure B.1.12 Case of Same-unit in the corpus (Chinese text)

B.2 CU Annotation Examples with Representative Words

B.2.1 misión & 任务

Spanish Word *(occurrences in total)*: misión (1)

Text Name: FCEC2

Text in Spanish: La Fundación Consejo España China es una institución privada y sin ánimo de lucro constituida en el año 2004, cuya **misión** se encuadra en el llamado ejercicio de la Diplomacia Pública entre nuestros dos países.

English: The Foundation Council Spain China is an institute private and without profit founded in the year 2014, its task is framed in the so-called exercises of the Public Diplomacy between our two countries.

Chinese Word(occurrences in total): 任务 (1)

Text Name: FCEC2

Text in 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.

B.2.2 propósito & 旨在

Spanish Word(occurrences in total): propósito (1)

Text Name: TERM38

Text in Spanish: El **propósito** de esta comunicación es hacer una reflexión sobre los retos a que se está enfrentando la neología terminológica en la realidad actual, ...[76]

English: The purpose of this communication is to make a reflection about the challenges that it is facing of neology terminological in the reality current.

Chinese Word(occurrences in total): 旨在 (5)

Text Name: TERM38

Text in Chinese: 本文**旨在**分析当今新生术语所面临的挑战, ...

English: This paper aims to analyze the current nascent terminology that faces of challenges, ...

B.2.3 finalidad & 目标

Spanish Word(occurrences in total): finalidad (1)

Text Name: TERM50

Text in Spanish: La **finalidad** de esta ponencia se centra en la exposición de algunos de los resultados de la investigación llevada a cabo por los grupos de trabajo de estas universidades en los últimos tres años.

English: The goal of this paper concentrates on the presentation of some research results that was carried out by the groups of working of these universities in the past three years.

Chinese Word(occurrences in total): 目标 (2)

Text Name: TERM50

Text in Chinese: 该报告的**目标**在与展示几所大学的工作小组在近三年进行的研究成果。

English: The paper of goal is to present some university working groups in the recent carried out research results.

[76] For some cases the annotated CUs, in a text, the CU is not a complete sentence but the part of the sentence that can represent the main idea of the text.

B.2.4 objetivo & 目标

Spanish Word (occurrences in total): objetivo (1)

Text Name:TERM25

Text in Spanish: Así pues, durante estos últimos años el **objetivo** de nuestro trabajo ha sido lograr la confluencia en un mismo texto, por un lado, de las diferentes técnicas en el campo de la traducción y, por otro lado, del conocimiento en profundidad de los sistemas jurídicos que el euskera tiene que asimilar (el Derecho español, el francés y el de las Comunidades Europeas).

English: Therefore, during these last years the object of our work has been to achieve the influence in the same text, on one hand, the differences techniques in the field of translation and, on the other hand, of the knowledge in-depth of those systems legal that the Basque language has to assimilate (Spanish, French, and European Communities).

Chinese Word (occurrences in total): 目标 (2)

Text Name:TERM25

Text in Chinese: 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）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。

English: Therefore, recent years our working goal is to make the translation process that uses various methods (reasonable terminology, creation of new terminology) and the Basque must be able to in depth to meet the each law content (Spain, French and the European Union Law) in a single document.

B.2.5 difusión & 传播/推广

Spanish Word (occurrences in total): difusión (3)

Text Name: ICP3

Text in Spanish: Además de nuestros cursos de lengua española, el Instituto Cervantes de Pekín se ocupa de la **difusión** de la cultura española e hispanoamericana.

English: In addition to our Spanish courses, the Institute Cervantes of Beijing occupies the diffusion of the culture Spanish and Latin American.

Chinese Word (occurrences in total): 传播 (2)

Text Name: ICP3

Text in 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 besides of offering Spanish courses, also in charge of the diffusion of Spanish and Latin American culture.

Spanish Word (occurrences in total): difusión (3)

Text Name: BMCS1

Text in Spanish: Desde su apertura, este centro viene dedicando sus esfuerzos a la promoción y enseñanza de la lengua española y a la **difusión** de la cultura hispanohablante a través de la oferta de cursos de español adaptados a diferentes necesidades, así como la organización de eventos culturales y talleres y la disposición de la Biblioteca Miguel de Cervantes a todos los públicos.

English: Since its opening, this center has been dedicating its efforts to the promotion and teaching of the language Spanish and the diffusion of Spanish-speaking culture through the offer of classes of Spanish adapted to different needs, as well as the organization of events cultural and workshops and the layout of the Library Miguel of Cervantes to all audiences.

Chinese Word (occurrences in total): 推广 (3)[77]

Text in 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 have been committed to through adapted to different needs of Spanish courses, the organization of different cultural activities and salon workshops, and to face the public of Miguel de Cervantes Library to promote Spanish teaching and to broadcast Spanish-speaking countries' culture.

B.2.6 difundir & 推广

Spanish Word (occurrences in total): difundir (1)

Text Name: ICP6

Text in Spanish: Pero la presencia del Instituto Cervantes en China se entiende no sólo como vehículo para **difundir** la lengua y la cultura en español, sino también como plataforma abierta al diálogo con personas e instituciones de este país que nos acoge, ...

English: But the presence of the Institute Cervantes in China is understood not only as a vehicle for the diffusion the language and the culture in Spanish, but also as platform open to dialogues with persons and institutes of this country that us carry...

---

[77] Among the 3 CUs that include the word *tuiguang* (推广), we find that its correspond Spanish words are *difusión* (1 time), *difundir* (1 time) and no translation (1 time). For the case of translation, the translation has been used, the word *tuiguang* is an inserted word for enriching the context without changing its original Spanish meaning. This special case can be consulted in the CU of the text FCEC2.

Chinese Word (occurrences in total): 推广 (3)

Text Name: ICP6

Text in Chinese: 然而，塞万提斯学院在中国的设立不仅要为西班牙语和其文化的**推广**提供一个媒介，更是要为中国的众多文化机构和文化人士提供一个开放的平台，...

English: However, Cervantes Institute in China of establishment not only for Spanish and its culture of diffusion to offer a medium, but also for China of various cultural institutions and cultural people offers an open platform, ...

B.2.7 plantear & 阐述

Spanish Word (occurrences in total): plantear (1)

Text Name: TERM19

Text in Spanish: En esta comunicación, a partir de la experiencia en trabajos de normalización de terminología catalana, se **planteará** la necesidad social de la normalización terminológica, se comentarán algunas de las dificultades con que se enfrenta y se apuntarán ideas para su enfoque dentro de la sociedad actual.

English:In this communication, from the experience in workof normalization of terminology Catalan, it will state the necessary society of the normalization terminological, it will comment some of the difficulties with that it faces and it will discuss the ideas for its focus within the society current.

Chinese Word (occurrences in total): 阐述 (1)

Text Name: TERM19

Text in 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 standard the implementation gained experience, to state the establishment terminology standardization of necessary, also will discuss faces some difficulties, and will current society of this situation give some ideas.

B.2.8 tratar & 描述/旨在

Spanish Word (occurrences in total): tratar (3)

Text Name: TERM32

Text in Spanish: Esta ponencia **trata** del cómo crear un acuñador terminológico automatizado y convertirlo en una parte integral del banco terminológico.

English: This paper describes about how to create a coinage terminological automated and turn it into one part integral of bank terminological.

Chinese Word (occurrences in total): 描述 (4)[78]

Text Name: TERM32

Text in 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 paper attempts to describe how to create computers support under of terminology builder, and how to turn it into terminology group integral of part.

Spanish Word (occurrences in total): tratar (3)

Text Name: TERM30

Text in Spanish: Esta ponencia **tratará** de los desafíos con los que nos hemos encontrado, de las oportunidades identificadas y de las soluciones sugeridas, ...

English: This paper will describe of those challenges with that we have encountered, of those opportunities identified and of those solutions suggested, ...

Chinese Word (occurrences in total): 旨在 (5)

Text Name: TERM30

Text in Chinese: 此报告**旨在**描述我们在多语言区域管理专业术语的各种情况中发现的挑战以及对应的解决办法，...

English: This report aims to describe we in multilingual area management terminology of various situations found of challenges and their corresponding solutions, ...

