EPiC Series in Language and Linguistics
Volume 2, 2017, Pages 315–324
Professional and Academic Discourse:
an Interdisciplinary Perspective

Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus

Shuyuan Cao[1] Iria da Cunha[2] and Mikel Iruskieta[3]
[1] Universitat Pompeu Fabra (UPF)
[2] Universidad Nacional de Educación a Distancia (UNED)
[3] University of the Basque Country (UPV/EHU)
shuyuan.cao@hotmail.com, iriad@flog.uned.es, mikel.iruskieta@ehu.eus

Abstract

Spanish and Chinese are two very different languages in all language levels. Therefore, translation (both human and machine translation) from one to another and learning one of them as a foreign language are challenging tasks. Some automatic translation systems exist for this pair of languages, but there is enough room to improve the translation quality between Spanish and Chinese. In addition, the accessible sources, such as a parallel corpus for studying and understanding this language pair, are still few. In this paper, we present how we have created a Spanish-Chinese parallel corpus designed for language learning and translation tasks at the discourse level. This corpus has been enriched automatically with part-of-speech (POS) and several queries based on morpho-syntactic information can be realized. We have made available the parallel corpus to the academic community.

1 Introduction

Nowadays Spanish and Chinese are two of the most spoken languages in the world. 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. In such a task, a parallel corpus can help for translation and language learning purposes.

There are various differences between Spanish and Chinese, starting from the characters to their syntax and discourse structure. Here we give two examples in order to show some discourse similarities and differences between Spanish and Chinese.

(1.1) Sp: Aunque él está enfermo, va a trabajar.
[**Aunque** él está enfermo;]Unit1 [va a trabajar.]Unit2

[**DM**[1] he is ill, go to work.[2]]
(1.2) Ch: 虽然他生病了，但是他去上班了。
[**虽然**他生病了，[3]]Unit1 [**但是**他去上班了。]Unit2
[**DM1** he ill; **DM2** he go to work.]
(1.3) ENG: Although he is ill, he goes to work.[4]

In example (1), Spanish and Chinese passages can be related with the same rhetorical relation (CONCESSION[5]), and the order of the two discourse units is the same. However, in Chinese, it is mandatory to include two discourse markers to express this relation: in this case, the marker *suiran* (虽然) (‘although’) at the beginning of the first unit and the marker *danshi* (但是) (‘but’) at the beginning of the second unit. By contrast, in Spanish, to express the CONCESSION relation, only one discourse marker is necessary, the marker *aunque* (‘although’) is used at the beginning of the first unit.

(2.1.1) Sp: Hace frío, aunque no llueve.
[Hace frío]Unit1 [**aunque** no llueve.]Unit2
[Makes cold, **DM** no rain.]
(2.1.2) Sp: Aunque no llueve, hace frío.
[**Aunque** no llueve,]Unit1 [hace frío.]Unit2
[**DM** no rain, makes cold.]
(2.2) Ch:很冷，虽然没有下雨。
[很冷，]Unit1 [**虽然**没有下雨。]Unit2
[It’s cold, **DM** there is no rain.]

(2.3) ENG: Although he is ill, he goes to work.

In example (2), the units of the Spanish and Chinese passages can be related with the same rhetorical relation (CONCESSION). However, under the same rhetorical relation, the discourse structure in the two languages can be the same can also be different. In the Chinese passage, the discourse marker *suiran* (虽然) (‘although’) at the beginning of the second unit shows a CONCESSION relation, and the order between the two units cannot be changed without changing the sense of the sentence. In the Spanish passage, the discourse marker *aunque* (‘although’) is also at the beginning of the second unit and shows the same discourse relation, but the order between the two units is changeable and this does not change the sense of the sentence.

The aforementioned examples show the complexity of the differences between Spanish and Chinese (the quantity of discourse markers in each language, the order of the units, among other issues). At present, there are some automatic systems that change the characters from one language to another in an acceptable way. Several studies show that discourse elements affect translation quality (Mayor et al., 2009; Wilks, 2009, among others), thus, we consider that a Spanish-Chinese discourse comparative study would offer clues to identify equivalent discourse structures in both languages that could affect the translation quality.

