How does discourse affect Spanish-Chinese Translation? A case study based on a Spanish-Chinese parallel corpus

Shuyuan Cao
Grupo COLE, Departamento de Informática
Universidade de Vigo
Campus As Lagoas, Ourense 32004, Spain
shuyuan.cao@uvigo.es

Abstract

With their huge speaking populations in the world, Spanish and Chinese occupy important positions in linguistic studies. Since the two languages come from different language systems, the translation between Spanish and Chinese is complicated. A comparative study for the language pair can discover the discourse differences between Spanish and Chinese, and can benefit the Spanish-Chinese translation. In this work, based on a Spanish-Chinese parallel corpus annotated with discourse information, we compare the annotation results between the language pair and analyze how discourse affects Spanish-Chinese translation. The research results in our study can help human translators who work with the language pair.

1 Introduction

From the early history, people began to apply Natural Language Processing (NLP) techniques to language researches (Burstein, 2009). Different NLP studies make a great advance in different language aspects, such as pragmatics, semantics, speech, etc.

Among different NLP studies, the emphasis on the idea that discourse information may be useful for Natural Language Processing (NLP) has become increasingly popular. Discourse analysis is an unsolved problem in this field, although discourse information is crucial for many NLP tasks (Zhou et al., 2014). Plus, the greater the linguistic distance is between a pair of languages, the greater the number of differences in their syntax and discourse structure. Therefore, the translation between two very different languages can be potentially more difficult. Comparative or contrastive studies of discourse structures offer clues to identify properly equivalent discourse elements in two languages. These clues can be useful for human translation. The following examples show some of the discourse differences between Spanish and Chinese.

Example 1[1]:
(1.1) *Spanish*: [**Aunque** aún no contamos con resultados,]Unit$_1$ [intuimos que el modelo será más amplio que el del sintagma nominal.]Unit$_2$
[DM[2] still not 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,]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$_1$ [DM still not get results.]Unit$_2$
(1.3) *Chinese*: [**尽管**还没有取得最终结果，]Unit$_1$ [**但是**我们认为该模型已囊括了语段模型涉及的内容。]Unit$_2$
[DM1 still no get results,]Unit$_1$ [DM2 we consider that the model contains the sentence group nominal.]Unit$_2$
(1.4) *English*: Although we haven’t got the results yet, we consider that the model will be more extensive than the nominal sentence group.

In Example 1, we can see that the Spanish passage (1.1) and the Chinese passage (1.3) have a similar discourse structure. Both passages start with a discourse marker in the first unit. However, the DMs are used differently to show the same meaning in both languages. In Chinese, it is mandatory to include two DMs: the first one is *jinguan* (尽管), and it is located at the beginning of the first unit, and the other marker is *danshi* (但是), which is placed at the beginning of the second unit.

[1] We give an English literature translation for each example in this work.
[2] DM means discourse marker. We will give the specific definition of discourse marker in the methodology section.

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 1–10
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

These two discourse markers (DMs) are equivalent to the English DM 'although'. By contrast, in Spanish, just one DM *aunque* is needed to express the same meaning. Besides, as we can see in another Spanish passage (1.2), the order of the discourse units in can be changed and it makes sense syntactically, so the DM can appear both at the beginning of the first or the second unit. By contrast, the order cannot be changed in the Chinese passage, because neither syntactically nor grammatically makes sense.

Due to the certain considerable discourse differences between the two languages, it is essential to carry out a discourse comparative study for Spanish and Chinese. Therefore, based on a Spanish-Chinese parallel corpus, this work aims to give a discourse analysis with the intention to analyze how Spanish-Chinese translation can be affected from discourse level. This analysis can be beneficial for human translators who work with the language pair.

In the second section, we present the theoretical framework of this study. In the third section, we talk about some related works. In the fourth section, we give detailed information on the methodology of this work. In the fifth section, we evaluate the research results and give a qualitative analysis. In the last section, we conclude our work and look ahead at future work.

2 Theoretical Framework

The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a theory that was created especially for discourse analysis. It 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). Intra-sentence and inter-sentence annotation styles help to inform how discourse elements are being expressed in a language, and translation strategies (if any) can be detected in different levels of an RS-tree (da Cunha and Iruskieta, 2010; Iruskieta et al., 2015).

