Addressing the Gap: Building an RST Corpus for Japanese

Zhijie Song
Georgetown University
zs288@georgetown.edu

Abstract

This study addresses the scarcity of Rhetorical Structure Theory (RST) corpora in the Japanese language. Establishing a multi-genre Japanese RST corpus to support model training and analysis, this research explores key considerations in RST annotation, focusing on the nuances embedded in the grammatical structure of the Japanese language. In this paper I report on parsing experiments including monolingual training and multilingual training with Chinese and English data. The analysis reveals that joint training with Chinese yields superior results than English, particularly in enhancing intra-sentential level parsing.

1 Introduction

Understanding the rhetorical structure of language is pivotal for unraveling the intricate ways in which humans convey information, opinions, and narratives. The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), provides a comprehensive framework for analyzing and comprehending the organizational principles that underlie written and spoken discourse. While much progress has been made in applying RST to English and some other languages (Carlson et al., 2001; Redeker et al., 2012; Stede and Neumann, 2014; Zeldes, 2017; Toldova et al., 2017; Peng et al., 2022b), there exists a notable gap when it comes to comprehensive resources for languages with unique structures and expressive forms, such as Japanese.

Shinmori et al. (2003) applied RST to analyze the structure of Japanese patent claims, however, no research on annotation and corpus building for Japanese has been done yet. This research endeavors to bridge this gap by constructing a multi-genre RST corpus for Japanese. The Japanese language, with its distinctive syntactic structures and nuanced expressions, poses both challenges and opportunities for RST analysis. Building a rich and accurate RST corpus for Japanese holds the promise of unveiling the deeper layers of expression in this language, providing insights for linguistic studies.

This research encompasses various aspects of corpus construction, including data gathering, annotation, model training, and analysis. In this study, data are collected from three written genres: academic article, biography, and news. Following Peng et al.'s (2022b) research ideas on the Chinese RST corpus, this paper aims at alleviating the lack of training resources for hierarchical discourse parsing in Japanese. Besides, this research is also interested in analysing parsing results at intra- and inter-sentential levels, in hopes of providing new perspectives for RST research.

In section 2, I will review previous work on several issues related to the focal points of this study. Section 3 introduces the construction of the Japanese discourse treebank, focusing on some notable issues in annotation. I explain my methodology for parser training in section 4 and present the results and my analyses in section 5. Finally, conclusions are given in section 6.

2 Previous Work & Theoratical Backgrounds

2.1 Multi-grenre RST Datasets

As pointed out in Nishida and Matsumoto (2022) and Atwell et al. (2021), the challenge of handling out-of-domain samples has long been an obstacle in RST parsing. Ensuring the comprehensiveness and richness of the corpora across various genres is, therefore, of paramount importance. The Japanese corpus presented in this paper follows the construction of two multi-genre corpora: GUM (Zeldes, 2017) and GCDT (Peng et al., 2022b).

GUM stands out as a multi-genre English RST corpus. Up to its current version 9.2.0, it encompasses 12 distinct genres, making it an ideal bench-

mark for multi-genre corpora. GCDT, on the other hand, is a Chinese corpus created based on structures similar to GUM, providing valuable insights for Japanese language training and analysis in this study. Comprising 50 Chinese articles across five genres, this corpus contributes significantly to the breadth of our analysis.

2.2 Japanese Syntax

Since Japanese has not been a well-studied language in the domain of RST, there is lack of a guideline for annotating Japanese data for discourse parsing. Shinmori et al.'s (2003) discussions provide good examples for Elementary Discourse Units (EDU) segmentation and relation labeling, however, the study is restricted to structure analysis and term explanation with regard to patent claims. Therefore, referring to Japanese syntactic attributes is crucial to negotiating annotation guidelines.

Japanese, renowned for its unique syntactic structures, is a topic-prominent language characterized by the extensive use of the topic marker *wa* (は), which plays a crucial role in indicating the topic of a sentence. This feature has been extensively discussed in linguistic literature, highlighting its impact on information structure and discourse in Japanese (Kuno, 1973; Givon, 1976). A very well-known example for the distinction between a topic and a subject in Japanese is "zo-wa hana-ga nagai" (象は鼻が長い).

