Elaboration of a protocol to support Chinese-Spanish translation: an approach based on a parallel corpus annotated with discourse information

La elaboración de un protocolo de apoyo a la traducción chino-español: una aproximación basada en un corpus paralelo anotado con información discursiva

Shuyuan Cao
Institut Universari de Lingüística Aplicada,
Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018, Barcelona
shuyuan.cao01@estudiant.upf.edu

Resumen: La traducción chino-español es especialmente complicada debido a las grandes diferencias gramaticales, sintácticas y discursivas entre ambas lenguas. En este proyecto de tesis doctoral se propone contrastar el discurso producido en textos paralelos en estas lenguas y describir cómo la información discursiva se expresa formalmente en cada una de ellas. Se establecerá una tipología de diferencias discursivas entre estas lenguas para redactar un protocolo que pueda ser de utilidad tanto a traductores humanos como a investigadores en traducción automática. El marco teórico utilizado será la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988) y se utilizará la metodología de comparación de Iruskieta, da Cunha y Taboada (2014).
Palabras clave: traducción, análisis del discurso, traducción automática (TA)

Abstract: Mandarin Chinese-Spanish translation is particularly complicated because of the extensive grammatical, syntactic and discursive differences between the two languages. This PhD project proposes to contrast the discourse produced in parallel texts in these languages and to describe how the discursive information is formally expressed in both of them. A typology of discourse differences between the two languages is established in order to draft a protocol that can be useful for both human translators and researchers in machine translation (MT). The theoretical framework is Rhetorical Structure Theory (RST) by Mann and Thompson (1988) and the used comparison methodology is Iruskieta, da Cunha and Taboada (2014).
Keywords: translation, discourse analysis, machine translation (MT)

1 Motivation and Related work

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). In particular, the relation between MT and discourse analysis has only recently begun and works addressing this topic remain limited. A shortcoming of most of the existing systems is that discourse level is not considered in the translation, which therefore affects translation quality (Mayor et al., 2009; Wilks, 2009). Notwithstanding, some recent researches indicate that discourse structure improves MT evaluation (Fomicheva et al., 2012; Tu, Zhou and Zong, 2013; Guzmán et al., 2014).

Nevertheless, thus far there have not been many studies addressing this topic. The studies that use Rhetorical Structure Theory (RST) by Mann and Thompson (1988) as framework are a contribution for discourse analysis research. 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).

Some comparative studies between Chinese and English by employing RST exist. 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. There are few contrastive works between Spanish and Chinese. None of them uses RST. 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 human translation and audiovisual translation between the language pair.

Let's see two examples of discourse differences between Chinese and Spanish.

Ex. 1:

1.1. Ch: **虽然**他病得很重，**但是**他去上班了。

[**虽然**他病得很重，]EDU$_S$ [**但是**他去上了。]EDU$_N$

(**marker_1** he ill very, **marker_2** he goes to work.)

1.2. Sp: **Aunque** está muy enfermo, va a trabajar.

[**Aunque** está muy enfermo,]EDU$_S$ [va a trabajar.]EDU$_N$

(**marker_1** is very ill, goes to work.)

1.3. En: **Though** he is very ill, he goes to work.

In example 1, Chinese and Spanish passages show the same rhetorical relation (Concession), and the order of the Nucleus and the Satellite is also similar. However, in Chinese, it is mandatory to include two discourse markers to show this relation: one marker "*suiran*" (虽然) at the beginning of the Satellite and another marker "*danshi*" (但是) at the beginning of the Nucleus. These two discourse markers are equivalent to the English discourse marker *although*. By contrast, in Spanish, to show the Concession relation, only one discourse marker is used at the beginning of the Satellite (in this case, "aunque", *although*).

Ex. 2:

2.1. Ch: 很冷，虽然没有下雨。

[很冷，]EDU$_N$ [**虽然**没有下雨。]EDU$_S$

(It´s cold, **marker_1** there is no rain.)

2.2.1 Sp: Hace frío, aunque no llueve.

[Hace frío,]EDU$_N$ [**aunque** no llueve.]EDU$_S$

(Makes cold, **marker_1** no rain.)

2.2.2 Sp: Aunque no llueve, hace frío.

[**Aunque** no llueve,]EDU$_S$ [hace frío.]EDU$_N$

(**marker_1** no rain, has cold.)

2.3. En: It is cold, **though** there is no rain,

In example 2, the Chinese passage could have the same or the different rhetorical structure. In the Chinese passage, the discourse marker "*suiran*" (虽然) at the beginning of Satellite, which is equivalent to the English discourse marker *although*, shows a Concession relation, and the order between Nucleus and the Satellite cannot be changed. In the Spanish passage, "aunque" is also at the beginning of Satellite, which also corresponds to the English discourse marker *although*, and shows the same discourse relation, but the order between Nucleus and Satellite can be changed and this makes sense syntactically.