RST addresses both hierarchical and relational aspects of text structures for discourse analysis. Elementary Discourse Units (EDUs) (Marcu, 2000) and coherence relations are established in RST. Relations are recursive in RST and are held between EDUs, which can be Nuclei or Satellites, denoted by N and S. Satellites offer additional information about nuclei. EDUs can be linked among them holding a nucleus-satellite (e.g. Cause, Justify, Evidence) function or a multinuclear (e.g. Conjunction, List, Sequence) function. As relations are recursive, all the discourse units of the text have a function in a treelike structure, if and only if the text is coherent.

3 State of the Art

Some previous researches using RST for comparative discourse are, for instance, Chinese and English (Ramsay, 2000, 2001), Japanese and Spanish (Kumpf, 1986; Marcu et al., 2000), Arabic and English (Mohamed and Omer, 1999), French and English (Delin et al., 1994; Salkie and Oates, 1999), Dutch and English (Abelen et al., 1993), Spanish and Basque (da Cunha and Iruskieta, 2010; Imaz and Iruskieta, 2017).

RST contrastive studies that use more than two languages are not common; those that have included work on Portuguese-French-English (Salkie and Oates, 1999) and Basque-English-Spanish (Iruskieta et al., 2015).

Currently, only three works use RST for Spanish-Chinese discourse analysis. One work is from (Cao et al., 2016). They explore sentences that contain the Spanish discourse marker *aunque* ('although' in English) and their Chinese parallel sentences in the UN subcorpus. Another work is the creation of the language teaching and learning resources for Spanish-Chinese by (Cao and Gete, 2018). In their work, they create a system that gives tests to check the Spanish-Chinese students language level through erased DMs in texts. But, they only analyze the single sentences in the corpus, not the whole discourse structure of each text in the corpus. The last work talks about the Spanish-Chinese discourse analysis taking RST as framework is the the RST Spanish-Chinese Treebank (Cao et al., 2018). Cao et al. (2018) establish the first Spanish-Chinese discourse treebank with annotated discourse information under RST. Although the treebank can be useful for different NLP tasks, Cao et al. (2018) only create the treebank without any practical use.

To our knowledge, our work is the first one that analyzes the discourse structures of a whole text for both Spanish and Chinese and apply the analysis results to a language translation task.

4 Methodology

In this section, we present the methodology of this study. In the first subsection, we introduce the research corpus. In the second subsection, we explain how we carry out our analysis.

4.1 Corpus

In this research, we use the RST Spanish-Chinese Treebank created by Cao et al. (2018), a corpus annotated with discourse information under RST. As Cao et al. (2018) indicate, The RST Spanish-Chinese Treebank is the first Spanish-Chinese parallel corpus that guarantees the discourse structure diversity for the language pair. In their corpus, the texts are from different sources. The genres and topics of the corpus are different. Totally, 50 Spanish texts and their translated Chinese texts are selected.

Concerning the corpus annotation, Cao et al. (2018) make three annotation steps. All the annotations are completed by linguists with RST annotation training. As the initial step, they segment the corpus based on the elaborated criteria. After the segmentation work, authors annotate the discourse structure of the whole corpus following the method proposed by Pardo (2005).

Towards the annotation quality of the corpus, Cao et al. (2018) use Kappa to measure the segmentation part. For the evaluation of discourse structure annotation, they use a qualitative method (Iruskieta et al., 2015). Under the qualitative method, four elements are being examined by using F-measure: Nuclearity(N), Relation(R), Composition(C) and Attachment(A).

The K results of the segmentation annotation in the Spanish subcorpus is from 0.716 to 0.945 while the results of the Chinese subcorpus is from 0.616 to 0.815. The F results of the discourse structure annotation in the Spanish subcorpus are: N (from 0.761 to 1), R (from 0.641 to 1), C (from 0.761 to 0.933) and A (from 0.731 to 0.933). For the Chinese subcorpus discourse structure annotation, the F results are: N (from 0.864 to 0.978), R (from 0.727 to 0.844), C (from 0.864 to 0.978) and A (from 0.84 to 0.978).

The full annotation of the RST Spanish-Chinese Treebank can be find at ixa2.si.ehu.es/rst/zh/index.php. It is a free open access to the research community and all the corpus texts and annotations can be downloaded for research purposes. Moreover, in their corpus, authors give the part-of-speech (pos) information for each text.