B.2.9 pretender & 旨在

Spanish Word (occurrences in total): pretender (1)

Text Name: TERM28

Text in Spanish: Mediante la presente comunicación se **pretende** dar a conocer la metodología utilizada para la localización y análisis de la terminología jurídica en euskera en un corpus completo de un campo determinado.

English: Through the present communication it pretends give to know the methodology used for locating and analyzing of the terminology legal in Basque in a corpus complete of a field specific.

Chinese Word (occurrences in total): 旨在 (5)

Text Name: TERM28

Text in Chinese: 此次报告**旨在**说明在一个完全属于某一领域的语料中确定和分析巴斯克语法律术语使用的方法。

---

[78] For the text TERM40, the translation strategy has been used. The Chinese translation of the Spanish Word*nos referimos* (referirse) is *miaoshu* (描述).

English: The report aims to explain in a complete belonged to specific field of corpus to locate and analyze Basque legal terminology used of methodology.

B.2.10 intentar & 旨在

Spanish Word (occurrences in total): intentar (3)

Text Name: TERM18
Text in Spanish: En esta ponencia **intentamos** evaluar críticamente la tendencia señalada sobre estas líneas, tanto desde un punto de vista lingüístico como sociolingüístico.
English: In this report we intend to evaluate critically the tendency indicated about these lines, both from a point of view linguistic and sociolinguistic.

Chinese Word (occurrences in total): 旨在 (5)

Text Name: TERM18
Text in 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 linguistic view and sociolinguistic of views critically evaluate above mentioned tendency.

B.2.11 nuestro & 我们的

Spanish Word (occurrences in total): nuestro (5)

Text Name: BMCS2
Text in Spanish: Todos **nuestros** profesores son nativos, han recibido una formación específica en la enseñanza de español como lengua extranjera (ELE) y tienen experiencia docente en China.
English: All our teachers are natives, have received the training specific in teaching of Spanish as language (ELE) and have experiences teaching in China.

Chinese Word (occurrences in total): 我们的 (6)

Text Name: BMCS2
Text in 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: Our all teachers are all Spanish native teachers, have received ELE qualification training, and have in China teaching of experiences.

B.2.12 desde & 自...以来

Spanish Word (occurrences in total): desde (3)[79]

Text Name: EEP8

Text in Spanish: **Desde** su creación el año 1987, el grupo Capella de Ministrers, bajo la dirección de Carles Magraner, ha desarrollado una importante tarea investigadora y musicológica en favor del patrimonio musical español, desde el medioevo hasta el siglo XIX.

English: Since its creation the year 1987, the group Capella de Ministrers, under the direction of Carles Magraner, has developed an important task research and musicological in favor of heritage musical Spanish, since the Middle Ages to the century 19th.

Chinese Word (occurrences in total): 自...以来 (3)

Text Name: EEP8

Text in 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.

B.2.13 para & 为

Spanish Word (occurrences in total): para (12)[80]

Text Name: BMCS4

Text in Spanish: El Instituto Cervantes, siendo la institución que crea los exámenes DELE, ofrece los mejores cursos de preparación **para** los candidatos que desean obtener su diploma.

English: The Institute Cervantes, being the institution that creates the exams DELE, offers the better courses of preparation for the candidates that want to obtain their diploma.

---

[79] Totally, there are 3 annotated CUs in the Spanish subcorpus that contain the word *desde* ('from'), but only 1 CU contains this word whose Chinese translation is *zi...yilai* (自...以来).

[80] There are 12 CUs that contain the Spanish Word*para* ('for') in the Spanish subcorpus. However, in their parallel Chinese CUs, not all the translation is the Chinese Word*wei* (为), which also means 'for' in English.

Chinese Word (occurrences in total): 为 (17)[81]

Text Name: BMCS4

Text in 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 the main organizer, Cervantes Institute for who want to by exam to obtain level certificates of candidates offer high quality of preparation courses.

B.2.14 según & 据

Spanish Word (occurrences in total): según (2)

Text Name: CCICE2

Text in Spanish: España ha escalado hasta convertirse en el sexto mayor mercado de inversión inmobiliaria del mundo por delante de Francia, Canadá o China, **según** el último informe global publicado por la consultora CBRE.

English: Spain has climbed till convert itself the sixth largest market of investment real estate of world ahead of France, Canada or China, according to the last report global published by the consultancy CBRE.

Chinese Word (occurrences in total): 据 (3)[82]

Text in Chinese: **据**世邦魏理仕咨询公司 CBRE 发布的最新全球报告显示，西班牙已超越法国、加拿大或中国，跻身第六大房地产投资市场。

English: According to CBRE published the last global report shows, Spain has surpassed France, Canada or China, among the sixth largest real estate investment market.

B.2.15 este & 本/此

Spanish Word (occurrences in total): este (7)

Text Name: TERM40

Text in Spanish: Con **este** breve trabajo nos referiremos en particular a la herramienta lingüística que ofrece a los usuarios del sistema para la recuperación temática de la información: el Tesauro jurídico.

---

[81] The Chinese Word*wei* (为) is included in the 17 annotated CUs. this is because *wei* (为) is a multi meaning word. Besides of its meaning of 'for' (that corresponds to the Spanish Word*para*), it also means 'as' in English. Moreover, combining with other words, it can convert into different phrase.

[82] Among all the 3 annotated Chinese CUs that contain the word *ju* (据), there is a CU (text: CCICE5) whose parallel Spanish CU does not include *según* ('based'). The translator uses the translation strategy to add the word *ju* (据) without changing its original context in Spanish.

English: With this short work we will refer in particular to the tool linguistic that offers to the users of system for the recovery thematic of the information: the Thesaurus Legal.

Chinese Word (occurrences in total): 本 (11)

Text Name: TERM40

Text in 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 report aims to describe system users in making a thematic information recovery resorted linguistic tool: the Thesaurus Legal.

Spanish Word (occurrences in total): este (7)

Text Name: TERM31

Text in Spanish: El grupo IXA tiene la intención de desarrollar una herramienta de **este** tipo para el euskera.

English: The group IXA has the intention of developing a tool of this type for the Basque.

Chinese Word (occurrences in total): 此 (11)

Text Name: TERM31

Text in Chinese: IXA 研究组试图针对巴斯克语开发一个**此**类工具。

English: IXA research group intends for Basque to develop a this type tool.

B.2.16 esta & 本/此

SpanishWord (occurrences in total): esta (9)

Text Name: TERM51

Text in Spanish: Tras **esta** presentación la comunicación giraría en torno a: ...

English: After this presentation the communication would revolve around: ...

Chinese Word (occurrences in total): 本 (11)

Text Name: TERM51

Text in Chinese: 在陈述完上述观点后，**本**文将围绕下列问题展开：...

English: After stating the above mentioned views, this communication will revolve around the following issues: ...

Spanish Word (occurrences in total): esta (9)

Text Name: TERM18

Text in Spanish: En **esta** ponencia intentamos evaluar críticamente la tendencia señalada sobre estas líneas, tanto desde un punto de vista lingüístico como

sociolingüístico.

English: En **esta** ponencia intentamos evaluar críticamente la tendencia señalada sobre estas líneas, tanto desde un punto de vista lingüístico como sociolingüístico.

Chinese Word (occurrences in total): 此 (9)

Text in 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 linguistic view and sociolinguistic of views critically evaluate above mentioned tendency.

B.2.17 no solo... sino que también & (不仅仅)...同时

Spanish Word (occurrences in total): no solo... sino que también(1)

Text Name: ICP5

Text in Spanish: Estudiar español en nuestro instituto **no** es **solo** aprender el idioma, **sino que también** da la oportunidad de conocer y descubrir las diferentes culturas del mundo hispánico.

English: Studying Spanish in our institute not is only learn the language, but also give the opportunity of knowing and discovering the differences cultural of word Hispanic.

Chinese Word (occurrences in total): (不仅仅)...同时[83] (2)

Text Name: ICP5

Text in 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 study Spanish, not only is learn the language itself, but also learn Hispanic world culture.

[83] 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 does not change. Among all the annotated CUs, the phrase *bujinjin...tongshi* (不仅仅...同时) appear once, and the only word *tongshi* (同时) appear once, which equivalents to *bujinjin...tongshi* (不仅仅...同时) under its content.