[1] DM means discourse marker. In this work, we adopt the definiton of Portolés (2001). 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.

[2] In this work, for each example, we give an English literal translation in brackets for both languages in order to make the reader understand better the content of the passage.

[3] All the Chinese characters and punctuations occupy two spaces in a written text; therefore, readers can see a blank space between a Chinese character and its following bracket, or a blank space between a Chinese punctuation and its following bracket.

[4] All the examples in this work include an English translation that has been done by the authors of this study.

[5] The CONCESSION relation is a discourse relation under the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988).

In order to carry out this discourse comparative study, a Spanish-Chinese parallel corpus is necessary. Currently, Spanish-Chinese parallel corpora without annotation exist (Resnik, Olsen & Diab, 1999; Rafalovitch & Dale, 2009; Eisele & Chen, 2010; Wang et al., 2013). These parallel corpora do not have any discourse information, neither syntactic information nor part-of-speech (POS) information. As we will explain in the second section, all these corpora have limitations for a discourse comparative study.

Annotated corpora with relational discourse structure under the theoretical framework of RST exist for several languages. For example, the RST Discourse Treebank for English (Carlson, Marcu & Okurowski, 2001); the Postdam Commentary Corpus (Manfred & Neumann, 2014) for German; the RST Basque Treebank (Iruskieta et al., 2013; Iruskieta, Díaz de Ilarraza & Lersundi, 2013) for Basque, and the Multilingual RST Treebank (Iruskieta, da Cunha & Taboada, 2015), a parallel corpus for English, Spanish and Basque. Regarding the languages of this research, there is an annotated RST corpus for Chinese, the Caijingpinglun Corpus (CJPL) (Yue, 2006), and an annotated RST corpus for Spanish, the RST Spanish Treebank (da Cunha, Torres-Moreno & Sierra, 2011; da Cunha et al., 2011). The RST corpora for Chinese and Spanish are very useful to carry out monolingual discourse research. Nevertheless, none of them can be used as a Spanish-Chinese parallel corpus to study translation strategies from one language to another.

This paper aims to introduce a new constructed RST Spanish-Chinese parallel corpus and the methodology we have used on it. At present, it is annotated with POS information and several queries based on morpho-syntactic information can be done online. We will annotate the rhetorical structures in our corpus under RST, taking into account that Hovy and Lavid (2010) prove that the annotation of rhetorical structure is one of the most difficult challenges for annotation works. This corpus will be used to carry out a Spanish-Chinese contrastive discourse study, and the results will be applied to translation (both human and machine translation) tasks and language learning tasks between the two languages.

In the second section, we will present the analysis of the already existing Spanish-Chinese parallel corpora. In the third section, we will introduce the methodology that we have followed to elaborate the corpus. In the fourth section, we will give the detailed information of the new elaborated Spanish-Chinese parallel corpus. In the last section, we will underline the main conclusions and look ahead at the future work.

2 Analysis of the Already Existing Spanish-Chinese Parallel Corpora

As we have mentioned in the previous section, there are currently few parallel Spanish-Chinese 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 & Dale, 2009) and (c) Sina Weibo Parallel Corpus (Wang et al., 2013). An analysis of each corpus will be given in this section, with the purpose of explaining why they are not adequate for translation and language learning purposes between Spanish and Chinese.

(a) The Holy Bible (Resnik, Olsen & 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 a comparative discourse analysis. Lastly, the texts in *Bible* are very old and cannot represent the modern language style.

(b) The United Nations Multilingual Corpus (UN) (Rafalovitch & 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, as a not direct translation corpus, we consider that the UN Spanish-Chinese subcorpus is not adequate to carry out a Spanish-Chinese discourse comparative study.

(c) 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.

3 Methodology

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 our methodology.

Firstly, in order to build the corpus and avoid the limitations of the existing corpora, 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 (the final sources used in this study will be mentioned in Section 4). 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.

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

Finally, with the aim to do several queries based on POS information (Iruskieta et al., 2013), we have created a free online interface.