(1) 象は 鼻が 長い
Elephant-TOPIC nose-SUBJ long
"The elephant's nose is long"

Shibatani (1990) and Kuno (1973) clarified the distinction between the topic marked by *wa* (は) and the subject marked by *ga* (が). In the smaller subject-indicative structure (鼻が+長い), the subject is the focus of the expression "the nose is long", while the whole structure functions as a comment to the topic in the bigger topic-comment structure (象は+鼻が長い). Therefore, a literal translation that helps understanding the structure would be "(speaking of) the elephant, its nose is long". In other cases where the comment does not contain a subject-indicative structure, the topic itself usually functions as the subject. Topic marker *wa* (は) and topic-comment structures are very common in Japanese, and prominently marked in RST annotations.

Besides, the fluidity of constructing lengthy sentences in written Japanese, particularly through the utilization of the verb conjunctive form (連用形), has also garnered scholarly attention. Scholars have explored the syntactic intricacies and communicative functions of these constructions (Shibatani, 1990; Tsujimura, 2014).

Additionally, Japanese shares a common feature with Chinese (Peng et al., 2022b) as both are languages with prenominal relative clauses. In Chinese, nominal adjuncts and complements exhibit a host of differential behaviors, and the observed differences can be elegantly explained by reference to their differences in syntactic structure and the corresponding semantic composition procedure (Huang, 2016). This is also true with Japanese.

These studies contribute significantly to our understanding of Japanese syntax, as well as providing valuable references for the establishment of annotation standards in Japanese.

2.3 Multilingual RST Parsers

Multilingual RST parsers play a crucial role in the task of RST parsing, which involves constructing a labeled tree structure for a document by merging a sequence of gold or predicted EDUs. The inherent shared unlabeled constituent tree structure across diverse RST datasets provides an opportunity for effective joint training, especially in achieving State-of-the-Art (SOTA) results across various languages (Peng et al., 2022b).

Guided by the principle that more prominent units should function as nuclei to less prominent satellite units, the strategy of multilingual joint training has exhibited significant success in enhancing parsing performance. Techniques such as translating EDUs across languages (Cheng and Li, 2019; Liu et al., 2020) and mapping word embeddings into a unified space (Braud et al., 2017; Iruskieta and Braud, 2019; Liu et al., 2020, 2021) are commonly deployed for encoding EDUs across different languages during joint training (Peng et al., 2022b).

Liu et al. (2021) introduced a cutting-edge multilingual RST parser featuring a pointer-network decoder for top-down depth-first span splitting. Leveraging the multilingual xlm-roberta-base (Conneau et al., 2020), this model underwent joint training with six languages: English, Portuguese, Spanish, German, Dutch, and Basque. Following the methodology of Peng et al. (2022b), I employ the parser introduced by Liu et al. (2021) for monolingual training on my Japanese corpus and joint

Genre	#Docs	#Tokens	#EDUs	Source
Academic	10	14420	1164	National Institute of Informatics (国立情報学研究所)
Bio	10	16043	1814	Wekipedia (ウィキペディア)
News	10	11042	1121	The Nikkei (日本経済新聞)
Total	30	41505	4099	

Table 1: Summary of the Data

training with GCDT and GUM.

2.4 Intra- and Inter-sentential Level Differences

It has been noticed that RST parsing at the macro level is more challenging than at the micro level. Feng and Hirst (2012) first employed two levels of granularity in a document, i.e., intra- and inter-multi-sentence parsing models. Joty et al. (2013) introduced the CKY algorithm to obtain optimal trees with two levels of granularity in a document. They argued that distinguishing between these two conditions can result in more effective parsing, and created intra- and inter-sentential parsers separately and combined them to perform document-level discourse analysis. They further introduced CODRA (Joty et al., 2015), a complete probabilistic discriminative framework for performing rhetorical analysis in the RST framework. CODRA's discourse parser applies an optimal parsing algorithm for both intra- and multi-sentential parsing. Liang et al. (2020) implied that there are differences in the distribution and sense labeling of intra-sentential and inter-sentential implicit relations. Intra-sentential implicits have a more even distribution of senses, while inter-sentential implicits are more unequally distributed, with a few senses covering the majority of cases. This unequal distribution can lead to the model favoring the majority class and ignoring minority classes. Furthermore, Peng (2023) conducted studies on the macro cross-paragraph discourse structure in RST parsing and treebanking for Chinese and English. He examined the association between paragraphs and RST from the perspective of analyzing segmentation differences and intra- versus inter-paragraph relation distribution across genres and corpora.