B.2.18 y &并/以及/还

Spanish Word (occurrences in total): y (34)[84]

Text Name: TERM32

Text in Spanish: Esta ponencia trata del cómo crear un acuñador terminológico automatizado **y** convertirlo en una parte integral del banco terminológico.

English: This paper describes about how to create a coinage terminological automated and turn it into one part integral of bank terminological.

Chinese Word (occurrences in total): 并 (4)

Text Name: TERM32

Text in 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 paper attempts to describe how to create computers support under of terminology builder, and how to turn it into terminology group integral of part.

Spanish Word (occurrences in total): y (34)

Text Name: BCMS3

Text in Spanish: Nuestros programas incluyen cursos generales de español de 60 horas, incluidos cursos VIP (de un alumno), cursos intensivos, regulares **y** de fin de semana.

English: Our programs include courses general of Spanish of 60 hours, including courses VIP (of one student), courses intensive, regular and weekend of week.

Chinese Word (occurrences in total): 以及 (5)

Text Name: BCMS3

Text in Chinese: 我们的课程种类有各个等级的 60 小时的西班牙语课程，形式包括 VIP 课程（一个学生），紧凑课程，常规课程**以及**周末课程。

English: Our courses types has each level of 60 hours of Spanish courses, including VIP course (one student), intensive courses, regular courses and weekend courses.

Spanish Word (occurrences in total): y (34)

Text Name: TERM29

Text in Spanish: ...nos ha obligado a adoptar un determinado enfoque que, os

[84] Totally, there are 35 CUs that include the Saanish word *y* ('and'), however, due to the translation strategy, not all the Chinese translations of *y* are *bing* (并), *yiji* (以及), and *hai* (还). For some cases, there is no translation of the word *y*, the translator uses the coordinated clauses in the Chinese context, because *y* represents a LIST relation under RST.

permita afrontar dichos problemas **y** llevar a cabo el trabajo terminológico y la difusión de sus resultados.

English: us forced to adopt a approach focuses that, us allows to face mentioned problems and carry out the work terminological and the diffusion of their results.

Chinese Word (occurrences in total): 还 (4)[85]

Text Name: TERM29

Text in Chinese: 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，**还**应考虑在传播术语和信息学方面采用不同的方法论。

English: It also forces us in carrying out the specific research, not only in consideration of the above mentioned criteria, and also take account of in diffusion of terminology and information studies adopted different approaches.

B.3 Discourse Structure Examples

Category: N-S

B.3.1 Antithesis

Nuclear: The author favours the idea.

Satellite: The author disfavours the idea.

Text Name: TERM30

Spanish: [Los sistemas de gestión terminológica asequibles hoy en día han aliviado algunas de las labores de almacenamiento y recuperación asociadas con el archivo y presentación de términos especializados.]$_S$ [De todas formas, la recopilación, el análisis y la convalidación son funciones desempeñadas por personas cualificadas.]$_N$

English: [The system of management terminology affordable today have alleviated some of the labor of storage and recovery associated with archiving and presentation of terms specialized.]$_S$ [In any case, the collection, the analysis, and the validation are functions performed by persons qualified.]$_N$

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.]$_N$

B.3.2 Background

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

Satellite: Text for getting the understanding.

---

[85] Among all the annotated Chinese CUs, there are 4 CUs that contain the word *hai* (还). Among the 4 CUs, due to the translation strategy, there is only 1 CU whose parallel Spanish CU include the word *y* ('and') and is the Spanish translation of *hai* (还).

Text Name: TERM23

Spanish: [Con la ayuda del Comité Terminológico para la Lengua irlandesa (An Coiste Téarmaíochta) Fiontar y VOCALL están encauzando las necesidades terminológicas tanto de la educación universitaria en irlandés como de la formación profesional en esta lengua.]$_S$ [Nuestra ponencia estudiará la metodología utilizada por ambos grupos para crear nuevos términos.]$_N$

English: [With the help of the Committee Terminology for the Language Irish (An Coiste Téarmaíochta) Fiontar and VOCALL are channeling the needs terminological of both the education university in Irish and the training professional in this language.]$_S$ [Our paper will study the methodology used by both groups for creating new terms.]$_N$

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.]$_N$

B.3.3 Cause

Nuclear: A situation.

Satellite: Another situation that causes that one.

Text Name: CCICE1

Spanish: [En 2015, por la primera vez, la región de Norteamérica se convierte en el tercer feudo por primas de Mapfre,]$_N$ [desplazando en esa posición a Latam Sur.]$_S$

English: [In 2015, for the first time, the region of North America becomes the third stronghold for premiums of Mapfre,]$_N$ [moving in that position to Latam South.]$_S$

Chinese: [在保险方面，北美已超越南美，]$_S$ [上升为西班牙保险公司 Mapfre 第三大市场。]$_N$

English: [In insurance, North America has surpassed South America,]$_S$ [rising to the Spanish insurance company Mapfre the third largest market.]$_N$

B.3.4 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

Spanish: [Antes de detallar los distintos aspectos que trataremos en la comunicación]$_S$ [definiremos el concepto de topónimo en relación con el término geográfico.]$_N$

English: [Before detailing the distinct aspects that we will discuss in the communication]$_S$ [we will define the concept of place name in relation with the term

geographical.]$_N$

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.]$_N$

B.3.5 Concession

Nuclear: A situation confirmed by the author.

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

Text Name: TERM34

Spanish: [En múltiples ocasiones ha quedado en evidencia que para dar con los equivalentes en euskera de los adjetivos referenciales de las lenguas vecinas se han de seguir varios caminos diferentes (Ensunza, 1989; Loinaz, 1995),]$_N$ [si bien, de alguna manera, la elección de los recursos que se han de emplear se deja en la mayoría de las casos en manos del buen juicio, intuición y el sentido estético del escritor/traductor.]$_S$

English: [On multiple occasions it has been evident that for finding with the equivalent in Basque of the adjectives referential of the languages neighboring followed various paths different (Ensunza, 1989; Loinaz, 1995),]$_N$ [however, in some way, the selection of the resources that are to be used are left in the majority cases in the hands of good judgment, intuition and the sense aesthetic of the writer/translator.]$_S$

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).]$_N$ [However, to some extent, the choice of approach still largely depends on the good judgment, intuition and aesthetic of the author or translator.]$_S$

B.3.6 Condition

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

Satellite: A condition situation.

Text Name: BMCS3

Spanish: [Si optas por un aprendizaje lo más parecido posible a la inmersión, y necesitas mejorar tu nivel de español rápidamente,]$_N$ [los cursos intensivos son una bunea opción.]$_S$

English: [If you choose by learning as possible to immersion, and you need to improve your level of Spanish quickly,]$_N$ [the courses intensive are a good option.]$_S$

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

B.3.7 Elaboration

Nuclear: The basic information.

Satellite: An additional information of the basic information.

Text Name: EEP7

Spanish: [El 18 de septiembre el Embajador de España en la R. P. China inauguró en Pekín la sexta edición del festival "España de Cine", un ciclo cinematográfico que arrancó en 2010 y que, año tras año, permite mostrar una pequeña selección de cine español en China.]N [Desde el viernes 18 hasta el miércoles 23 de septiembre se han proyectado en las salas de cine Lumière Pavilions siete películas producidas recientemente en España y seleccionadas para este ciclo por Ángel Sala, director del prestigioso Festival de Cine Fantástico de Sitges.]S

English: [On 18th of September the Ambassador of Spain in the P. R. China inaugurated in Beijing the sixth edition of the "Spain of Film", a cycle cinematographic that started in 2010 and that, year after year, allows to show a small selection of film Spanish in China.]N [From Friday the 18th to Wednesday the 23rd of September, it has projected in the auditorium of the cinema Lumière Pavilions seven films produced recently in Spain and selected for this cycle by Ángel Sala, director of the prestigious Festival of Film Fantastic of Sitges.]S

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

B.3.8 Enablement

Nuclear: An action.

Satellite: The information aims to perform the action.