The evaluation results for each annotation step show that the corpus is annotated with high quality. Based on the the reliable annotation results, we decide to use the RST Spanish-Chinese as the research corpus.

4.2 Discourse differences in translation strategies

In the study of Iruskieta et al. (2015), besides of creating the qualitative method for the discourse annotation evaluation, they also find how discourse elements affect language translation and the causes are defined as translation strategies.

- Marker change. Marker change means for the parallel passages, the DMs in both texts are different.

- Clause structure change. During the translation process, a non-finite verb phrase is changed to finite verb structure.

- Unit shift. In the parallel passages, the punctuations are different.

Therefore, we follow the method of Iruskieta et al. (2015) to detect the possible translation strategies which can affect translation Spanish and Chinese from discourse level.

Additionally, for the marker change case, we confirm the definition of DM in our analysis. One of the DM definitions that address RST is from Eckle-Kohler et al. (2015). They consider that, from textual level, DMs are used to signal discourse relations in a text segment, as cohesive relationships between the utterances. Specifically, under RST, da Cunha (2013) proposes three types of DMs: (i) Traditional markers; (ii) Markers including lexical units; and (iii) Markers including verbal structures. Adopting the definitions from the two works, we use the concept of traditional markers and markers including verbal structures. Both types of markers signal a discourse relation in a segmented text.

5 Evaluation and Analysis

In this section we analyze the results based on the discourse differences. Based on the annotation of each text in the corpus, we compare all the parallel passages and find the following differences discourse differences in the corpus:

- Marker change. Marker change means for the parallel passages, the DMs in both texts are different.

- Unit shift. In the parallel passages, the punctuations in the original passage and the translated passage are different.

- Unit shift plus marker change. For the parallel passages, the punctuation and the DM in the translated passage are different from the original passage.

- Different order EDUs. Although the Spanish passage and its Chinese parallel passage include the same content, the order of the expressions are different in two languages.

- Added discourse. A new discourse is added in the translated passage and causes the relation change between the parallel parts.

Figure 1 concludes the statistical information of the translation strategies in our study. We can see that, among the 26 cases that we collected from the annotated corpus, marker change is the most frequent translation strategy.

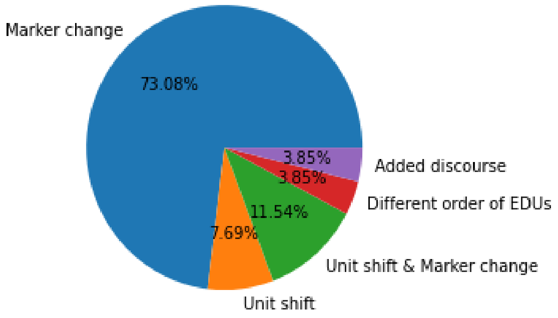


Figure 1: Statistical conclusion of the translation strategies in the corpus

5.1 Marker change

Totally, there are 19 cases related to marker change. There are 6 cases that the DMs in the Spanish passages are changed in their Chinese parallel passages. For the other 13 collected cases, the Spanish passages don't contain any DM. In contrast, there is a DM in each of their Chinese parallel passages. Table 1 sums up the cases of the change of the DMs. Meanwhile, Table 2 summarizes the facts of the cases whose Spanish passages don't contain any DM but their parallel Chinese passages contain DMs.

From Table 1 we can see that, the types of discourse relations in the parallel passages can be same or different when there is a marker change process. In Table 2, we realize that with the new added DM, the type of the discourse relations between the parallel passages are all changed. For example, in the text EEP2, the discourse relation in the Spanish is implicit (Elaboration) because of the lack of the DM. The relation Elaboration is beyond to N-S type under RST. Notwithstanding, in its Chinese parallel passage, the DM *lingyifangmian* (另一方面) represents a List relation, and the relation is beyond to N-N type under RST.

The change of DMs causes changes of relation definition and relation type. As a result, the sentence meaning can be different between the parallel parts. In this work, we assume that the semantics of a discourse relation should transfer from source to target language, as indicated in da Cunha and Iruskieta (2010) and Laali and Kosseim (2014). For this reason, we consider that the semantic aspect doesn't affect the discourse meaning when the translation strategies are used for the Spanish-Chinese translation.

5.2 Unit Shift

As regards the translation strategy of unit shift, we only find 2 cases in total. Below, are the 2 cases and our analysis.

