Using Discourse Information for Education with a Spanish-Chinese Parallel Corpus

Shuyuan Cao▲*, Harritxu Gete*
▲Universitat Pompeu Fabra (UPF)
*University of Basque Country (UPV/EHU)
shuyuan.cao@hotmail, harritxu.gete@gmail.com

Abstract

Nowadays, with the fruitful achievements in Natural Language Processing (NLP) studies, the concern of using NLP technologies for education has called much attention. As two of the most spoken languages in the world, Spanish and Chinese occupy important positions in both NLP studies and bilingual education. In this paper, we present a Spanish-Chinese parallel corpus with annotated discourse information that aims to serve for bilingual language education. The theoretical framework of this work is Rhetorical Structure Theory (RST). The corpus is composed of 100 Spanish-Chinese parallel texts, and all the discourse markers (DM) have been annotated to form the education source. With pedagogical aim, we also present two programs that generate automatic exercises for both Spanish and Chinese students using our corpus. The reliability of this work has been evaluated using Kappa coefficient.

Keywords: discourse analysis, education, corpus, bilingual language learning

1. Introduction

Using natural language processing (NLP) for educational applications starts from the early history (Burstein, 2009). Different NLP studies make a great advance in different educational areas, for instance, translation studies, text retrieval, text mining or speech recognition.

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).

As two of the most spoken languages in the world, Spanish and Chinese occupy important positions in NLP development. Due to the great linguistic distance that between this pair of languages, the number of differences in their discourse structure is also considerable. The following examples show some of the discourse differences between Spanish and Chinese.

Ex.1:

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

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

[DM[1] still no get still no get results,]Unit$_1$ [we consider that the model will more extensive than the sentence group nominal[2].]Unit$_2$

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

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

[We consider that the model will more extensive than the sentence group nominal.]Unit$_1$ [DM still no get still no get results]Unit$_2$

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

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

In Example 1, we can see that the Spanish passage and the Chinese one have a similar discourse structure. Both passages start with a discourse marker in the first unit. However, the discourse markers 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. These two discourse markers are equivalent to the English discourse marker ‘although’. By contrast, in Spanish, just one DM “*aunque*” is needed to express the same meaning. Although in 1.1 it is being used at the beginning of the first unit, as we can see in 1.2, the order of the discourse units in this Spanish passage 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.

Ex.2:

2.1 Sp: Si optas por un aprendizaje lo más parecido posible a la inmersión, y necesitas mejorar tu nivel de español rápidamente, los cursos intensivos son una buena opción.

[1] DM means discourse marker. We will give the specific definition of discourse marker in the methodology section.

[2] In this work, for all the presented examples, we will give an English literature translation for each example.

[Si optas por un aprendizaje lo más parecido posible a la inmersión,]Unit$_1$ [y necesitas mejorar tu nivel de español rápidamente,]Unit$_2$ [los cursos intensivos son una buena opción.]Unit$_2$

[DM1 you opt for a learning as more similar possible to immersion,]Unit$_1$ [DM2 you need to improve your level of Spanish quickly,]Unit$_2$ [the courses intensive are a good option.]Unit$_3$

2.2 Ch: 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，紧凑课程是一个很好的选择。

[若您希望进行全面集中的语言学习，]Unit$_1$ [或者您希望短时间内提高您的语言水平，]Unit$_2$ [紧凑课程是一个很好的选择。]Unit$_3$

[DM1 you want to focus on completely intensive of language learning,]Unit$_1$ [DM2 you wish in short time to improve your language level,]Unit$_2$ [the intensive courses are a good option.]Unit$_3$

2.3 Eng: If you want to focus on language intensive learning or if you want to improve your language skills in a short time, a compact program is a good choice.

In Example 2, we can see that there are two DMs in the Spanish passage, one is "*si*" ('if' in English) and another one is "*y*" ('and' in English). The DM "*y*" represents a LIST relation between the first unit and the second unit. Meanwhile, the DM "*si*" connects the first two units (Unit$_{1\text{-}2}$) and the third unit (Unit$_3$) as a CONDITION relation. In the Chinese passage, there are also two discourse markers. One of them is "*ruo*" (若), which means 'if' in English; the other one is "*huozhe*" (或者) and corresponds to 'or' in English. The DM "*huozhe*" (或者) represents a DISJUNCTION relation between the first and second unit. Same as the Spanish passage, The DM "*ruo*" (若) also gives a CONDITION relation between the first two units (Unit$_{1\text{-}2}$) and the third unit (Unit$_3$).