Text Name: ICP7

Spanish: [Además de esto, en su espacio personal del MOPAC, el usuario puede hacer una reserva o renovar un préstamo, comunicarse con el responsable de la biblioteca o dejar su comentario en las referencias bibliográficas de los libros que ha leído o de las películas que ha visto,]N [y todo desde su teléfono móvil.]S

English: [In addition to this, in his space personal of MOPAC, the user can make a reservation or renew a loan, communicate with the head of the library or leave his comments in the references bibliographical of the books that he has read or of the movies he has seen,]N [and everything from his cell phone.]S

Chinese: [除此之外，在 MOPAC 的个人空间里，读者用户可以进行预约或者

续借操作，可以和图书馆的负责人取得联系，还可以在图书参考中写下对读过的书或看过的电影的精彩评论，]$_N$ [所有这些，都可以通过您的手机完成。]$_S$

English: [In addition, in MOPAC personal space, users can make the reservation or renew loan, and with library head get contact, and also in the book references bibliographical write down books that he has read or the movies he has seen of comments,]$_N$ [all these, can be realized by cell phone.]$_S$

B.3.9 Evidence

Nuclear: A claim.

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

Text Name: TERM34

Spanish: [De cualquier modo, en numerosas lenguas, existen adjetivos que cubren los tres espacios descritos.]$_N$ [Para ilustrarlo recordaremos el ejemplo de Levi: musical voice, musical criticism, musical comedy.]$_S$

English: [In any case, in many languages, exist adjectives that cover the three spaces described.]$_N$ [To illustrate, we will remember the example of Levi: musical voice, musical criticism, musical comedy.]$_S$

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.]$_S$

B.3.10 Evaluation

Nuclear: A situation.

Satellite: An evaluative comment about this situation.

Text Name: FCEC1

Spanish: [Esta tarea nos corresponde a todos]$_N$ [y es mejor substracto para nuestra actividad exterior.]$_S$

English: [This task belongs to all of us]$_N$ [and it is better subtract for our activity external.]$_S$

Chinese: [这是我们的任务，]$_N$ [是我们对外交往工作的精髓。]$_S$

English: [This is our task,]$_N$ [it is our external works of essence.]$_S$

B.3.11 Interpretation

Nuclear: A situation.

Satellite: An interpretation of the situation.

Text Name: TERM30

Spanish: [De alguna forma, la producción escrita especializada está asociado a la "producción técnica" de términos,]$_N$ [un modelo discursivo que a su vez está

asociado a máqunias.]$_S$

English: [In some way, the production written specialized is associated with the "production technical" of terms,]$_N$ [a model discursive that in its turn is associated with machines.]$_S$

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.]$_S$

B.3.12 Justify

Nuclear: A text.

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

Text Name: TERM30

Spanish: [Pero éste no es el caso de otras lenguas con menor número de hablantes; la gestión terminológica, en estos casos,]$_N$ [a menudo está estrechamente vinculada con la políticamente motivada y, a menudo, emocionalmente cargada planificación lingüística.]$_S$

English: [But this no is the case of other languages,with minor number of speakers; the management terminological, in these cases,]$_N$ [it is often closely linked with the politically motivated and, often emotionally charged planning language.]$_S$

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.]$_S$

B.3.13 Means

Nuclear: An event or an idea.

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

Text Name: FCEC2

Spanish: [Nuestro objetivo es el fomento de un mayor contenido de intercambio bilateral desde una aproximación global y en sus más diversos ámbitos: económico, cultural, educativo, legal, etc.,]$_N$ [contando para ello con programas y actividades propias encaminadas a crear un foco de atención permanente hacia China.]$_S$

English: [Our objective is the promotion of a greater content of exchange bilateral from a approach global and its most diverse areas: economic, culture, education, law, etc.,]$_N$ [counting with its own programs and activities aimed at creating a focus of attention permanent towards China.]$_S$

Chinese: [我们的目标是加强双边在全球化进程中的交流和合作，这一合作涉及各个领域，经济，文化，教育，司法等，]$_N$ [为此我们推出了不同的计划和活动，强调发展对华关系的重要性。]$_S$

English: [Our goal is to strengthen bilateral exchanges global process of communication and cooperation, which involves various fields such as economy, culture, education, law and so on,]$_N$ [to this end, we have carried out different plans and activities, to emphasize to China the development of importance.]$_S$

B.3.14 Motivation

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

Text Name: TERM39

Spanish: [En el caso de este tipo de lenguas minoritarias, los recursos en lengua escrita como bancos léxicos y terminológicos no han sido suficientemente desarrollados como apoyo para el estudiante.]$_N$ [El producto multimedia será idéntico para todas las lenguas del proyecto, y se comercializará como una herramienta de autoaprendizaje para estudiantes de lengua extranjera, así como para estudiantes no aventajados en la L1, dentro de la formación profesional y en las áreas arriba señaladas.]$_S$

English: [In the case of this type of languages minority, the resources in language written as banks lexical and terminological not have been sufficiently developed as support for the students.]$_N$ [The product multimedia will be identified for all the languages of the project, and it will commercialized as a too of self-learning for students of language foreign, as well as for students not advantaged in the L1, within the training professional in the areas above indicated.]$_S$

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.]$_N$

B.3.15 Otherwise[86]

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

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

Text Name: RST webpage example

Spanish: [Los líderes del proyecto deben enviar inmediatamente sus entradas para el folleto revisado.]$_N$ [De lo contrario, se utilizará la entrada existente.]$_S$

English: [The leaders of project must submit immediately their entries for the brochure revised.]$_N$ [Otherwise, it will use the entrance existing.]$_S$

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.]$_S$

B.3.16 Purpose

Nuclear: An intended situation.

Satellite: The intent behind the situation.

Text Name: ICEG2

Spanish: [Visite la Memoria de los Talleres]$_N$ [para conocer todas los talleres que hemos realizado en el Instituto Confucio de la Universidad de Granada.]$_S$

English: [Visit the memory of the workshops,]$_N$ [to know all the workshops that we have realized in the Institute Confucius of the University of Granada.]$_S$

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.]$_S$

B.3.17 Preparation

Nuclear: Content to be presented.

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

Text Name: FICB4

Spanish: [Organiza]$_N$ [Federación de Asociaciones Chinas de Cataluña]$_S$

English: [Organization]$_N$ [Federation of Associations Chinese of Catalonia]$_S$

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]$_S$

B.3.18 Restatement

Nuclear: A situation.

Satellite: A reexpression of the situation.

Text Name: TERM30

Spanish: [habremos de trabajar sobre la modelización de los términos técnicos,]$_N$ [es decir, hemos de reducir las características de los mismos.]$_S$

English: [we will have to work on the modelling of the terms technical,]$_N$ [that is, we have to reduce the characteristics of them.]$_S$

Chinese: [我们也对技术术语进行建模，]$_N$ [即精简这些术语的特性。]$_S$

English: [We also model technical terms,]$_N$ [that is, to simplify the characteristics of these terms.]$_S$

B.3.19 Result

Nuclear: A situation.

Satellite: Another situation which is caused by that one.

Text Name: EEP8

Spanish: [Desde su creación el año 1987, el grupo Capella de Ministrers, bajo la dirección de Carles Magraner, ha desarrollado una importante tarea investigadora y musicológica en favor del patrimonio musical español, desde el medioevo hasta el siglo XIX.]$_N$ [El resultado, transformado en testimonio musical, conjuga a la perfección tres factores clave: el rigor histórico, la sensibilidad musical y, muy especialmente, un incontenible deseo de comunicarnos y hacernos partícipes de estas experiencias.]$_S$

English: [Since its creation the year 1987, the group Capella de Ministrers, under the direction of Carles Magraner, has developed an important task research and musicological in favor of heritage musical Spanish, since the Middle Ages to the century 19th.]$_N$ [The result, transformed into testimony musical, combines perfectly three factors key: the rigor historical, the sensitivity musical and, most especially, an uncontainable desire to communicate and participate of these experiences.]$_S$

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。]$_N$ [这些成果都转变成了音乐的见证， 汇聚成为三个因素的完美结合：历史的严谨、音乐的感性和无法抑制的沟通欲望，正是这些因素给我们以身临其境的感受。]$_S$

English: [Since 1987 founded, by Carles Magraner directed of the group Capella de Ministrers from the Middle Ages to the 19th century for Spanish musical heritage makes many research musicology works.]$_N$ [These achievements all have turned into the music testimony, have become three aspects that perfect combination: the history rigors, the musical sensibilities, and unrestricted of communication desire, all these elements give us immersive feelings.]$_S$

B.3.20 Solutionhood

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

Satellite: A question, request.

