# Discourse Segmentation for Building a RST Chinese Treebank

**Shuyuan Cao**
Universitat Pompeu Fabra (UPF)
shuyuan.cao@hotmail.com

**Nianwen Xue**
Brandeis University
xuen@brandeis.edu

**Iria da Cunha**
Universidad Nacional de
Educación a Distancia (UNED)
iriad@flog.uned.es

**Mikel Iruskieta**
University of Basque Country
(UPV/EHU)
mikel.iruskieta@ehu.eus

**Chuan Wang**
Brandeis University
cwang24@brandeis.edu

## Abstract

Corpus-based discourse analysis of Chinese, as the most spoken language in the world, could be useful for language learning and translation studies. We present here the development of the first free open access Chinese discourse segmented corpus following RST, which can help in the evaluation of automatic segmentation systems and in the development of rhetorical parsers, among other tasks. Our research includes six stages. First, we compile different texts to include in the corpus. Second, we establish discourse segmentation criteria for Chinese. Third, two annotators segment the texts following these rules. Fourth, we calculate the segmentation agreement with Kappa and we analyze the disagreements, including the annotation errors. Fifth, we improve our segmentation criteria. Finally, we elaborate the gold standard discourse segmentation for Chinese, which can be consulted online.

## 1 Introduction

The emphasis on the idea that discourse information may be useful for Natural Language Processing (NLP) has been increasingly discussed. Discourse information and discourse-based studies are crucial for many NLP tasks (Zhou et al., 2014), such as machine translation (MT) and language learning.

Segmentation is a crucial step of discourse analysis, since it can affect the result of the relational discourse structure. Moreover, discourse segmentation can be useful for different NLP tasks, for instance, the evaluation of automatic segmentation systems, and the development of discourse parsers and automatic summarizers.

Corpus-based research is another important aspect for NLP tasks. As Wu (2014) indicates, corpora offer a large amount of language information in a quick and effective way. Corpus-based approach has been applied to different NLP tasks, such as information retrieval, parsing and machine translation (MT), among others.

Chinese is the world's most spoken language and occupies an important position in the NLP research field. However, corpus-based studies with discourse information for Chinese are still few, especially for Chinese discourse segmentation. This paper aims to present the first accessible segmented Chinese corpus according to RST and enriched with part-of-speech (POS) information.

In the second section, we introduce the theoretical framework of this study. In the third section, we discuss some related works. In the fourth section, we present the detailed information of our corpus. In the fifth section, we explain the methodology for elaborating the segmentation criteria. In the sixth section, we show results and limitations of this work. In the seventh section, we show our final segmentation criteria and present an error analysis. Finally, conclusions and future work are outlined in the last section.

*Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms*, pages 73–81,
Santiago de Compostela, Spain, September 4 2017. ©2017 Association for Computational Linguistics

## 2 Theoretical Framework

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a theory that was created especially for discourse analysis and it has been selected as the theoretical framework of this work. 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, da Cunha and Taboada, 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 hold between EDUs, which can be Nuclei or Satellites. Satellites offer additional information about nuclei. 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. 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

### 3.1 RST Based Discourse Segmentation

On the one hand, several corpora for different languages have been annotated under RST. Authors of these corpora have established their own segmentation criteria for different discourse analysis tasks. Some of these corpora are: (i) for English, the RST Discourse Treebank (Carlson, Marcu and Okurowski, 2001)[1] and the Discourse Relations Reference Corpus (Taboada and Renkema, 2008)[2]; (ii) for German, the Potsdam Commentary Corpus (Stede and Neumann, 2014)[3]; (iii) for Spanish, the RST Spanish Treebank (da Cunha, Torres-Moreno and Sierra, 2011; da Cunha et al., 2011)[4]; (iv) for Basque, the RST Basque Treebank (Iruskieta et al., 2013[5]; (v) for Portuguese, the CorpusTCC (Pardo, Nunes and Rino, 2008) and *Rhetalho* (Pardo and Seno, 2005)[6]; (vi) for Spanish, Basque and English, the Multilingual RST Treebank (Iruskieta, da Cunha and Taboada, 2015)[7].

On the other hand, some available discourse segmentation systems based on RST exist. For example: i) for English (Tofiloski, Brooke and Taboada, 2009)[8], ii) for Spanish (da Cunha et al., 2012)[9], and iii) for Basque (Iruskieta and Zapirain, 2015) [10].