Therefore, the discourse analysis between the language pair Chinese-Spanish is very important, otherwise, erroneous results will appear for the translation between these two languages. The motivation of this PhD research is to help to improve the Chinese-Spanish translation quality by analyzing discourse level.

2 Aims and Hypothesis

Main objective: Develop a protocol including guidelines to correctly show discourse

information for human translation and MT between the two languages.

Specific objectives:

1) Contrast the discourse produced in Chinese-Spanish parallel texts in order to describe how discourse information is formally expressed in both languages.

2) Establish the types of differences between discourse in Chinese and Spanish.

These goals are related with the following hypotheses:

1) The similarities and differences between the discourse produced in Chinese and Spanish have to be modeled by using discourse information given in the framework of RST, such as discourse segmentation, position of discourse markers in Nuclei and Satellites, and discourse relations.

2) The discourse information should be considered for both human translation and MT between Chinese and Spanish.

3 Methodology

Our methodology includes the following steps:

1) Find a parallel Chinese-Spanish corpus and use RST to annotate it. Specifically, we will annotate EDUs, discourse relations (including discourse markers) and discourse structure. We will use official documents from the United Nations Multilingual Corpus (Eisele and Chen, 2010) and patent abstracts included in the Lumera's (2009) corpus. Table 1 presents the detail information of the research corpus.

Name	**UN corpus**	**Patent abstracts**
Text types	Official documents	Patent abstracts
Number of Chinese texts	65,022	50
Number of Spanish texts	70,509	50
Number of parallel texts	62,738	50
Domain	Wars, cooperation regional, development of culture, etc.	Chemistry, technic, medicine, etc.
Available to access	Public	Ask for the permission of the author

Table 1: Detail information of the research corpus

2) Compare annotated discourse structures manually in both languages, following the method proposed by Iruskieta, da Cunha and Taboada (2014). The selected RST relations for this PhD research are in the following table.

N-S	**N-N**
Circumstance	Contrast
Solutionhood	Joint
Elaboration	List
Background	Sequence
Enablement	Same-unit
Motivation	
Evidence	
Justify	
Antithesis	
Concession	
Interpretation	
Cause	
Result	
Otherwise	
Purpose	
Restatement	
Summary	

Table 2: Selected RST relations for PhD research

3) Describe the main discourse similarities and differences found in the Chinese-Spanish corpus, in relation with: a) discourse segmentation, b) nuclearity and discourse relations and c) rhetorical trees.

4) Elaborate a typology of similarities and differences between Chinese and Spanish.

5) Establish a protocol including guidelines to correctly show Chinese-Spanish discourse information, useful for human translators and researchers that work on MT.

4 Specific research questions wishing to discuss at the Doctoral Symposium

(a) Which type of RST discourse elements would show relevant Chinese-Spanish discourse differences?

(b) How could our results contribute to human translators and MT?

(c) Is the corpus for this research appropriate? Which characteristics should have the corpus? How many words or / and texts should be enough to achieve our goals?

References

Chien, Y. S. 2012. Análisis contrastivo de los marcadores condicionales del español y del chino. PhD thesis. Barcelona: Universitat Autònoma de Barcelona.

Cui, S. R. 1985. Comparing Structures of Essays in Chinese and English. Master thesis. Los Angeles: University of California.

Eisele, A.; Chen, Y. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. In *Language Resources and Evaluation Conference 2010*. 2868-2872.

Fomicheva, M.; da Cunha, I.; Sierra, G. 2012. La estructura discursiva como criterio de evaluación de traducciones automáticas: una primera aproximación. In *Empiricism and analytical tools for 21 century applied linguistics: selected papers from the XXIX International Conference of the Spanish Association of Applied Linguistics (AESLA)*. 973-986.

Guy, R. 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, R. 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, F.; Joty, S.; Màrquez, Ll.; Nakov, P. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 687-698.

Iruskieta, M.; da Cunha, I.; Taobada, M. 2014. A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora. *Language resources and evaluation.* 1-47. To appear.

Kong, K. C. C. 1998. Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text* 18(1). 103-141.

Lumeras, M. A. 2009. *Estudio descriptivo multilingüe del resumen de patente: aspectos contextuales y retóricos*. In Lang, Peter (ed.). German.

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

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

Mayor, A.; Alegria, I.; Díaz de Ilarraza, A.; Labaka, G.; Lersundi, M.; Sarasola, K. 2009. Evaluación de un sistema de traducción automática basado en reglas o porqué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural* 43. 197-205.

Tu, M.; Zhou, Y.; Zong, C. Q. 2013. A Novel Translation Framework Based on Rhetorical Structure Theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 370-374.

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.

Wilks, Y. 2009. *Machine Translation: Its scope and limits*. 3ª ed. New York: Springer.

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.

Zhou, L. J.; Li, B. Y.; Wei, Z. Y.; 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.* 942-949.