Text Name: TERM28

Spanish: [Prescindiendo de la metodología empleada habitualmente en el tratamiento de la terminología,]$_S$ [hemos basado nuestra investigación en tres pilares: los recursos informáticos, la lingüística del corpus y la traductología.]$_N$

English: [Regardless of the methodology used commonly in the treatment of terminology,]$_S$ [we have based our research in three pillars: the resources computational, the linguistic corpus and the translation theory.]$_N$

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.]$_N$

B.3.21 Summary[87]

Nuclear: Text.

Satellite: A short summary of that text.

Text Name: CCICE2

Spanish: [...]$_N$ [A nivel mundial, la inversión inmobiliaria alcanzó los 407.000 millones de dólares en la primera mitad del año (unos 377.646 millones de euros), el mejor semestre desde el anterior pico cíclico en 2007, cuando se sumaron 441.000 millones de dólares (cerca de 409.193 millones de euros).]$_S$

English: [...]$_N$ [To level world, the investment real estate reached 407, 000 million dollars in the first half of the year (about 377, 646 millions euros), the best semester since the previous cycle peak in 2007, when they added 441, 000 million dollars (about 409, 193 million euros).]$_S$

Chinese: [...]$_N$ [根据报告，今年上半年全球房地产投资总额高达 4070 亿美元（折合约 3776.46 亿欧元），为 2007 年以来表现最好的半年，当时为 4410 亿美元（约 4091.93 亿欧元）。]$_S$

English: [...]$_N$ [According to the report, this year in the first half global real estate investment totaled 407 billion U.S dollars (377.646 billion euros), since 2007 the highest in six months, compared to 441 billion U.S dollars (409.193 billion euros).]$_S$

[87] To detect the SUMMARY relation, we give the screenshot of the annotated texts and the English literature translation of the Satellite, as the content of the Satellite summarizes all the text.

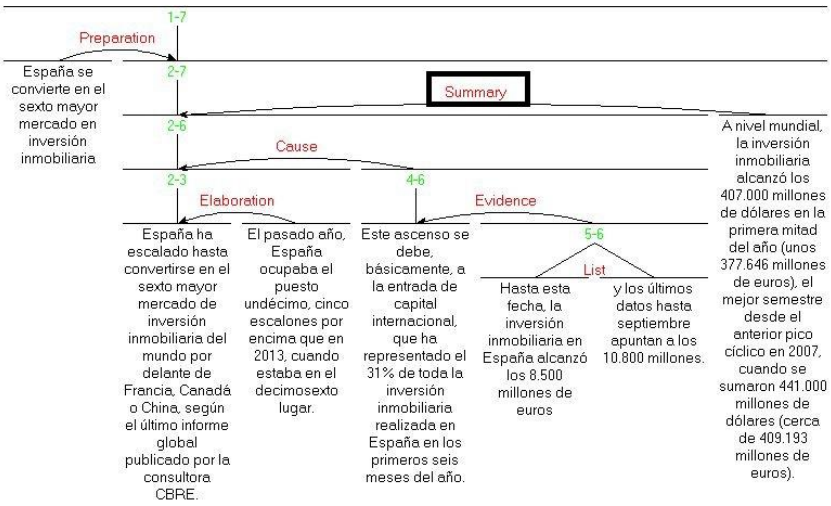

Figure B.3.21 Example of SUMMARY relation in the corpus (Spanish text)

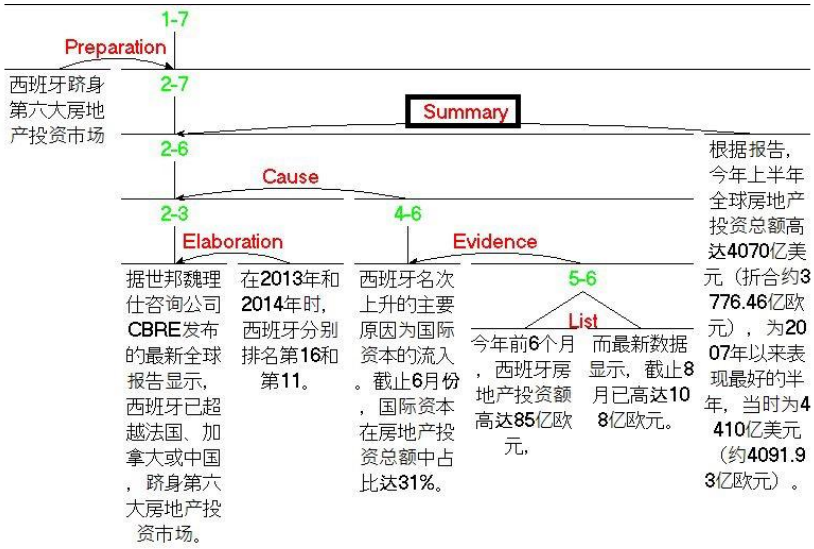

Figure B.3.22 Example of SUMMARY relation in the corpus (Chinese text)

Category: N-N

B.3.22 Conjunction

Nuclear: A situation or an action.

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

Text Name: ICP4

Spanish: [Las convocatorias de plazas de profesor titular para los Institutos Cervantes aparecen publicadas en la prensa española.]$_N$ [También aparecen todas las convocatorias en la página web del Instituto Cervantes www.cervantes.es.]$_S$

English: [The calls for the position of professor tenured for the Institute Cervantes appear published in the press Spanish.]$_N$ [Also appear all the announcements in the page web of Institute Cervantes www.cervantes.es.]$_N$

Chinese: [塞万提斯学院正式教师职位招聘在西班牙媒体上公布。]$_N$ [同时也

在塞万提斯学院网站发布信息。]N

English: [Cervantes Institute official teacher position in the Spanish press announced.]N [Also in Cervantes Institute web page announced.]N

B.3.23 Contrast

Nuclear: One option.
Nuclear: The other option.

Text Name: CCICE1

Spanish: [Los resultados a septiembre constatan que Mapfre generó primas por valor de 2.103,1 millones de euros en el área de Norteamérica, frente a los 1.822,5 millones de Latam Sur.]N [Hace un año, la situación era inversa: la facturación de Latam Sur (2.095,2 millones) excedía en más de 500 millones a la de Norteamérica (1.573,4 millones).]N

English: [The results of September show that Mapfre generated premiums worth of 2,103.1 million euros in the area of Northamerica, compared to 1,822.5 million in Latam Sur.]N [A year ago, the situation was reversed: the turnover of Latam Sur (2,095.2 million) exceeded by more than 500 million to that of North America (1,573.4 million).]N

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

B.3.24 Disjunction

Nuclear: An alternative.
Nuclear: Another alternative.

Text Name: ICP4

Spanish: [Las solicitudes para trabajar como profesor colaborador en el Instituto Cervantes de Pekín pueden enviarse a la Jefa de Estudios mediante correo electrónico a la dirección acpek@cervantes.es]N [o entregarse en el propio correo: ...]N

English: [Applications for position as professor assistant in the Institute Cervantes of Beijing can send to the Manage of Research by email to acpek@cervantes.es]N [or deliver in the itself mail: ...]N

Chinese: [申请北京塞万提斯学院助理教师工作可以发送申请至教学负责人，邮箱为 acpek@cervantes.es]N [或直接递交到北京塞万提斯学院：...]N

English: [Applying for Beijing Cervantes Institute assistant professor position can be send to teaching manager, the email is acpek@cervantes.es]N [or directly to Beijing Cervantes Institute: ...]N

B.3.25 List
Nuclear: An item.
Nuclear: A next item.