4 Corpus

As we have mentioned in the previous section, finally we have selected 50 Spanish texts that have parallel Chinese translations, which have been done by professional Chinese translators. Therefore, the final corpus includes 100 texts. 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.

Table 1 shows the genre statistical information of the corpus.

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

Table 1: Genre corpus information

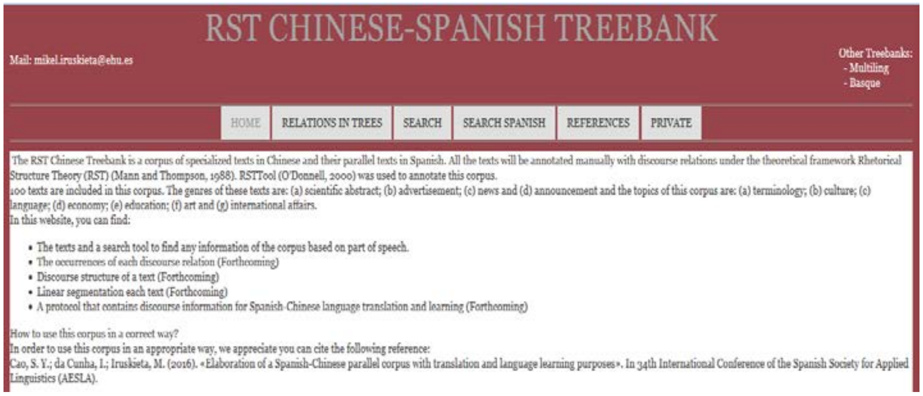
Furthermore, the texts have been divided into the following seven domains: (a) terminology, (b) culture, (c) language, (d) economy, (e) education, (f) art and (g) international affairs.

Table 2 shows the domain statistical information of the corpus.

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

Table 2: Domain corpus information

This corpus is freely open, and the access is the following link: http://ixa2.si.ehu.es/rst/zh/. Figure 1 shows a screenshot of the main page of the corpus.



The RST Chinese Treebank is a corpus of specialized texts in Chinese and their parallel texts in Spanish. All the texts will be annotated manually with discourse relations under the theoretical framework Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). RSTTool (O'Donnell, 2000) was used to annotate this corpus.
100 texts are included in this corpus. The genres of these texts are: (a) scientific abstract; (b) advertisement; (c) news and (d) announcement and the topics of this corpus are: (a) terminology; (b) culture; (c) language; (d) economy; (e) education; (f) art and (g) international affairs.
In this website, you can find:

- The texts and a search tool to find any information of the corpus based on part of speech.
- The occurrences of each discourse relation (Forthcoming)
- Discourse structure of a text (Forthcoming)
- Linear segmentation each text (Forthcoming)
- A protocol that contains discourse information for Spanish-Chinese language translation and learning (Forthcoming)

How to use this corpus in a correct way?
In order to use this corpus in an appropriate way, we appreciate you can cite the following reference:
Cao, S. Y.; da Cunha, I.; Iruskieta, M. (2016). «Elaboration of a Spanish-Chinese parallel corpus with translation and language learning purposes». In 34th International Conference of the Spanish Society for Applied Linguistics (AESLA).

Figure 1: Webpage of the new Spanish-Chinese parallel corpus

The corpus can also be consulted by means of a search tool for both Spanish and Chinese. This search tool can be used to find information based on POS categories. The search tool for Spanish is available at: http://ixa2.si.ehu.es/rst/zh/bilatzailea_spa/. The search tool for Chinese is available at: http://ixa2.si.ehu.es/rst/zh/search.php. With this tool, it is possible to search for lexical units (by lemmas, tokens and POS categories) and their corresponding textual contexts. Here we give an example by using the search tool for Chinese. For example, the adversative Chinese marker *dan* (但) (‘but’ / ‘although’) can be searched as a token with the option “exact match” (see Table 3), as a token with the same beginning (“starts with”) (see Table 4) or as a token with the same ending (“ends with”) (see Table 5), obtaining different results.