Although this is a parallel example, we can see that the discourse relation between the first unit and the second unit are different in the Spanish and Chinese passage. In the Spanish passage, the relation within the first two units is LIST while in the Chinese passage the first two units hold a DISJUNCTION relation. This is because of the translation strategy. The discourse relations are different, but the main idea in both passages is the same.

These two examples show some of the differences in the discourse structure of Spanish and Chinese. Since this pair of languages have considerable differences in this aspect, a comparative discourse analysis between Spanish and Chinese is essential for language learning. Therefore, this work aims to give a discourse analysis with a Spanish-Chinese parallel corpus. This analysis could be beneficial for Spanish-Chinese language learning education, from discourse level point of view. In addition, with educational purposes, we have also developed two programs for automatic exercise generation. The generated exercises can be used by Spanish and Chinese language learning students in order to practice the usage of DMs in these languages.

In the second section, we introduce the theoretical framework of the work. In the third section, we talk about the state of the art. In the fourth section, we present the methodology. Firstly, we explain how we create and annotate the corpus and then, we explicate the methodology used for the automatic exercise generation. In the fifth section, we describe the evaluation method and we show the obtained results. Finally, in the last section, we explain our conclusions and the possible future work.

2. Theoretical Framework

Different theories and approaches have been applied to discourse analysis, the most used are: a) The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), b) the Penn Discourse Tree-Bank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2004) and c) the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003).

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 defined to hold between EDUs; the EDUs can be Nuclei or Satellites. Satellites offer additional information about nuclei. The EDUs can be linked among them holding a nucleus-satellite (e.g. CAUSE, JUSTIFY, EVIDENCE, CONCESSION) function or a multinuclear (e.g. CONJUNCTION, LIST, SEQUENCE) function and 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.

PDTB is a large-scale annotation project and extends the work of the Penn Treebank (PTB) (Marcus, Santorini and Marcinkiewicz, 1993) and Propbank (Kingsbury and Palmer, 2002). Grounded on a lexicalized approach to discourse, the discourse connectives and their arguments have been annotated in PDTB (for instance, contingency, temporal, expansion). Sense labels for each relation with a hierarchical classification scheme. The senses annotation works detect the polysemy of connectives and make the PDTB helpful for sense disambiguation tasks (Miltsakaki et al., 2005).

SDRT explores the logical form between discourse interpretation and discourse coherence by using rhetorical relations to model the semantics/pragmatics of a text. In SDRT, the logic of information content is used to represent the logical forms of discourse and the glue logic is being applied to construct logic forms. The logic forms represent the syntax and dynamic semantics of a language. SDRT can model the complexity of the interaction between semantics and pragmatics under some discourse relations (e.g. CAUSE, EXPLANATION, CONTRAST).

RST has been selected as the theoretical framework of this work. Comparing to PDTB and SDRT, RST focuses on the hierarchical structure of a whole text, where discourse relations can be annotated within a sentence (intra-sentence style) and between sentences (inter-sentence style). The intra-sentence annotation and inter-sentence annotation styles help to inform how discourse elements are being expressed in a language and translation strategies (if there are) can be detected in different levels of an RS-tree (da Cunha and Iruskieta, 2010; Iruskieta, da Cunha and Taboada 2015).

3. State of the art

3.1 RST Corpus and Education

Studies addressing RST for education have been applied to different language pairs. 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 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 expressed by Chinese native speakers by means of RST, and the study 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 the theoretical framework to explore the different features with 60 English essays written by Chinese students. Beside of 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.

Regarding the language pair Spanish-Chinese, few works exist and can be useful for Spanish-Chinese education. Yao (2008) uses film dialogues to elaborate an annotated corpus, and compares the Chinese and Spanish discourse markers in order to give some suggestions for teaching and learning Spanish and Chinese. In this work, Yao does not use a particularly detailed framework and only offers a comparative analysis of Spanish and Chinese discourse markers, followed by his conclusions. Taking different newspapers and books as the research corpus, Chien (2012) compares the Spanish and Chinese conditional discourse markers to give some conclusions of the conditional discourse marker for foreign language teaching between Spanish and Chinese. Wang (2013) uses Pedro Almodóvar's films *La mala educación* and *Volver* as the corpus to analyze how the subtitled Spanish discourse markers can be translated into Chinese, so as to make a guideline for translation education between the language pair. However, none of these works use RST as its theoretical framework. The only work for the Spanish-Chinese education under RST is the work of Cao, da Cunha and Bel (2015). They explore the different Chinese translations of the Spanish DM "*aunque*" ('however/but') in the UN corpus to indicate how to translate the Spanish DM "*aunque*" into Chinese during the foreign language study. Yet, this work only focuses on the discourse structure of the single sentences instead of the whole texts.