### 3.2 Discourse Segmentation for Chinese

Few works focus on the Chinese segmentation from the discourse level. The Penn Chinese Treebank (Xue, 2005) is especially designed for Chinese discourse analysis with the Penn Discourse TreeBank (PDTB) (Miltsakaki et al. 2004) style. In this work, segmentation criteria are based on connectives and different types of conjunctions. Under RST, there are three works that use form-based criteria that based on punctuation marks to elaborate segmentation rules for Chinese (Yue, 2006; Qiu, 2010; Li, Feng and Zhou 2013).

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

[1] https://catalog.ldc.upenn.edu/LDC2002T07 [Last consulted: 06 of July of 2017]

[2] http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html [Last consulted: 06 of July of 2017]

[3] http://angcl.ling.uni-potsdam.de/resources/pcc.html [Last consulted: 06 of July of 2017]

[4] http://corpus.iingen.unam.mx/rst/citar.html [Last consulted: 06 of July of 2017]

[5] http://ixa2.si.ehu.es/diskurtsoa/en/ [Last consulted: 06 of July of 2016]

[6] http://www.icmc.usp.br/~taspardo/projects.htm [Last consulted: 06 of July of 2017]

[7] http://ixa2.si.ehu.es/rst/ [Last consulted: 06 of July of 2017]

[8] https://www.sfu.ca/~mtaboada/SLSeg.html [Last consulted: 06 of July of 2017]

[9] http://dev.termwatch.es/esj/DiSeg/WebDiSeg/ [Last consulted: 06 of July of 2017]

[10] http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl [Last consulted: 06 of July of 2017]

Previous segmentation criteria were based on linguistic form, but our segmentation criteria for Chinese are also based in linguistic function.

## 4 Research Corpus

Complexity of discourse structure and heterogeneity are the main characteristics taken into account for the corpus development. The specific considerations are the following: (a) texts with different sizes (between 100 and 2,000 words), (b) specialized texts and non-specialized texts, (c) texts from different domains, (d) texts from different genres, (e) texts from different original publications, and (f) texts from different authors.

Based on the mentioned aspects, finally, we have selected 50 Chinese texts to form our research corpus. The genres of the texts are four: (a) abstracts of research papers, (b) news, (c) advertisements, and (d) announcements. The longest text of the corpus contains 1,774 words and the shortest one contains 111 words. Table 1 shows the genre statistics of the corpus.

The 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.

The corpus includes texts related to seven domains: (a) terminology (15 texts), (b) culture (6 texts), (c) language (8 texts), (d) economy (7 texts), (e) education (4 texts), (f) art (5 texts), and (g) international affairs (5 texts).

The corpus was enriched automatically with POS information by using the Stanford parser (Levy and Manning, 2003) for Chinese.

Finally, we have created an online interface to access the research corpus: `http://ixa2.si.ehu.es/rst/zh/`. Users can search POS information[11] and discourse segments of each text in the research corpus. Moreover, users can also download the texts of the corpus.

| Genre | Texts | Original publication |
|---|---|---|
| Abstract of research paper | 15 | International Conference about Terminology (1997) |
| News | 15 | Shanghai Miguel Cervantes Library, Chamber of Commerce and Investment of China in Spain, Spain Embassy in Beijing, Confucius Institute Foundation in Barcelona |
| Advertisement | 13 | Shanghai Miguel Cervantes Library, Spain-China Council Foundation, Beijing Cervantes Institute, Granada Confucius Institute |
| Announcement | 7 | Spain Embassy in Beijing, Confucius Institute Foundation in Barcelona, Beijing Cervantes Institute |
| Total | 50 | |

**Table 1:** Corpus source information

## 5 Methodology

First of all, we elaborate a preliminary discourse segmentation criteria proposal for Chinese based on linguistic function (the function of the syntactic components) and linguistic form (punctuation category and verbs). We have not considered the meaning (of any coherence relation between propositions) to segment EDUs to avoid circularity in the annotation process. For the function and form perspective, we adopt the segmentation criteria from Iruskieta, da Cunha and Taboada (2015).

The following segmentation criteria are used in out work:

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

*(Ex.1) Text name*: FCEC1

*Text*: [亲爱的朋友们，] [...]

*English*: [Dear friends,] [...]

*Explanation*: The Chinese passage starts with a greeting, it is followed by a comma and there is a line break.

[11] For more detailed information about the POS information about the corpus, consult Cao, da Cunha and Iruskieta (2016) and Cao, da Cunha and Iruskieta (2017).