Text name: BMCS2
Spanish: [Metodología actual.][3]$_{\text{S\_Interpretation}}$ [El material de enseñanaza procede de España, ...]$_{\text{N\_Interpretation}}$
English: [Methodology current.] [The material of teaching comes from Spain, ...]
Chinese: [领先的教学方法]$_{\text{S\_Preparation}}$ [我们的教材为西班牙原版教材]$_{\text{N\_Preparation}}$
English: [Leading teaching method] [Our material is the Spanish original]

In this case, there is a period between two EDUs in the Spanish passage. But in its Chinese passage, the two EDUs don't contain any punctuation, and the relation definition between the two EDUs in the Chinese text is different from the Spanish passage. Yet, both Spanish and Chinese passages show the information of the teaching method.

[3] In our work, for some comparison analysis, we use color blue to detect the discourse differences.

Marker change (A DM → Another DM)				
Text name	DM		Relation (Relation type)	
	Spanish	Chinese	Spanish	Chinese
BMCS3	y (eng: and)	huozhe (或者) (eng: or)	List (N-N)	Disjunction (N-N)
FCEC1	y (eng: and)	zhizai (旨在) (eng: aims to)	List (N-N)	Purpose (N-S)
TERM31	igualmente (eng: and)	tongshi (同时) (eng: at the same time)	List (N-N)	Conjunction (N-N)
TERM18	para (eng: for)	ruo (若) (eng: if)	Purpose (N-S)	Condition (N-S)
TERM32	para que (eng: for)	ruo (若) (eng: if)	Purpose (N-S)	Condition (N-S)
TERM38	por lo tanto (eng: therefore)	dan (但) (eng: but)	Result (N-S)	Concession (N-S)

Table 1: Summary of cases that the DMS are different in the parallel passages

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**.**]$_{\text{N\_Evidence}}$ [Las tendencias globalizadas exigen a los renovados autores y talentos que incorporen las nuevas tecnologías y que desarrollen una innovadora experimentación genérica.]$_{\text{S\_Evidence}}$
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，]$_{\text{N\_Evaluation}}$ [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。]$_{\text{S\_Evaluation}}$
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.]

In the above case, the period splits the Spanish passage into two sentences, and the relation between the two sentences is Evidence. Concurrently, there is a comma between the two EDUs in the Chinese passage. The two EDUs hold a Evaluation relation. Same as the prior case, the two different relations in the parallel passages doesn't affect the text idea, we can get the information about the Spanish film industry from both Spanish and Chinese passages.

5.3 Unit shift plus marker change

Comparing to Iruskieta et al. (2015), the first work that analyzes the language translation strategies from discourse level, unit shift plus marker change in our study is a newly discovered translation strategy. We find 3 cases corresponding to this phenomenon.

Text name: ICP3
Spanish: [Los actores son en su mayoría gratuitos]$_{\text{S\_Concession}}$ [**pero** para las actividades que se realizan en nuestro auditorio se recomienda acudir unos minutos antes del cominezo, ya que el aforo de la sala es limitado a 90 personas.]$_{\text{N\_Concession}}$
English: [The acts are mainly free] [but for the

Marker change (No DM → A DM)				
Text name	DM		Relation (Relation type)	
	Spanish	Chinese	Spanish	Chinese
CCICE3	/	yinci (因此) (eng: therefore)	Elaboration (N-S)	Reault (N-S)
FCEC1	/	yinci (因此) (eng: therefore)	Elaboration (N-S)	Result (N-S)
TERM31	/	yinci (因此) (eng: therefore)	Elaboration (N-S)	Result (N-S)
TERM18	/	bing (并) (eng: and)	Condition (N-S)	List (N-N)
TERM32	/	bing (并) (eng: and)	Condition (N-S)	List (N-N)
TERM38	/	bing (并) (eng: and)	Condition (N-S)	List (N-N)
BMCS2	/	wei (为) (eng: and)	Elaboration (N-S)	Purpose (N-S)
EEP2	/	lingyifangmian (另一方面) (eng: on the other hand)	Elaboration (N-S)	List (N-N)
ICP5	/	yucitongshi (与此同时) (eng: meanwhile)	Summary (N-S)	Conjunction (N-N)
TERM31	/	ruo (若) (eng: if)	Evaluation (N-S)	Condition (N-S)
TERM31	/	ye (也) (eng: and)	Elaboration (N-S)	List (N-N)
TERM50	/	dang (当) (eng: when)	Contrast (N-N)	Circumstance (N-S)
TERM50	/	er (而) (eng: however)	Contrast (N-N)	Contrast (N-N)

Table 2: Summary of cases that add a new DM in the Chinese passages

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.]