Spanish	Chinese	English translation
Aunque[6] muchos adjetivos referenciales se pueden dar en euskera a través de palabras compuestas, está claro que ese método no ofrece formas para todos los adjetivos referenciales. **[TERM 34]**	尽管巴斯克语中很多关系形容词均可通过复合词构成，**但**这种办法确定不能用于所有的关系形容词.	Although in Basque many referential adjectives can be formed by compound words, it is clear that this word-formation mechanism does not produce forms for all the referential adjectives.
Si bien este aspecto es común al progreso científico y técnico y, por lo tanto, característico de la neología terminológica, la especificidad del área tratada confiere a la neología que le es propia unas particularidades que cabe tener en cuenta. **[TERM 38]**	对于科技进步来说，这种现象的产生并不稀奇，**但**需要注意的是，介于术语新词的特点， 各领域的专业性要求又赋予了新词一定的特殊性。	While this aspect is common to scientific and technical progress and, therefore, characteristic of terminological neology, the specificity of the field analysed gives the corresponding neologism some peculiarities that should be taken into account.

Table 3: The search tool: an example of the Chinese DM *dan* (但) ('but' / 'although') obtained with the "exact match" option

Spanish	Chinese	English translation
Los adjetivos referenciales siempre derivan del nombre **pero**, muchas veces, se mezclan el origen y el carácter referencial y entran en el mismo saco los adjetivos que son completamente predicativos con los adjetivos que sin duda alguna son absolutamente referenciales. **[TERM 34]**	关系形容词通常由名词衍生而来，**但是**多数情况下会将词源和关系特征相结合，将完全作为表语使用的形容词与毫无疑问作为关系描述的形容词混为一谈 。	Relational adjectives usually derive from nouns, **but** in many cases etymological and referential factors are intermingled and, as a result, adjectives that are completely predicative are grouped together with adjectives that are absolutely referential.
Por el contrario, los que se englobarían en la clasificación de "nominal nonpredicating adjectives", **a pesar de** ser por categoría adjetivos, tendrían un comportamiento similar al de los sustantivos: linguistic difficulties/language difficulties. **[TERM 34]**	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)。	On the contrary, those which are classified as "nominal nonpredicating adjectives", **despite** belonging to the category 'adjective', have a behaviour very similar to nouns: linguistic difficulties / language Difficulties.

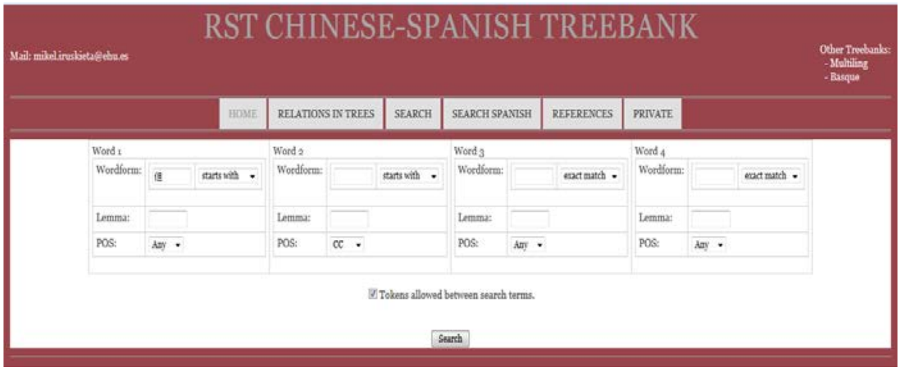
Table 4: The search tool: an example of the Chinese DM *dan* (但) ('but' / 'although') obtained with the "starts with" option

[6] In Tables 3-5 the DM is in bold in the three languages. Due to the use of translation strategies and the different usage of DMs in two languages, in some cases there is no corresponded DM in the Spanish examples.

Spanish	Chinese	English translation
Aunque muchos adjetivos referenciales se pueden dar en euskera a través de palabras compuestas, está claro que ese método no ofrece formas para todos los adjetivos referenciales. **[TERM 34]**	尽管巴斯克语中很多关系形容词均可通过复合词构成，**但**这种办法确定不能用于所有的形容词.	Although in Basque many referential adjectives can be formed by compound words, it is clear that this word-formation mechanism does not produce forms for all the referential adjectives.