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

*(Ex.2) Text name*: ICP4

*Text*: [塞万提斯学院正式教师职位招聘在西班牙媒体上公布。] [同时也在塞万提斯学院网站发布信息。

*English*: [Cervantes Institute official professor recruitment notice publishes on Spanish media.] [Meanwhile, also publishes on the Cervantes Institute webpage.]

*Explanation*: After the word "*gongbu*" (公布) ('publish'), there is a period, followed by another sentence.

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

*(Ex.3) Text name*: TERM34

*Text*: [区分界限在哪里？] [区分表语及非表语的关键在哪里？] [涉及文字关系、背景联系、物主关系还是其它方面？]

*English*: [Distinguish boundary in where?] [Distinguish predicative and non-predicative of key in where?] [About characters relation, background relation, possessive relation or other aspect?]

*Explanation*: At the end of each sentence, there is a question mark.

- Other EDUs should have a main verb or an adjunct verb phrase.[12] This is a basic segmentation criterion and segmentation criteria bellow should follow this rule.

*(Ex.4) Text name*: CCICE3

*Text*: [10 月份，西班牙财政部共**筹集** 143.99 亿欧元，共拍卖国债四次。]

*English*: [The month of October the Treasury **raised** 14.399 millions in four issues.][13]

*Explanation*: The Chinese word "*chouji*" (筹集) is a verb and means 'raise' in English.

- Discourse Marker (DM) [14], verb and comma. If there is a DM at the beginning of a sentence and, this sentence is divided into two parts by a comma (each one including a verb), both parts are considered independent EDUs.

*(Ex.5) Text name*: TERM31

*Text*: [**由于**经常使用词法句型模式，] [用以分析文本或者至少说明性略语较为合适。]

*English*: [**Due to** often uses morph-syntax models,] [to analyze texts or at least illustrative abbreviations.]

*Explanation*: The Chinese DM "*youyu*" (由于) ('due to') is placed at the beginning of the first EDU, and a comma is included in the sentence. Besides, the first EDU includes the Chinese verb "*shiyong*" (使用) ('use'), while the second EDU includes the verb "*fenxi*" (分析) ('analyze').

*(Ex.6) Text name*: TERM19

*Text*: [此时，标准不但会失效，] [**而且**也不能发挥作用。]

*English*: [In this condition, standardization not only ceases to be effective,] [**but also** could not play its role.]

*Explanation*: The Chinese DM "*er*" (而且) ('but also') appears after a comma in the sentence. In addition, verbs are included in both EDUs: "*shixiao*" (失效) ('lose effectiveness') in the first EDU, and "*fahui*" (发挥) ('exert') in the second EDU.

- Semicolon plus adjunct verb phrase.

*(Ex.7) Text name*: TERM34

*Text*: [例如，形容词 marginal（边上的）在英语中可用于参照语和谓语，例如"边缘注释 (marginal not)"以及"边缘个案 (marginal case)"；] [相反，在"名词非表语性形容词"一类中，尽管采用了形容词的定义，但是与名词发挥的作用类似，比如：linguistic difficulties（语言上的困难）/language difficulties（语言困难）。]

*English*: [For example, adjective marginal (something besides) in English can be used referential and predicate, for example, "marginal note" and "marginal case";] [in contrast, in "noun but not predicative adjective" category, although adapts adjective definition, with noun works function similar, such as, linguistic difficulties/language difficulties]

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

[13] Here we give an English literal translation for each example in order to let the readers understand.

[14] In this work, the definition of DM that we follow is based on Portolés (2001). DMs are invariable linguistic units that depend on the following aspects: (a) distinct morph-syntactic properties, (b) semantics and pragmatics and (c) inferences that are made in the communication.

*Explanation*: A semicolon separates the text into two parts, and each EDU includes a Chinese verb: the verb "*yong*" (用) ('apply to') in the first EDU and the verb "*shiyong*" (使用) ('use') in the second EDU.

- Parenthetical and dash. Only when a parenthetical unit does not modify a noun neither an adjective and it includes a verb, it is an independent segment; if within the parenthetical unit there are coordinated parts, the coordinated parts are also segmented[15].