Text Name: EEP3

Spanish: [El programa de su visita a China incluyó visitas al Observatorio Astronómico Nacional de la Academia China de las Ciencias, a la Adeministración Estatal de Oceanografía y a la China Certification and Inspection Group,]$_N$ [y en las que se abordaron en detalle temas en materia oceanográfica, astrofísica y de energías renovables respectivamente.]$_N$

English: [The program of her visit to China included visits to the Observatory Astronomical National of the Academy China of the Sciences, to the Administration State of Oceanography and to the China Certification and Inspection Group,]$_N$ [and in which addressed in detail topics in material oceanographic, astrophysical and energy renewable respectively.]$_N$

Chinese: [此次访华行程还包括参观中科院国家天文台，国家海洋局和中国检验认证集团，]$_N$ [并分别就海洋，宇航以及可再生能源材料的问题进行了讨论。]$_N$

English: [This visit to China included visits to the Observatory Astronomical National of the Academy China of the Sciences, to the Administration State of Oceanography and to the China Certification and Inspection Group,]$_N$ [and issued with material oceanographic, astrophysical and energy renewable discussed.]$_N$

B.3.26 Sequence
Nuclear: An item.
Nuclear: A next item.

Text Name: FICB2

Spanish: [La sesión de la tarde contó con las participantes del Sr. Joaquín Mataró, director de La Salle Girona, quien presentó la experiencia del proyecto “Lengua y cultura chinas” en La Salle Girona.]$_N$ [La profesora de chino de la Salle Girona, la Sra. Hong Haifeng, compartió su experiencia y métodos docentes para jóvenes y adolescentes.]$_N$

English: [The section of the afternoon tells with the participants of Mr. Joaquín Mataró, the head of the La Salle Girona, who presented the experience of the project “Language and culture Chinese” in La Salle Girona.]$_N$ [The professor of Chinese of La Salle Girona, Ms. HongHaifeng, shared her experience and methods teaching for young people and adults.]$_N$

Chinese: [下午由拉萨耶教学集团的赫罗纳中学的校长谈论了其学校教授和学习汉语的经验，]$_N$ [来自其学校的本土老师洪海峰当场示范了如何在当地开展针对儿童的汉语教学。]$_N$

English: [Afternoon by La Salle Group of Girona Secondary School of the head talked about teaching and learning Chinese of experience,]$_N$ [from his school of native teacher HongHaifeng in present demonstrated how to locally carry out to children of Chinese teaching.]$_N$

Appendix C
Special Discourse Comparison Cases

In this part, we will show the cases that cannot make the recommendations for Spanish-Chinese translation.

Recommendation C.1

Text Name: EEP4

Spanish: [La iniciativa establece nexos para unir, por un lado, China, Asia Central, Rusia y Europa, y, por otro, China con el Mediterráneo a través del Golfo Pérsico.]$_{N\_Elaboration}$ [Para España, la iniciativa tiene un gran interés para trasladar un mensaje de compromiso de nuestro país con el proyecto de la Nueva Ruta de la Seda y, en general, con el fortalecimiento de nuestras relaciones bilaterales con China.]$_{S\_Elaboration}$

English: [The initiative establishes links to unite, on one hand, China, Central Asia, Russia and Europe, and on the other, China with the Mediterranean through the Gulf Persian.] [For Spain, the initiative has a great interest to convey a message of commitment of our country with the project of the New Silk Road and, in general, with the strengthening of our relations bilateral with China.]

Chinese: [“新丝绸之路”的倡议将为中国、中亚、俄罗斯和欧洲，以及中国通过波斯湾和地中海国家建立联系纽带。]$_{N\_Result}$ [该倡议给中西两国带来进一步巩固双边关系的新领域。]$_{S\_Result}$

English: [The “New Silk Road” initiative will for China, Central Asia, Russia and Europe, as well as China through Persian Gulf and the Mediterranean countries establish connections.] [The initiative bring China and Spain two countries brings bilateral relations of new area.]

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	/		Chinese	Result
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-S		Chinese	N-S
Recommendation for translation: None					
Note: The annotation of discourse relation for Spanish and Chinese passage is different. In the Spanish passage, the first EDU introduces the basic information of the New Silk Road, and the second EDU gives more additional information about the establishment of the New Silk Road, especially the influence for Spain and China. Therefore, the annotators define the ELABORATION relation for the Spanish case. In the Chinese passage, the second EDU declares the result of the establishment of the New Silk Road for Spain and China, the description in the Chinese passage is different from the Spanish passage.					

Recommendation C.2

Text Name: FICB2

Spanish: [Como conclusión de la formación, los asistentes compartieron dudas y experiencias.]$_{N\_Elaboration}$ [Todos los asistentes recibieron los certificados de participación de Hanban y de la FICB.]$_{S\_Elaboration}$

English: [As the conclusion of the training, the assistants shared doubts and experiences.] [All the attendees received the certificate of the participation of Hanban and the FICB.]

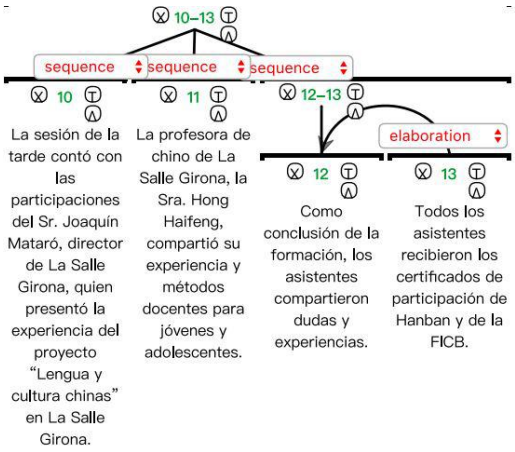

Figure C.2.1 Partly annotation of the text FICB2 (Spanish part)

Chinese: [...之后进行了圆桌会议的讨论，全体与会教师就汉字书写问题等进行了讨论，并就海外汉语教学中的疑惑和经验展开了深入的交流。]$_{N\_Sequence}$ [培训结束之后，我院为参加本次培训的每位教师颁发了汉办制作和巴塞罗那孔子学院制作的教学培训证书。]$_{N\_Sequence}$

English: [After that, held the roundtable discussion, all the participating teachers Chinese characters writing and other problems discussed, and oversea Chinese teaching process of doubts and experiences further communication.] [After the training, our institute awarded each teacher with Hanban and the FICB made certificate.]

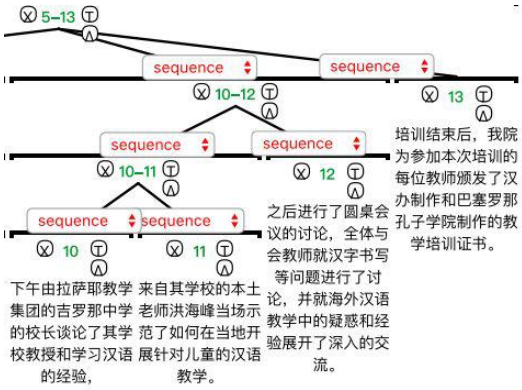

Figure C.2.2 Partly annotation of the text FICB2 (Chinese part)

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	Elaboration
	Chinese	/		Chinese	Sequence
Relation type	Spanish	N-S	**EDUs order**	Spanish	N-S
	Chinese	N-N		Chinese	N-N
Recommendation for translation: None					
Note: In this case, we can see that in the Spanish passage, the EDU12 and the EDU13 hold an ELABORATION relation, the two EDUs are at the same discourse level. Anyhow, in the Chinese passage, the EDU12 and the EDU13 are not at the same discourse level, the EDU12 and the EDU(10-11) from a SEQUENCE relation, then the EDU(10-12) and the EDU13 hold a SEQUENCE relation. Although the different languages cause the different annotations, the main idea in the passages is the same, which introduce the activities during the round table discussion and gives the certificates to the participants after of the training.					

Recommendation C.3

Text Name: ICP5

Spanish: [Estudiar español en nuestro instituto no es solo aprender el idioma,]$_{N\_List}$ [sino que también da la oportunidad de conocer y descubrir las diferentes culturas del mundo hispánico.]$_{N\_List}$

English: [Studying Spanish in our institute is not only learning the language,] [but also gives the opportunity of knowing and discovering the differences cultural of the world Hispanic.]