The studies above show the importance of distinguishing between discourse parsing within sentences and across sentences. This paper is especially interested in training performances at macro and micro levels, as well as intra- and inter-sentential relation type differences across languages.

3 Building RST Treebank for Japanese

3.1 Data collection

This study creates a multi-genre RST dataset in modern Japanese. Following the design of GUM (Zeldes, 2017) and GCDT (Peng et al., 2022b), my Japanese dataset (JP) [1] selects 30 documents, with 10 in each of the three written genres: academic articles, biography, and news. The dataset covers 41.5K tokens and 4.1K EDUs, with an average count of about 1.4k tokens and 1366 EDUs in one document, trying to be parallel to the sizes of GUM and GCDT. In order to achieve a reasonable length, certain documents are excerpts taken from longer original texts. Document sizes and sources are shown in Table 1.

3.2 EDU Segmentation and Relation Annotation

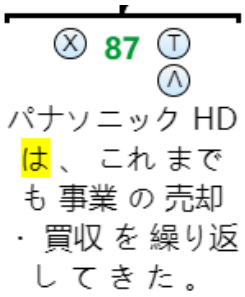
The EDU segmentation and relation annotation for the Japanese treebank in this study were done using rstWeb (Zeldes, 2016). Consistent with GUM and GCDT, I use the enhanced two-level relation labels with 15 coarse and 32 fine-grained relations.

Due to the absence of literature on Japanese RST annotations, discussing annotation guidelines is crucial. I refer to the widely accepted guidelines for English (Carlson and Marcu, 2001) and GCDT's Chinese manual (Peng et al., 2022a) to negotiation standards for Japanese. This paper aims to initiate this discussion by introducing some notable points about EDU segmentation and relation annotation criteria based on various syntactic structures in Japanese.

The topic-comment structure is a significant source of multi-nuclear relations, especially when the comment section is complex. In principle, EDU segmentation is not expected between the topic and the comment, since the basic EDU is a clause

[1] Source texts and annotations are available at https://github.com/Zhijie-Song/Japanese-Discourse-Treebank.git

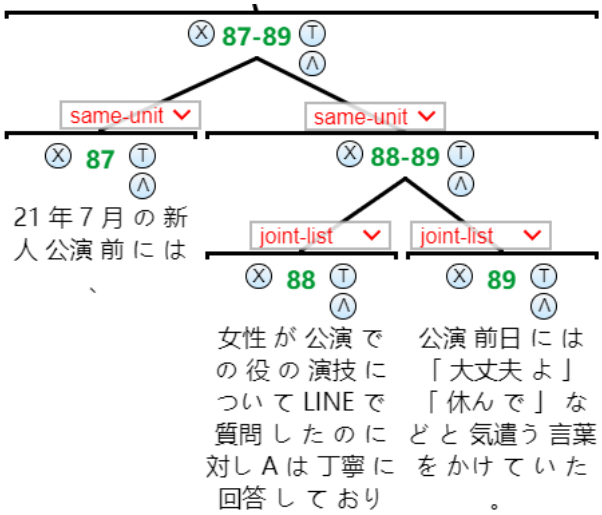
(Carlson and Marcu, 2001), and the topic does not contain indicatives (except for embedded EDUs) as shown below.