*(Ex.8) Text name*: TERM18

*Text*: [确实，术语数据库的设计和管理无论在理论和方法论] [ (如何表示一个术语？] [有最简单的表达方法吗？ ] [术语之间如何分类？ )] [⋯]

*English*: [Indeed, the design and management of the terminology database no matter in theory and methodology,] [(how to express a terminology?] [is there the easiest way to express?] [how to distinguish among terminologies?] ) […]

*Explanation*: The parenthetical unit does not modify its previous part; it should be an independent segment. The sentences "*ruhe biaoshi yige shuyu*?" (如何表示一个术语？) (How to express a term?), "*you zuijiandan de fangfa ma*?" (有最简单的方法吗？) (Is there the easiest way to express?) and "*shuyu zhijian ruhe fenlei*?" (术语之间如何分类？) (How to distinguish among terminologies?) include a verb and are coordinated parts in this parenthetical unit with verbs and question marks.

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

*(Ex.9) Text name*: TERM25

*Text*: […] [自 1994 年以来**我们**在德武斯特大学进行法律领域专业文件的翻译工作，] [**我们**希望能按照实际情况呈现出这些年工作中碰到的问题以及取得的成就。] […]

*English*: [From 1994 until now we in Deusto University **carry out** law campus professional document of translation works,] [we **hope** can follow real situation present these years works encounter problems and achievement] […]

*Explanation*: In the Chinese text, the two coordinated clauses include verbs ("*jinxing*" [进行] ['to carry out'] and "*xiwang*" [希望] ['hope']).

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

*(Ex.10) Text name*: BMCS5

*Text*: [现代化的交流工具（**聊天，论坛，博客，wiki 和电子邮件**），辅助学生在任何地方都与组内同伴交流互动。]

*English*: [Modern communications tools (chats, forums, blogs, wiki and emails), helps students in anywhere with inside group companions interact.]

*Explanation*: The names of the communication tools in the parenthetical part are appositives of the "*xiandaihua de jiaoliugongju*" (现代化的交流工具) ('modern communication tools').

- Reported speech. In this study, we do not consider reported speech as an independent EDU.
- Truncated EDUs. For the cases of truncated EDUs, we use the non-relation label of Same-unit (Carlson, Marcu and Okurowski, 2003).

## 6 Result

In this work, we use Cohen Kappa to measure inter-annotator agreement between the two corpus annotators (A1 and A2). Previous works use Kappa to measure the agreement between two annotators in RST discourse segmentation (Iruskieta, Diaz de Ilarraza and Lersundi 2015). Kappa calculates the agreement between annotators as:

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

where (A) represents the current observed agreement, and P(E) represents chance agreement. Kappa was calculated by considering titles, parentheses, and verbs, as EDUs candidates. Table 2 includes

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

the statistics used to measure the agreement between both annotators.

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

| Annotator | | A2 | | Total |
|---|---|---|---|---|
| | | Yes | No | |
| A1 | Yes | 765 | 101 | 866 |
| | No | 204 | 1888 | 2092 |
| Total | | 969 | 1989 | 2958 |

**Table 2:** Segmentation cross tabulation

Table 3 includes the Kappa agreement results regarding each part of the corpus. The highest agreement between both annotators is 0.815, and the lowest agreement is 0.616. The agreement for the whole corpus is 0.76, which means the preliminary segmentation criteria are reliable for Chinese.

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

**Table 3:** K results regarding each part of the corpus

## 7 Analysis of Corpus Annotation

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

- **Title**

A1: [2.] [术语构建] (×)

[2.] [Terminology construction]

A2: [ 2. 术语构建] (√)[16]

[2. Terminology construction]

*Analysis*: A1 has divided the title into two parts due to the period. However, we do not segment any element in a title or subtitle.

- **Comma + DM + verb**

A1: [这些内容不仅丰富了术语内容，] [**同时**还引起了一些术语基本定义的争论。] (√)

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

A2: [这些内容不仅丰富了术语内容，**同时**还引起了一些术语基本定义的争论。] (×)

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

*Analysis*: A1 has divided the sentence into two parts due to the comma. This segmentation is correct, because the discourse marker "*tongshi*" (同时) ('meanwhile') appears after the coma. Besides, the two parts have the same subject, and there is a verb "*fengfu*" (丰富) ('enrich') in the first EDU and another verb "*yinqi*" (引起) ('cause') in the second EDU.

- **Colon**

A1: [各种语言中唯一一致的命名参照物的情况是：] [术语均从英语中来。] (√)

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

A2: [各种语言中唯一一致的命名参照物的情况是：术语均从英语中来。] (×)

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

*Analysis*: A1 has divided the sentence into two parts due to the colon. In the preliminary version of segmentation criteria, colon was not considered; therefore, there is a disagreement regarding this punctuation mark between both annotators. We decide to segment the part after colon, because both EDUs include verbs: "*mingming*" (命名) ('to give name') in the first EDU and "*lai*" (来) ('come') in the second EDU.