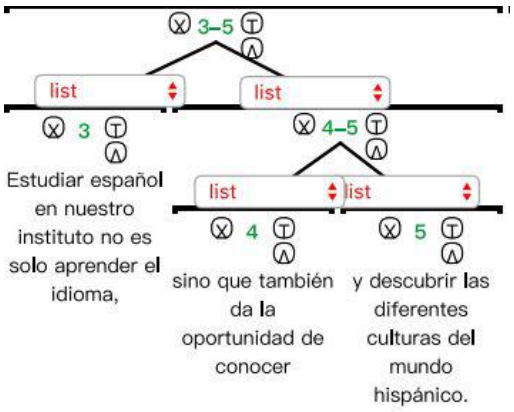

Figure C.3.1 Partly annotation of the text ICP5 (Spanish part)

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。]$_{N\_Summary}$ [给予你一个了解和发掘西班牙语世界不同文化的机会。]$_{S\_Summary}$

English: [In our institute study Spanish, is not only about learning the language itself, it is also about learning Spanish-speaking culture.] [Giving you an knowing and exploring Hispanic world different cultures opportunity.]

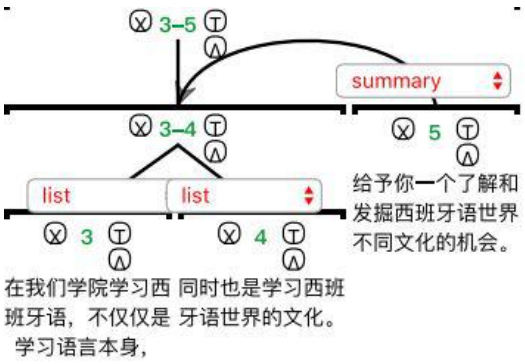

Figure C.3.2 Partly annotation of the text ICP5 (Chinese part)

Discourse markers (DM)	Spanish	/	**Relation**	Spanish	List
	Chinese	/		Chinese	List & Summary
Relation type	Spanish	N-N	**EDUs order**	Spanish	N-N
	Chinese	N-N		Chinese	N-N
Recommendation for translation: None					
Note: From the screen shoots we can see that, in the Spanish passage, the EDU3, the EDU4 and the EDU5 are three parts of a complete sentence. The EDU3 and the EDU(4-5) are at the same discourse level and hold a LIST relation, meanwhile there is another LIST relation within the EDU4 and EDU5. However, in the Chinese passage, the EDU(3-4) and the EDU5 are at the same discourse level and hold a SUMMARY relation, the EDU3 and the EDU4 are two parts of a complete sentence and there is a LIST relation between them. The different way of expression causes the annotation difference between the parallel passages.					

Appendix D
List of Annotated Texts Links

This part gives the link of each text in the corpus.

Text name	Link
BMCS1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_CHN1-GS.txt
BMCS2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_CHN2-GS.txt
BMCS3	http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_ESP3-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_CHN3-GS.txt
BMCS4	http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_ESP4-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_CHN4-GS.txt
BMCS5	http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_ESP5-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/BMCS_CHN5-GS.txt

Table D.1 Text annotation links of the BMCS part

Text name	Link
CCICE1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_CHN1-GS.rs3
CCICE2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_CHN2-GS.rs3
CCICE3	http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_ESP3-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_CHN3-GS.rs3
CCICE4	http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_ESP4-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_CHN4-GS.rs3
CCICE5	http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_ESP5-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/CCICE_CHN5-GS.rs3

Table D.2 Text annotation links of the CCICE part

Text name	Link
EEP1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN1-GS.rs3
EEP2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN2-GS.rs3
EEP3	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP3-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN3-GS.rs3
EEP4	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP4-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN4-GS.rs3
EEP5	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP5-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN5-GS.rs3
EEP6	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP6-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN6-GS.rs3
EEP7	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP7-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN7-GS.rs3
EEP8	http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_ESP8-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/EEP_CHN8-GS.rs3

Table D.3 Text annotation links of the EEP part

Text name	Link
FICB1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_CHN1-GS.rs3
FICB2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_CHN2-GS.rs3
FICB3	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_ESP3-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_CHN3-GS.rs3
FICB4	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_ESP4-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_CHN4-GS.rs3
FICB5	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_ESP5-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FICB_CHN5-GS.rs3

Table D.4 Text annotation links of the FICB part

Text name	Link
FCEC1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FCEC_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FCEC_CHN1-GS.rs3
FCEC2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/FCEC_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/FCEC_CHN2-GS.rs3

Table D.5 Text annotation links of the FCEC part

Text name	Link
ICP1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN1-GS.rs3
ICP2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN2-GS.rs3
ICP3	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP3-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN3-GS.rs3
ICP4	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP4-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN4-GS.rs3
ICP5	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP5-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN5-GS.rs3
ICP6	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP6-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN6-GS.rs3
ICP7	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP7-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN7-GS.rs3
ICP8	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_ESP8-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICP_CHN8-GS.rs3

Table D.6 Text annotation links of the ICP part

Text name	Link
ICEG1	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICEG_ESP1-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICEG_CHN1-GS.rs3
ICEG2	http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICEG_ESP2-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/ICEG_CHN2-GS.rs3

Table D.7 Text annotation links of the ICEG part

Text name	Link
TERM18	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM18_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM18_CHN-GS.rs3
TERM19	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM19_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM19_CHN-GS.rs3
TERM23	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM23_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM23_CHN-GS.rs3
TERM25	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM25_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM25_CHN-GS.rs3
TERM28	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM28_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM28_CHN-GS.rs3
TERM29	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM29_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM29_CHN-GS.rs3
TERM30	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM30_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM30_CHN-GS.rs3
TERM31	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM31_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM31_CHN-GS.rs3
TERM32	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM32_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM32_CHN-GS.rs3
TERM34	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM34_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM34_CHN-GS.rs3
TERM38	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM38_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM38_CHN-GS.rs3
TERM39	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM39_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM39_CHN-GS.rs3
TERM40	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM40_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM40_CHN-GS.rs3
TERM50	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM50_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM50_CHN-GS.rs3
TERM51	http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM51_ESP-GS.rs3 http://ixa2.si.ehu.es/rst/zh/zh.RS3/TERM51_CHN-GS.rs3

Table D.8 Text annotation links of the TERM part

Appendix E
Encoding of the Task for Spanish-Chinese Language Learning

This part gives the encoding of the task for Spanish-Chinese Language Learning

Script 1 Generation Programming for Chinese texts

```
import random
import sys
ls = []
lista = open('dm_list.txt')
for line in lista:
    ls.append(line.rstrip('\n'))
# Create a list with the names of the files to create the exercises.
filenames = []
texts = open(sys.argv[1])
for line in texts:
    filenames.append(line.rstrip('\n'))
# Create a new file for each exercise and describe it.
for file in filenames:
    exercise = open('Exercises/'+file, 'w+')
    exercise.write('\n##############################################################################\n')
    exercise.write('Exercise: Fill in the blanks, with one of the options after the text.\n')
    exercise.write('##############################################################################\n\n')
    # Remove the DMs from the list and append them to a list, in order to create the
    # list with the options.
    options=[]
    with open('Texts/'+file) as fh:
        for line in fh:
            for i in ls:
                if i in line:
                options.append(i)
                    line = line.rstrip()
                    newline = line.replace(i, "____")
                    line=newline
            exercise.write(line)
    # Shuffle the options list and write in the file.
    exercise.write('\n\nOptions:\n')
    random.shuffle(options)
```

```
    for opt in options:
        exercise.write('- '+opt+'\n')
```