3.2 Exercise Generation and Multi Choice Question

Regarding the exercise generation aspect, some successful studies have been applied to education filed by using different approaches. For example, under statistical ranking module, address the challenge of automatically generating questions from reading materials for educational practise and assessment; Heilman and Smith (2010) give a rule-based system to rank the output to give the *wh-* question. 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. Concept map is another approach can be used for question generation. 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 that focuses on the discourse structure of the whole texts under RST for Spanish-Chinese language education, and contains the exercise generation function.

4. Methodology

We carry out different steps for this study and the following subsections details our methodology.

4.1 Research Corpus

Cao, da Cunha and Iruskieta (2017) indicate that there is not an adequate Spanish-Chinese parallel corpus for discourse analysis under RST; therefore, we construct a new Spanish-Chinese parallel corpus. We have 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 have searched for texts with these characteristics in different sources. To obtain a 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 with the people in charge of the organizations that had been 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. Finally, 50 Spanish texts and their parallel Chinese texts have been selected for our study. The longest text includes 1,201 words and the shortest text contains 91 words.

The original sources of these texts are: (a) International Conference about Terminology (1997), (b) Shanghai Miguel Cervantes Library, (c) Chamber of Commerce and Investment of China in Spain, (d) Spain Embassy in Beijing, (e) Spain-China Council Foundation, (f) Confucius Institute Foundation in Barcelona, (g) Beijing Cervantes Institute and (h) Granada Confucius Institute. Moreover, in order to guarantee the representativeness of our corpus, we have selected different types of texts from several domains. The genres of the texts are four: (a) abstracts of research papers, (b) news, (c) advertisements and (d) announcements.

4.2 Segmentation Annotation

In this work, we first segment the whole corpus with RSTTool (O'Donell, 2000) manually. We adopt the segmentation criteria by Cao et al. (2017), which can be applied to both Spanish and Chinese. The segmentation criteria are the following:

- Paragraphs and line breaks. A line break will be taken as an independent EDU to segment the titles (and subtitles).
- Sentences and periods. A period will be taken as an independent EDU.
- Question mark and exclamation mark. Both marks are signals of a sentence boundary.
- Other EDUs should have an adjunct verb phrase. This is a basic segmentation criterion and all other following segmentation criteria should follow this rule.
- DM, comma and adjunct verb phrase. If there is a DM at the beginning of the sentence and the sentence is being divided into two parts by a comma. In addition, if a DM is after a comma and the EDU has a verb, we also segment.
- Semicolon plus adjunct verb phrase.
- Parenthetical and dash. Only when the parented unit does not modify noun neither adjective, it is an independent segment; if within the parenthetical unit there are coordinated parts, we also segment the coordinated parts.
- Coordination and ellipsis with verbs. Coordinated clauses with verbs represent the independent EDUs, including where the subject is eliminated in the following EDUs.

For those non-EDU segmentation criteria are the following:

- Relative, modifying and appositive clauses. Relative clauses, clauses that modifies a noun or adjective, or appositive clauses are not considered as EDUs.
- Reported speech. Reported speech cannot be considered 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 corpus is accessible to the academic community. More detailed information of the corpus can be consulted at: `http://ixa2.si.ehu.es/rst/zh/`, including the segmentation examples of each text in the corpus.

4.3 Discourse Markers Annotation

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

We have categorized different types of DMs as following show[3]:

➢ N-S type

- Antithesis

<u>Nuclear</u>: The author favors the idea.
<u>Satellite</u>: The author disfavors the idea.
<u>Spanish DM(s)</u>: aunque; por el contrario; sino
<u>Chinese DM(s)</u>: 但是

- Cause

<u>Nuclear</u>: A situation.
<u>Satellite</u>: Another situation that causes that one.
<u>Spanish DM(s)</u>: como; debido a; ya que
<u>Chinese DM(s)</u>: 因为; 由于

- Circumstance

<u>Nuclear</u>: The text shows the ideas or the events that occur in the interpretive text.
<u>Satellite</u>: An interpretive context of situation or time.
<u>Spanish DM(s)</u>: cuando
<u>Chinese DM(s)</u>: 作为; 如同

- Concession

<u>Nuclear</u>: A situation confirmed by the author.
<u>Satellite</u>: Another situation inconsistent but also affirmed by the author.
<u>Spanish DM(s)</u>: pero; sino que; si bien
<u>Chinese DM(s)</u>: 尽管; 然而