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

- **Temporal adverb clause + comma + verb clause**

A1: [当上述内容均能在同一片文章中准确描述后，] [我们便能做到建立巴斯克语的“法律论述体系”。] (√)

[**When** all the previous mentioned can be **described** in the same passage,] [we can **establish** the “legal discourse system” for Basque.]

A2: [当上述内容均能在同一片文章中准确描述后，我们便能做到建立巴斯克语的“法律论述体系”。] (×)

[**When** all the previous mentioned can be **described** in the same passage, we can **establish** the “legal discourse system” for Basque.]

*Analysis*: A1 has divided the sentence into two parts due to the comma. The temporal adverb “*dang*” (当) (‘when’) and the comma can be considered as a segmentation boundary, because both EDUs include a verb: “*miaoshu*” (描述) (‘describe’) in the first EDU and “*jianli*” (建立) (‘establish’) in the second EDU.

- **Wrong EDU without verbs**

A1: [**包括** 12 副绘画作品和 2 副达利的原创作品，] [以及 205 份杂志、报纸及宣传单。] (×)

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

A2: [**包括** 12 副绘画作品和 2 副达利的原创作品，以及 205 份杂志、报纸及宣传单。] (√)

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

*Analysis*: A1 has divided the sentence into two parts because it is a coordinated sentence. However, the segmentation of the annotator A1 is not correct because there is no verb in the second EDU. The only verb in this sentence is “*baokuo*” (包括) (‘include’).

Based on the error analysis, we have improved our segmentation criteria. Meanwhile, we give a debate between discourse experts and, taking our segmentation criteria into account, we have chosen the best segmentation option in case of disagreement.

Hence, we have created the gold standard segmented corpus for Chinese. This gold standard will be the basis for the discourse annotation of the corpus.

Table 4 shows the final criteria used for the discourse segmentation. We have divided the segmentation criteria into two types: EDU criteria and Non-EDU criteria.

| Criteria to form an EDU | Non EDU criteria |
|---|---|
| Every EDU should have an adjunct verb clause | Relative, modifying and appositive clauses |
| Paragraphs with line breaks (titles) | Reported speech |
| Period and question exclamation marks | Truncated EDUs (same-unit) |
| Comma + adjunct verb clause | |
| Semicolon + adjunct verb clause | |
| Colon + adjunct verb clause | |
| Parenthetical & dash + adjunct verb clause | |
| Coordination with two adjunct verb clauses | |

**Table 4:** Final discourse segmentation criteria

## 8 Conclusion and Future Work

In this work, we have presented the RST discourse segmentation criteria used to annotate a Chinese corpus including texts from different domains, textual genres, sources, authors and length. Two annotators have annotated the corpus and inter-annotator agreement has been measured with Kappa, obtaining adequate results. Moreover, we carry out an error analysis to obtain the final gold standard discourse segmented corpus for Chinese following RST. This corpus can be downloaded and consulted online. Users can use the search tool to find information in the corpus related to discourse segments and POS categories in Chinese.

In the future, we will carry out the annotation of the coherence RST relations of these texts, which is one of the most difficult challenges for annotation works (Hovy and Lavid, 2010).

## References

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2016. A Corpus-based Approach for Spanish-Chinese Language Learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA3)*, 97-106.

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel. 2017. Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *EPiC Series in Language and Linguistics*, 2: 315-324.

Carlson Lynn, Marcu Daniel, and Okurowski Mary Ellen. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse Dialogue*, 1-10.

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

da Cunha Iria; SanJuan, Eric; Torres-Moreno, Juan-Manuel; Lloberes, Marina; and Castellón, Irene. 2012. DiSeg 1.0: The First System for Spanish Discourse Segmentation. *Expert Systems with Applications (ESWA)*, 39(2): 1671-1678.

da Cunha Iria, Torres-Moreno Juan-Manuel, and Sierra, Gerardo. 2011. On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop at ACL'2011*, 1-10.

da Cunha Iria; Torres-Moreno Juan-Manuel; Sierra Gerardo; Cabrera-Diego Luis Adrián; Castro Rolón Brenda Gabriela; and Rolland Bartilotti Juan Miguel. 2011. The RST Spanish Treebank On-line Interface. In *Proceedings of Recent Advances in Natural Language Processing (RANLP' 2011)*, 698-703.

Fournier Chris. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL' 2013)*, 1702-1712.