In the text ICP3, the Spanish passage is an independent sentence and is divided into two EDUs. The relation between the two EDUs is Concession because of the DM *pero* ('but' in English), which is at the beginning of the second EDU. Nevertheless, in the Chinese passage, there is a comma at the end of the first EDU, which makes the Chinese passage contain two sentences. Besides, in the Chinese passage, the DM is erased during the translation process. In the Chinese passage, the relation is Elaboration, which is different from the relation in the Spanish passage.

Text name: TERM31
Spanish: [En las lenguas de flexión compleja,

el tratar solamente el tratar solamente el aspecto formal de las palabras acarrerá malos resultados]$_{\text{N\_List}}$ [**y** será necesaria la lematización.]$_{\text{N\_List}}$
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。]$_{\text{N\_Circumstance}}$ [**此时**词根分析就变得更为不可或缺。]$_{\text{S\_Elaboration}}$
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.]

In the Spanish passage, we can see that the DM *y* splits the sentence into two parts. And the two EDUs hold a List relation. In its parallel Chinese passage, there is a comma at the end of the first EDU. Moreover, the DM in the Chinese passage is *cishi* (此时), whose meaning is 'when' in English and represents a Circumstance relation under RST.

Text name: ICP5
Spanish: [Estudiar español en nuestro instituto no es solo aprender el idioma,]$_{\text{N\_List}}$ [sino que **también** da la oportunidad de conocer.]$_{\text{N\_List}}$ [**y** descubrir las diferentes culturas del mundo hispánico.]$_{\text{N\_List}}$
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.]
Chinese: [[在我们学院学习西班牙语，不仅仅是学习语言本身，]$_{\text{N\_List}}$ [**同时**也是学习西语世界的文化。]$_{\text{N\_List}}$]$_{\text{N\_Summary}}$ [给予你一个了解和发掘西班牙西语世界不同文化的机会。]$_{\text{S\_Summary}}$
English: [In our institute study Spanish, is not only about learning the language itself,] [at the same time it is also about learning Spanish-speaking culture.] [Giving you an knowing and exploring Hispanic world different cultures opportunity.]

In this case, the Spanish passage is divided into three parts by the DMs *también* ('also' in English) and *y* ('and' in English). The three EDUs form a List relation[4]. Differently, although the parallel Chinese passage also contains three EDUs, due to the comma at the end of the second EDU, the Chinese passage contains two sentences. The first two EDUs form a sentence and the last EDU is a single sentence. Like the Spanish passage, the first two EDUs in the Chinese passage hold a List relation because of the DM *tongshi* (同时) ('at the same time' in English). Unlike the Spanish passage, the third EDU in the Chinese passage doesn't contain any DMs and is an additional information of the first two EDUs. The relation between the first two EDUs and third EDU is Summary.

5.4 Different order of EDUs

Different order of EDUs is another new translation strategy that doesn't exist in the work of Iruskieta et al. (2015). Based on the annotation results, we detect a case of this translation strategy.

Text name: CCICE1
Spanish: [En 2015, por la primera vez, la región de Norteamérica se convierte en el tercer feudo por primas de Mafre,]$_{\text{N\_Cause}}$ [desplazando en esa posición a Latam Sur.]$_{\text{S\_Cause}}$
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，]$_{\text{S\_Cause}}$ [上升为西班牙保险公司 Mafre 第三大市场。]$_{\text{N\_Cause}}$
English: [In insurance, North America has surpassed South America,] [rose to the Spanish insurance company Mapfre third largest market.]

In this example,the translation of the first EDU in the Spanish passage is the second EDU of the Chinese passage. In the meantime, the second EDU in the Spanish passage matches the first EDU in the Chinese parallel passage.

5.5 Added discourse

Added discourse is also a new identified translation strategy in this study. During the analysis process, we find only one case about added discourse.