- Condition

<u>Nuclear</u>: Action or situation whose occurrence results

[3] We have annotated all the DMs in the research corpus, for each text, we have annotated the DMs within a sentence and between the sentences. Due to the translation strategies, not all the discourse relations contain the DMs for both languages. We use "/" to indicate the cases that donot have the DMs in the corpus.

from the occurrence of the conditioning situation.
Satellite: A condition situation.
Spanish DM(s): si
Chinese DM(s): 若; 如果

- 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

- 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

- Interpretation

Nuclear: A situation.
Satellite: An interpretation of the situation.
Spanish DM(s): en concreto
Chinese DM(s): /

- 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

- Restatement

Nuclear: A situation.
Satellite: A re-expression of the situation.
Spanish DM(s): es decir
Chinese DM(s): 即

- 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

- 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

- Conjunction

Nuclear: A situation or an action.
Nuclear: 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

- Contrast

Nuclear: One alternate.
Nuclear: The other alternate.
Spanish DM(s): por el contrario
Chinese DM(s): 而; 相反

- Disjunction

Nuclear: An alternative.
Nuclear: Another alternative.
Spanish DM(s): o
Chinese DM(s): 或; 或是; 或者; 亦或

- List

Nuclear: An item.
Nuclear: The 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

- Sequence

Nuclear: An item.
Nuclear: 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

4.4 Exercise Elaboration

4.4.1. Exercise for L2 Spanish Learner

The exercises to practice Spanish language consist of different texts with some blanks within them. After each text, there are multi-choice answers for the user, who can choose between several DMs. These exercises have been generated automatically by removing the annotated DMs from the texts of the corpus. The distractors are DMs that can be used in the same context as the correct answer. Apart from the automatic generation of the exercises, the system can also grade the answers of the user and it gives the correct ones.

We have used the Python programming language to generate the texts automatically by removing the annotated DMs.

First of all, we have annotated all the Spanish DMs in the research corpus. Secondly, we have made a program to generate the Spanish exercises one by one automatically (an exercise is created from each text). The following steps have been carried out to make the program: (a) We have elaborated a list of the DMs we want to remove. This list is created with all the annotated DMs that appear in the texts we are using to develop the program. (b) With a simple program developed with Python, we remove the DMs following this rules: some of the DMs of the list have to be removed always, other ones only if they appear at the beginning of the sentence and finally, there are some DMs that we have to be removed only if in the same sentence appears another specific word (this is the case of the composed DMs). (c) Finally, to select the possible answers of the exercises, we have created 7 groups of DMs depending on their discourse meaning. When a DM is removed, the distractors of the exercise are selected from those in the same group. However, within each group, the DMs are grouped if it is almost impossible to distinguish between them. In this case, one cannot be used as a distractor of the other.

Thirdly, we have made another program to grade the answers.

4.4.2. Exercise for L2 Chinese Learner

Similarly, we have made a small program to take out all the discourse markers in the Chinese texts. However, the exercise design is different from Spanish language exercises. In this case, instead of giving different options for each blank, the system gives all the erased DMs at the end of the text, as choices for the Chinese language student.

Firstly, we have annotated all the DMs in each Chinese text and then, we have designed a simple program to erase from the texts all the DMs in the list.

The reason to make two different designs for Spanish and Chinese texts is because, although the texts are parallel, comparing with the Spanish texts, the Chinese texts are more difficult to understand because of the different meanings but the same word (including some annotated DMs[4]). Therefore, we consider that, for a Chinese text, it is better to remove all the DMs and mix the correct answers to let the users to choose so that they can understand the text better by filling the DMs.

5. Evaluation

In this work, we use the Cohen's Kappa coefficient to evaluate the correctness of our automatic exercise generation program. Previous works use Kappa to measure the annotation agreement for RST studies (Iruskieta, Diaz de Ilarraza and Lersundi, 2015; Cao et al., 2017), Kappa gives the agreement of annotation as follows:

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

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

We have developed our program using only a part of the annotated corpus (60% of the corpus, 30 Spanish texts and their parallel Chinese texts). Then, the program has been applied to the rest of the corpus in order to measure its accuracy, so the Kappa coefficient evaluates the correctness of our program deleting the DMs in those texts.

Table 1 shows the K results of the 20 tested Spanish texts and their parallel Chinese texts[5].

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 1: Program accuracy of the 40 tested Spanish-Chinese parallel texts

From table 1, 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 analysing 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 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 the texts that contain low results and we find some common limitations for these texts, here we give CCICE1 as the example. In this text, the Chinese character "*wei*" (为) appear 7 times, however, none of the them can be considered as the DM, whose discourse 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 do' in English), "*wei*" (为) cannot be considered as a DM.