Hovy Eduard, and Lavid Julia. 2010. Toward a 'Science' of Corpus Annotation: A New Methodology Challenges for Corpus Linguistics. *International Journal of Translation*, 22(1): 13-36.

Iruskieta Mikel, Aranzabe María Jesús, Diaz de Ilarraza Arantza, Gonzalez-Dios Itziar, Lersundi, Mikel and Lopez de Lacalle Oier. 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 Mikel, da Cunha Iria, and Taboada Maite. 2015. A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2): 263-309.

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

Iruskieta Mikel and Zapirain Benat. 2015. EusEduSeg: A Dependency-Based EDU Segmentation for Basque. *Procesamiento del Lenguaje Natural*, 55: 41-48.

Levy Roger and Manning Christopher. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL' 2003)*, 439-446.

Li Yancui, Feng Wenhe, and Zhou Guodong. 2012. Elementary Discourse Unit in Chinese Dsicourse Structure Analysis. *Chinese Lexical Semantics*, 7717: 186-198.

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

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

Miltsakaki Eleni, Prasad Rashmi, Joshi Aravind, and Webber Bonnie. 2004. The Penn Discourse Treebank. In *Proceedings of 4th International Conference on Language Resources and Evaluation* (*LREC' 2004*), 2237-2240.

Mayor Aingeru, Alegría Iñaki, Díaz de Llarrarza Sánchez, Labaka Goka, Lersundi Mikel, and Sarasola Kepa. 2009. Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural*, 43: 197-205.

Pardo Thiago A. S., Nunes Maria Maria das Graças V., and Rino Lucia H. M. 2008. Dizer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Lecture Notes in Artificial Intelligence*, 3171:224-234.

Pardo Thiago A. S. and Seno Eloize R. M. 2005. Rhetalho: um corpus dereferência anotado retoricamente. *Anais do V Encontro de Corpora*. São Carlos-SP, Brasil.

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

Qiu Wusong. 2010. *Jiyu xiucijiegoulilun de hanyuxinwenpinglun yupianjiegou yanjiu* (基于修辞结构理论的汉语新闻评论语篇研究 *[*Analysis of Discourse Structure in Chinese News Commentaries under Rhetorical Structure Theory*]*). Master thesis. Nanjing: Nanjing Normal University.

Sidarenka Uladzimir, Peldszus Andreas, and Stede Manfred. 2015. Discourse Segmentation of German

Texts. *Journal for Language Technology and Computational Linguistics*, 30(1): 71-98.

Stede Manfred and Neumann Arne. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the International Conference on Language Resources and Evaluation* (*LREC' 2014*), 925-929.

Taboada Maite and Renkema Jan. 2008. *Discourse Relations Reference Corpus* [Corpus]. Simon Fraser University and Tilburg University.

Tofiloski Milan, Brooke Julian, and Taboada Maite. 2009. A Syntactic and Lexical-Based Discourse Segmenter. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL' 2009),* 77–80.

Wilks Yorick. *Machine Translation: Its scope and limits*. New York: Spring.

Wu Shangyi. 2014. On Application of computer-based corpora in translation. In *Proceedings of 2nd International Conference on Computer, Electrical, and Systems Sciences, and Engineering* (*CESSE' 2014*), 173-178.

Xu Shengqin, and Li Peifeng. 2013. Recognizing Chinese Elementary Discourse Unit on Comma. In *Proceedings of International Conference on Asian Language Processing (IALP' 2013)*, 3-6.

Xue Nianwen. 2005. Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky at ACL' 2005*, 84-91.

Xue Nianwen and Yang Yaqin. 2011. Chinese sentences segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL' 2011)*, 631-635.

Yang Yaqin and Xue Nianwen. 2012. Chinese Comma Disambiguation for Discourse Analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL' 2012)*, 786-794.

Yue Ming. 2006. *Hanyu caijingpinglun de xiucijiegou biaozhu ji pianzhangyanjiu* (汉语财经评论的修辞结构标注及篇章研究 *[*Annotation and Analysis of Chinese Financial News Commentaries in terms of Rhetorical Structure*]*). PhD thesis, Beijing: Communication University of China.

Zhou Lanjun, Li Binyang, Wei Zhongyu, and Wong Kam-Fai. 2014. The CUHK Discourse Treebank for Chinese: Annotating Explicit Discourse Connectives for the Chinese Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC' 2014)*, 942-949.

Zhou Yuping and Xue Nianwen. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL' 2012)*, 69-77.

Zhou Yuping and Xue Nianwen. 2015. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2): 397-431.