パナソニックHDは、これまでも事業の売却・買収を繰り返してきた。

”Panasonic HD has repeatedly sold and acquired businesses.” (from document News_Panasonic)

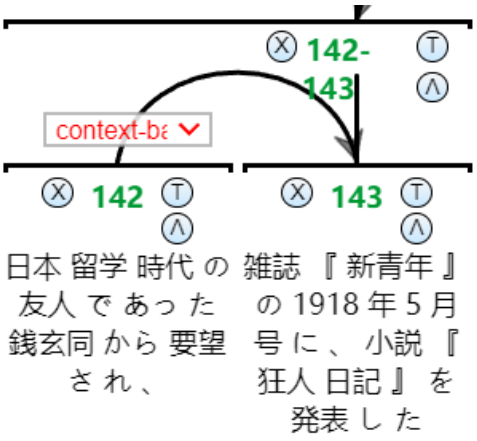
However, the comment section is usually complicated. When the comment is segmented because of embedded discourse units or multi-nucleus structures, the ”same-units” relation is needed to link the topic and the comment. There are plenty of combinations of a simple topic and a complicated comment.



21年7月の新人公演前には、女性が公演での演技についてLINEで質もしたのに対しAは丁寧に回答しており、公演前日には「大丈夫よ」「やすんで」などと気遣う言葉をかけていた。

”Before the rookie performance in July 2021, when the woman inquired about her acting in the performance via LINE, A responded politely, and on the day before the performance, A used concerning words such as ’It’s okay’ and ’Rest well’.” (from document News_Takarazuka)

In the example above, the topic marked by *wa* (は) is a temporal prepositional phrase ”Before the rookie performance in July 2021”, while the comment section consists of two clauses — ”when the woman inquired about her acting in the performance via LINE, A responded politely” and ”on the day before the performance, A used concerning words such as ’It’s okay’ and ’Rest well’” — connected by the verb conjunctive form おり. The two events form a multi-nucleus joint structure, while both contributing to commenting on the topic ”Before the rookie performance in July 2021”; therefore, they are considered within the same EDU. This is also an example of how verb conjunctive forms construct multi-nucleus relations; however, their usage is flexible and is able to build nucleus-satellite relations, as shown below.

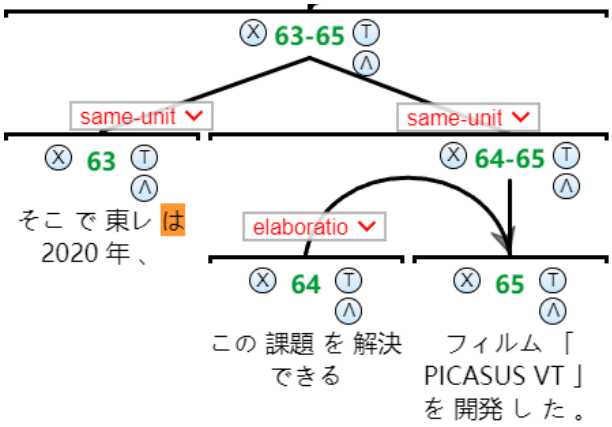


日本留学時代の友人であった銭玄同から要望され、雑誌「新青年」の1918年5月号に、小説「狂人日記」を発表した。

”At the request of Qian Xuantong, his friend during study abroad days in Japan, he published the novel ’Diary of a Madman’ in the May 1918 issue of the magazine *New Youth*.” (from document Bio_Luxun)

The verb conjunctive form 要望され connects two parts, each of which represents a grammatically complete expression. However, the first part, literally meaning "he was made a request by Qian Xuantong, his friend during study abroad days in Japan", does not convey information that is as important as the second part does, since it merely provides background information (at what situation he published the novel) for the content in the second part. Since a nucleus represents the more salient or essential piece of information in the relation, while a satellite indicates supporting or background information (Carlson and Marcu, 2001), this sentence should be annotated as context-background.