Table 5: The search tool: an example of the Chinese DM *dan* (但) (‘but’ / ‘although’) obtained with the “ends with” option

Many other queries (see Figure 2) can be carried out by using the POS categories in both languages, such as: nouns, adjectives, verbs, adverbs, pronouns, and so on.



	Word 1	Word 2	Word 3	Word 4
Wordform:	但 starts with	starts with	exact match	exact match
Lemma:				
POS:	Any	CC	Any	Any

☑ Tokens allowed between search terms.

Figure 2: The Chinese search tool

5 Conclusions and Future Work

In this paper, we have introduced a new annotated Spanish-Chinese parallel corpus for the study of the rhetorical discourse structures in both languages. At the current stage, users can find aligned concordances of the two languages. Besides, we have shown the online search interface of the corpus, which has been annotated automatically with POS information.

At present, the corpus is being annotated with the RSTTool interface (O’Donnell, 2000). We will annotate the 100 parallel texts manually in order to obtain equivalent discourse structures between Spanish and Chinese. The whole corpus and the rhetorical information will be made available once we finish the annotation work.

Acknowledgements

This work has been supported by a Ramón y Cajal contract (RYC-2014-16935) and the following research projects: i) APLE 2 (FFI2009-12188-C05-01) of the Institute for Applied Linguistics (IULA, UPF), ii) Turner (TIN2015-65308-C5-1-R) and iii) Elkarola (KK-2015/00098: Eusko Jaurlaritza / Gobierno Vasco). We also give our acknowledgements to Esther Miranda and Kike Fernandez because of their efforts on the construction of the RST Chinese Treebank (RCTB) web page.

References

Carlson, L., Marcu, D., & Okurowski, M. E. (2011). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 1-10.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 04)*, 239-242.

Costa-jussà, M. R., Henríquez, C. A., & Banchs, R. E. (2004). Evaluation Indirect Stretesgies for Chinese-Spanish Statistical Machine Translation. *Journal of Artificial Intelligence Research*, *45*, 761-780.

da Cunha, I., Torres-Moreno, J. M., & Sierra, G. (2011). On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop at ACL 2011*, 1-10.

da Cunha, I., Torres-Moreno, J. M., Sierra, G., Cabrera-Diego, L. A., Castro, B. G., & Rolland, J. M. (2011). The RST Spanish Treebank On-line Interface. In *Proceedings of Recent Advances in Natural Language Processing* (*RANLP*), 698-703.

Hovy, E., & Lavid, J. (2010). Toward a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal Of Translation*, *22*(1), 13-36.

Iruskieta, M., Aranzabe, M. J., Diaz de Ilarraza, A., Gonzalez-Dios, I., Lersundi, M., & Lopez de Lacalle, O. (2013). The RST Basque TreeBank: an online search interface to check rhetorical relations. In *Proceedings of IV Workshop A RST e os Estudos do Texto*, 40-49.

Iruskieta, M., da Cunha, I. & 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., Díaz de Illarraza, A., & Lersundi, M. (2013). Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory* (*CLLT*), *11*(2), 303-334.

Levy, R., & Manning, C. D. (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.

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

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.

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.

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

Rafalovitch, A., & Dale, R. (2009). United Nations general assembly resolutions: A six-languages parallel corpus, In *Proceedings of Machine Translation Summit XII*, 292-299.

Resnik P., Olsen, M. B. & Diab, M. (1999). The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, *33*(1-2), 129-153.

Stede, M., & Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation* (*LREC 14*), 925-929.

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

Wang, L., Guang, X., Dyer, C., Black A. & Trancoso, I. (2013). Mircoblogs as Parallel Corpora In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL 2013*), 176-186.

Yue, M. (2006). *Annotation and Analysis of Chinese Financial News Commentaries in terms of Rhetorical Structure*. PhD thesis, Beijing: Communication University of China.