The sequence of the phrases in Chinese 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 do something' 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)[6]. Since there is no space between Chinese characters in a text, hence, our programming considers "*zhi*" (旨) and "*zai*" (在) as a DM.

In conclusion, these are the limitations of our program in

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

[5] As section 4.1 shows, the research 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.

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

the Chinese subcorpus:

- A character could have different meanings depending on the text content, so that sometimes it is a DM and sometimes not, but our system cannot understand the text content.
- Some DMs are composed of two Chinese characters, however, our programming just annotates one character. For instance, in the case of "*yiji*" (以及), the first character "*ji*" (及)[7] is removed from the output. Other similar cases exist.
- The possible phrase sequences can cause the characters combine as an annotated DM, for example, the case of "*zhuzhi zaiyu*" (主旨//在于) that we explained before.
- 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.

6. Conclusion and Future Work

In this work, we have presented the first RST Spanish-Chinese parallel corpus that can be used for language exercise with multichoices for each text.

In this work, we have presented the first RST Spanish-Chinese parallel corpus, which can be used for the important task of discourse analysis. Experts have annotated it manually so that DMs are indicated and classified.

With educational aim, we have used the corpus to create a program that automatically generates language exercises with multiple choices for each text. These exercises are useful to learn the usage of discourse markers for Spanish and Chinese language students.

Despite the simplicity of our program, we get very good result for the Spanish subcorpus. In the case of the Spanish exercises, our system can also grade the users' answers automatically. While for Chinese subcorpus we get some limitations, it can also give good L2 Chinese language exercises.

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

7. Acknowledgements

We give our acknowledgements to the IXA group of the University of Basque Country, for the construction of the corpus webpage. Especially, we want to thank Itziar Aldabe, for her insightful ideas on this work.

8. References

Asher, N., and Alex, L. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Burstein, J. (2009). Opportunities for Natural Language Processing in Education. 2009. *Lecture Notes in Computer Science*, 6-27

Cao, S. Y., da Cunha, I., and Bel, N. (2015). 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.

Cao, S. Y., da Cunha, I., and Iruskieta, M. (2017). Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *EpiC Series in Language and Linguistics*, 2: 315-324.

Cao, S. Y., Xue, N. W., da Cunha, I., Iruskieta, M., and Wang, C. (2017). Discourse Segmentation for Building a RST Chinese Treebank. In *Proceedings of 6th Workshop "Recent Advances in RST and Related Formalisms"*, 73-81.

Chen W., Aist G., and Mostow J. (2009). Generating Questions Automatically from Informational Text. In *The 2nd Workshop on Question Generation, volume 1 of AIED 2009 Workshops Proceedings*, 17-24.

Chien, Y. S. (2012). Análisis contrastivo de los marcadores condicionales del español y del chino. *MarcoELE: Revista de Didáctica Español Lengua Extranjera (14)*.

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

Eckle-Kohler, J., Kluge, R., and Gurevych, I. (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.

Fang, Y. (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.

Heilman M., and Smith N. 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.

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

Iruskieta, M., Diaz de Ilarraza, A., and Lersundi, M. (2015). Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*, 11(2) : 303-334.

Kingsbury, P., and Palmer M. (2002). From Treebank to PropBank. In *Proceddings of International Conference on Language Resources and Evaluation (LREC' 2002)*, 1989-1993.

Li, H. K., and Liao, Z. H. (2015). A Rhetorical Structure Theory Based Approach to Texture Coherence in

[7] "*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.

Chinese EFL Learners' Argumentative Writing. *Overseas English*, 16: 201-204.

Liang, S. S., and Yang, Z. L. (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, W. C., and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text&Talk*, 8(3): 243-281.

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

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

Miltsakaki E., Prasad R., Joshi A., and Webber B. (2004). The Penn Discourse Treebank. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC' 2004)*, 2237-2240.

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

Olney A. M., and Graesser A. C., and Person N. K. (2012). Question Generation from Concept Maps. *Dialogue and Discourse*, 3(2) : 75-99.

Pórtoles, J. (2001). *Marcadores del discursivo*. 4th edition. Barcelona: Ariel.

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

Wang, Y. C. 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.

Yao, J. M. (2008). *Estudio comparativo de los marcadores del discurso en español y en chino a través de diálogos cinematográficos*. PhD thesis. Valladolid: Universidad de Valladolid.

Zhang, Y. (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, L. J., Wei, Z. Y., and Wong, K. F. (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.