Script 2 Generation Programming for Spanish texts

```
import os
import random
import json
import sys
reload(sys)
sys.setdefaultencoding('utf8')
# Create a list with the Discourse Markers we want to remove from the text and
# sort them according to the number of words of the Discourse Marker, highest to
# lowest.
file_list=open('dm_list.json')
ls0=json.load(file_list)
file_list.close()
ls=[mem.split() for mem in ls0]
ls.sort(lambda x,y: cmp(len(y), len(x)))
# Load the information about DMs.
file_info=open('dm_info.json')
dm_info=json.load(file_info)
file_info.close()
# Create a list with the files we want to use to create the exercises.
filenames = []
list_texts = open(sys.argv[1])
for line in list_texts:
    filenames+=set((line.rstrip('\n')).split())
# Load more information about DMs (second parts of DMs which have two parts)
file2=open('2parts.json')
secondparts=json.load(file2)
dirs = set(os.listdir('Texts'))-set(['.DS_Store'])
for dir in dirs:
    for file in os.listdir('Texts/'+dir):
        if file not in filenames:
            continue
        text = []
        text2 = []
        # For each file, create a new one containing the exercise and another one
        # with the answers.
        newfile = open('exercises/'+file,'w+')
        newfile.write('\n############################################################'
                      '###############\n')
        newfile.write('Exercise: You have to choose one of the three options to fill '
```

```
            'in the blank.\n')
newfile.write('Write your answer (a, b or c) after "YOUR ANSWER:"\n')
newfile.write('########################################################################'
            '#############\n\n')
answers = open('answers/' + file, 'w+')
l=[]
# Save the text as a list of words and punctuation marks
# (punctuation marks are needed to analyse the context of the MDs)
with open('Texts/'+dir+'/'+file) as fh:
    num = 1
    text0=[]
    for line in fh:
        text0.append(line.split())
    for i in range(len(text0)):
        text.append('#')
        for word in text0[i]:
            text.append(word)
    for word in text:
        if word[0] in ['¡', '(', '"', '¿'] and word[-1] in\
                [',', ';', '.', ':', '?', ')', '!', '"']:
            text2 += [word[0], word[1:-1], word[-1]]
        elif word[0] in ['¡', '(', '"', '¿']:
            text2 += [word[0], word[1:]]
        elif word[-1] in [',', ';', '.', ':', '?', ')', '!', '"']:
            text2 += [word[0:-1], word[-1]]
        else:
            text2.append(word)
    # If a word is a MD from the list and fulfills the conditions, replace it
    # by a blank with a number to indicate the options after the text.
    for j in range(len(text2))[1:]:
        for md in ls:
            md_upper = [word.upper() for word in md]
            md_up_low = [md[0][0].upper() + md[0][1:]] + md[1:]
            if len(md) >= len(text2) - j:
                break
            if text2[j:j+len(md)] == md or text2[j:j+len(md)] == md_upper \
                    or text2[j:j+len(md)] == md_up_low:
                if dm_info[' '.join(md)]=="always":
                    l.append([text2[j - 1], md, text2[j + len(md)]])
                    text2[j] = "___["+str(num)+"]___"
                    num += 1
                    for k in range(len(md) - 1):
                        del text2[j + k + 1]
                    break
```

```
            # If the MD info says that it has to be removed only at the
            # beginning of a sentence, analyse the context.
            elif dm_info[' '.join(md)] == "BOS" and (text2[j-1] == '.' \
                    or text2[j-1]=='#'):
                l.append([text2[j - 1], md, text2[j + len(md)]])
                text2[j] = "___[" + str(num) + "]___"
                num += 1
                for k in range(len(md) - 1):
                    del text2[j + k + 1]
                break
            # If the MD info says that it has two parts, remove both of them
            # and put the same number in both blanks, because the question is
            # the same.
            elif dm_info[' '.join(md)]=="2p":
                second_comp=False
                k=j+len(md)
                while (text2[k]!="." and second_comp==False):
                    for secondpart in \
                        secondparts[unicode(' '.join(text2[j:j+len(md)])).lower()]:
                        if text2[k:k+len(secondpart)]==secondpart:
                            second_comp=True
                            break
                    k+=1
                if second_comp==True:
                    l.append([text2[j - 1], [str(md[0])+' / '+str(secondpart[0])],
                            text2[j + len(md)]])
                    text2[j] = "___[" + str(num) + "]___"
                    text2[k-1] = "___[" + str(num) + "]___"
                    num += 1
                    for r in range(len(secondpart)-1):
                        del text2[k+r+1]
                    for r in range(len(md)-1):
                        del text2[j+r+1]
                    break
    # Write the text with the blanks replacing the DMs.
    for i in range(len(text2)):
        if text2[i] in ['¡', '(', "", '¿']:
            text2[i + 1] = text2[i] + text2[i + 1]
        elif text2[i] in [',', ';', '.', ':', '?', ')', '!', ""]:
            text2[i - 1] = text2[i - 1] + text2[i]
    text3 = [word for word in text2 if word not in \
            ['¡', '(', "", '¿'] + [',', ';', '.', ':', '?', ')', '!', ""]]
    text3 = text3[1:]
    for word in text3:
```

```
        if word == '#':
            newfile.write('\n')
        else:
            newfile.write(word+' ')
# Select three options for each blank depending on the information given about
# the distractors.
newfile.write('\n\n########\n')
newfile.write('Options:\n')
newfile.write('########\n\n')
distfile = open('distr_list.json')
distractors=json.load(distfile)
distfile.close()
i = 1
for md in l:
    if (md[0]==[","] or md[0]==[";"]) and md[2]!=["."] and md[2]!=[","] and md[2]!=[";"]:
        group = 1
    elif (md[0]==["."] or md[0]==["#"]) and md[2]==[","]:
        group = 2
    elif (md[0]==["."] or md[0]==["#"]) and md[2]!=[","]:
        group = 3
    elif md[0]!=["."] and md[0]!=[","] and md[0]!=[";"] and md[2]!=["."] and \
                md[2]!=[","] and md[2]!=[";"]:
        group = 4
    elif md[0] == [","] and md[2] == [","]:
        group = 5
    if md[1]==["y"] or md[1]==["o"] or md[1]==["e"]:
        group = 6
    if md[1]==["tanto / como"] or md[1]==["ni / ni"] or md[1]==["no sólo / sino"] or \
                md[1]==["no solamente / sino"]:
        group = 7
    distgroup=distractors['Group'+str(group)]
    choosefrom = []
    for group in distgroup:
        if [' '.join(md[1])] not in group:
            choosefrom += group
    choosefrom0=[o[0] for o in choosefrom]
    opt = random.sample(choosefrom0,2)+[' '.join(md[1])]
    opta = random.sample(opt,1)
    opt2=[]
    for option in opt:
        if option not in opta:
            opt2.append(option)
    optb = random.sample(opt2,1)
    opt3=[]
```

```
        for option in opt2:
            if option not in optb:
                opt3.append(option)
        optc = random.sample(opt3,1)
        # Write the possible options in the file, and a section to put the
        # student's answer
        newfile.write(str(i)+'. a) '+str(opta[0])+' b) '+str(optb[0])\
                    +' c) '+str(optc[0])+'\n')
        newfile.write("YOUR ANSWER: \n\n")
        i += 1
        if md[1]==opta:
            answers.write('a\n')
        elif md[1]==optb:
            answers.write('b\n')
        else:
            answers.write('c\n')
    newfile.close()
```

Script 3 Grade Programming for Spanish texts

```
import sys
# Open the files containing the student's answers and the correct answers.
exercise = open('exercises/'+sys.argv[1]).readlines()
correct = open('answers/'+sys.argv[1]).readlines()

# Create a list containing the student's answers.
answers=[]
for line in exercise:
    if line.startswith('YOUR ANSWER:'):
        if (line.rstrip('\n')).rstrip()== 'YOUR ANSWER:':
            answers.append(0)
            continue
        ans=(line.rstrip('\n')).split(':')[1]
        while ans[0]==' ':
            ans=ans[1:]
        answers.append(ans)
# Create a list containing the correct answers.
correct_answers=[]
for line in correct:
    correct_answers.append(line.rstrip('\n'))
# Compare the correct answers with the student's answers. Depending
# on them, generate a different feedback and calculate the grade.
grade=0
feedb=[]
```

```
for i in range(len(answers)):
    if answers[i] == 0:
        feedb.append('Expected answer: '+str(correct_answers[i]))
        continue
    if answers[i] == correct_answers[i]:
        feedb.append('Correct!')
        grade += 1
    elif answers[i] != correct_answers[i]:
        feedb.append('Incorrect... Expected answer: '+str(correct_answers[i]))
        grade += -0.3
# Calculate the final grade over 10 points.
final_grade = round(10*grade/len(answers),2)
with open('exercises/'+sys.argv[1]) as file:
    exercise=file.readlines()
# Write in the document the feedbacks generated before.
j=0
for i in range(len(exercise)):
    if exercise[i][0:12]=='YOUR ANSWER:':
        exercise[i]+=feedb[j]+'\n'
        j+=1
# Write the final grade at the end of the document.
exercise.append('\n###########\nFinal grade: '+str(final_grade)+'/10\n###########')
with open('exercises/'+sys.argv[1],'w') as file:
    file.writelines(exercise)
```