Text name: FICB2
Spanish: [Como conclusión de la formación,

[4] Although the three EDUs are annotated at different discourse level (see Figure 2 in the Appendices part), following van Kuppevelt and Smith (2012), EDUs that form the multi-nuclear relation type are at the same discourse level, so as in this study.

los asistentes compartieron dudas y experiencias.]$_{\text{N\_Elaboration}}$ [Todos los asistentes recibieron los certificados de participación de Hanban y de la FICB.]$_{\text{S\_Elaboration}}$
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.]
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。]$_{\text{N\_Sequence}}$ [**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。]$_{\text{N\_Sequence}}$
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.]

In this case we can see both Spanish and Chinese passages include 4 EDUs. Notwithstanding, in the Chinese passage, a new discourse *peixun jieshu hou* (培训结束后) is inserted at the beginning of the last EDU. The phrase *peixun jieshu hou* (培训结束后) means 'after the training' in English, which composes a Sequence relation with other EDUs. Concurrently, the last EDU in the Spanish passage is an additional information of the third EDU and the relation between the two EDUs is Elaboration (see Figure 3).

6 Conclusion

In this paper, based on the annotations from the RST Spanish-Chinese Treebank, we compare all the annotated parts to find the discourse differences between Spanish and Chinese. The comparison results in this study match the conclusions in Iruskieta et al. (2015). Furthermore, we find some new translation strategies under RST: **unit shift plus marker change**, **different order of EDUs**, and **added discourse**. The research results can help the Spanish-Chinese human translation.

Regarding future work, we will apply our results to the shallow discourse parsing for Spanish and Chinese, with the intention to improve the Spanish-Chinese machine translation (MT) from discourse level.

Acknowledgments

We thank Dr. Iria da Cunha and Dr. Mikel Iruskieta for their insightful comments on this work. This work is partially funded by the Spanish Ministry of Economy and Competitiveness through project TIN2017-85160-C2-2-R, and by the Galician Regional Government under project ED431D 2017/12.

References

Eric Abelen, Gisela Redeker, and Sandra A Thompson. 1993. The rhetorical structure of us-american and dutch fund-raising letters. *Text-Interdisciplinary Journal for the Study of Discourse*, 13(3):323–350.

Jill Burstein. 2009. Opportunities for natural language processing research in education. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 6–27. Springer.

Shuyuan Cao, Iria da Cunha, and Nuria Bel. 2016. An analysis of the concession relation based on the discourse marker aunque in a spanish-chinese parallel corpus. *Procesamiento del Lenguaje Natural*, (56):81–88.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The rst spanish-chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166.

Shuyuan Cao and Harritxu Gete. 2018. Using discourse information for education with a spanish-chinese parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2254–2261.

Iria da Cunha. 2013. A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers. *Research in Computer Science*, 70:95–106.

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

Judy Delin, Anthony Hartley, Cécile Paris, Donia Scott, and Keith Vander Linden. 1994. Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

Oier Imaz and Mikel Iruskieta. 2017. Deliberation as genre: Mapping argumentation through relational discourse structure. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 1–10.

Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Lorraine Edith Kumpf. 1986. Structuring narratives in a second language : descriptions of rhetoric and grammar.

Jan CJ van Kuppevelt and Ronnie W Smith. 2012. *Current and new Directions in Discourse and Dialogue*, volume 22. Springer Science & Business Media.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 610–619.

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

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

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Aysha H Mohamed and Majzoub R Omer. 1999. Syntax as a marker of rhetorical organization in written texts: Arabic and english. *IRAL, International Review of Applied Linguistics in Language Teaching*, 37(4):291.

Thiago Alexandre Salgueiro Pardo. 2005. *Métodos para análise discursiva automática*. Ph.D. thesis, Universidade de São Paulo.

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):241–256.

Guy Ramsay. 2001. What are they getting at? placement of important ideas in lengthy chinese newstext: A contrastive analysis with australian newstext. *Australian Review of Applied Linguistics*, 24(2):17–34.

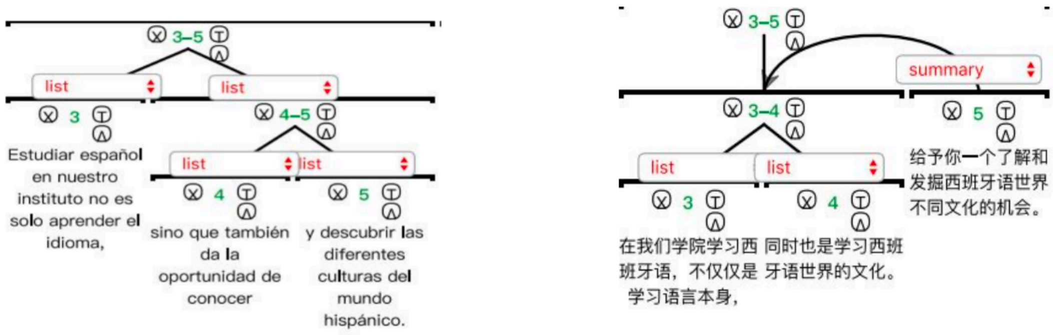
Raphael Salkie and Sarah Louise Oates. 1999. Contrast and concession in french and english. *Languages in Contrast*, 2(1):27–56.

Lanjun Zhou, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. The cuhk discourse treebank for chinese: Annotating explicit discourse connectives for the chinese treebank. In *LREC*, pages 942–949.

A Appendices

Supplied material 1: Figure 2 reflects the annotation case that, although EDU3, EDU4 and EDU5 are annotated at different discourse level, since they hold a multinuclear relation (List), the three EDUs can be considered as the same-discourse level EDUs.

Supplied material 2: Figure 3 shows the annotation of the added discourse case. From the annotation we can see that, in the Chinese passage, the last EDU starts with the added phrase *peixun jieshu hou* (培训结束后) changes the discourse level of the last EDU. Also, the new added discourse changes the relation between the last EDU and its previous EDUs, comparing to the Spanish parallel passage.



Estudiar español en nuestro instituto no es solo aprender el idioma,

sino que también da la oportunidad de conocer

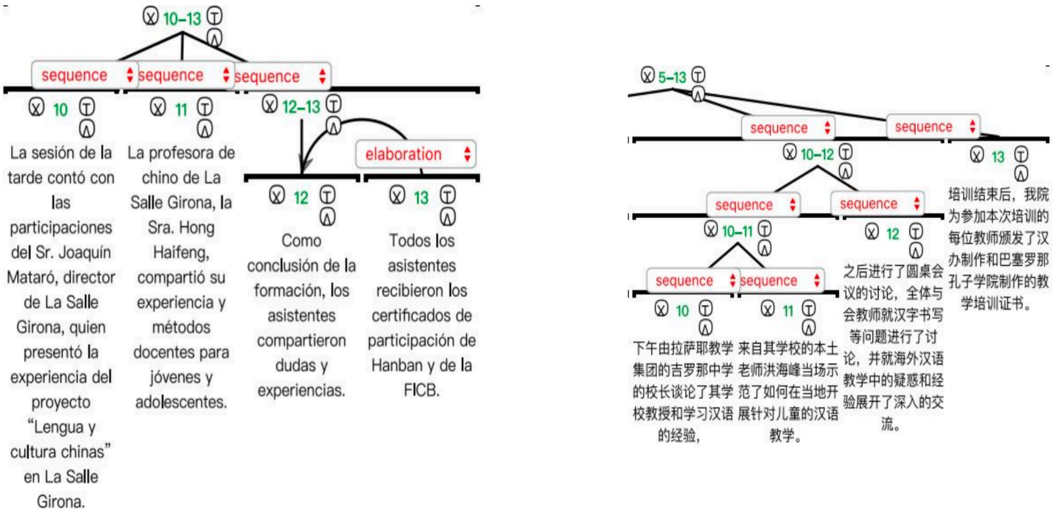
y descubrir las diferentes culturas del mundo hispánico.

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。

给予你一个了解和发掘西班牙语世界不同文化的机会。

Figure 2: The annotation of the parallel parts of text ICP5



La sesión de la tarde contó con las participaciones del Sr. Joaquín Mataró, director de La Salle Girona, quien presentó la experiencia del proyecto “Lengua y cultura chinas” en La Salle Girona.

La profesora de chino de La Salle Girona, la Sra. Hong Haifeng, compartió su experiencia y métodos docentes para jóvenes y adolescentes.

Como conclusión de la formación, los asistentes compartieron dudas y experiencias.

Todos los asistentes recibieron los certificados de participación de Hanban y de la FICB.

下午由拉萨耶教学集团的吉罗那中学的校长谈论了其学校教授和学习汉语的经验，

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。

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。

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。

Figure 3: The annotation of the parallel parts of text FICB2