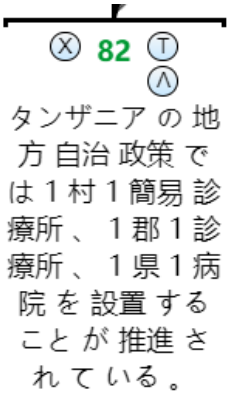
Both Carlson and Marcu (2001) and Peng et al. (2022a) considered it a principle to treat relative clauses as embedded discourse units. Peng et al. (2022a) pointed out that relative clauses in Chinese are prenominal, usually constructed using combinations of same-unit + elaboration-attribute. This is also often the case in Japanese, as shown below.



そこで東レは2020年、この課題を解決できるフィルム「PICASUS VT」を開発した。

"Therefore, in 2020, Toray developed 'PICASUS VT', a film that can solve this problem." (from document News_EV)

In this example, an attributive clause "that can solve this problem" is inserted into the middle of the sentence to modify "'PICASUS VT', a film", separating the sentence into two parts. Japanese prenominal relative clauses are constructed using the verb dictionary form (辞書形) or the past tense ta-form (タ形). However, there are exceptions where the syntax forms a relative clause that cannot be considered as an EDU.



タンザニアの地方自治政策では1村1簡易診療所、1郡1診療所、1県1病院を設置することが推進されている。

"In Tanzania's local governance policy, the establishment of 'one village, one simple clinic; one district, one clinic; one region, one hospital' is being promoted." (from document Academic_Tanzania)

1村1簡易診療所、1郡1診療所、1県1病院を設置する"establish the policy of 'one village, one simple clinic; one district, one clinic; one region, one hospital'" is not considered an attributive clause to こと(thing) even though it grammatically modifies it, because こと represents a subjective clause, and clauses that are subjects or objects of a main verb are not treated as EDUs (Carlson and Marcu, 2001).

Based on the examples discussed above, these are some noteworthy points summarized during the annotation process. However, we still need to establish more comprehensive annotation guidelines for Japanese, leaving room for future research.

4 Experiments

Using the SOTA multilingual parser DMRST (Liu et al., 2021), this study performs monolingual training tasks on the Japanese corpus (JP), as well as multilingual training combined with GCDT or GUM, testing on each corpus separately.

4.1 Datasets

This paper draws upon ten articles from each of the three same written genres (academic, bio, news) in

Corpus	Embedding	Span	Nuc	Rel
JP	*rinna/japanese-roberta-base*	68.04	50.69	45.73
	xlm-roberta-base	70.52	53.17	45.73

Table 2: Monolingual Experiments on Japanese Corpus

Experiment	Embedding	Span	Nuc	Rel
JP + GCDT (Test on JP)	*xlm-roberta-base*	70.25	54.82	48.38
JP + GCDT (Test on GCDT)	*xlm-roberta-base*	73.77	50.46	47.07
JP + GUM (Test on JP)	*xlm-roberta-base*	69.15	52.07	46.01
JP + GUM (Test on GUM)	*xlm-roberta-base*	65.06	51.20	44.28

Table 3: Multilingual Experiments on Combined Corpora

Experiment	Span	Nuc	Rel	Span	Nuc	Rel
	Intra-sentential			Inter-sentential		
JP	86.40	70.61	60.52	43.70	23.70	20.74
JP + GCDT (Test on JP)	86.90	72.93	64.19	41.79	23.88	21.64
JP + GCDT (Test on GCDT)	83.01	60.29	56.46	56.96	32.61	30.00
JP + GUM (Test on JP)	84.28	69.43	60.26	43.28	22.39	21.64
JP + GUM (Test on GUM)	80.00	65.13	56.92	43.80	31.39	26.28

Table 4: Intra- and Inter-sentential Differences

both GCDT and GUM V9.2.0. The GCDT dataset follows the division of Peng et al.'s (2022b) training, development and test sets, while JP and GUM datasets are partitioned into training, development, and test sets with an 8:1:1 ratio, using 1 document with closest token count to the average as the test set.

4.2 Metrics

Consistent to Peng et al. (2022b) and following the recommendation of Morey et al. (2017), this study adopts the 15 coarse relation classes common to both GCDT and GUM, tests on documents using gold EDU segmentation, and reports precision rate as metric for evaluating Span, Nuclearity (Nuc), and Relation (Rel) using the original Parseval .

4.3 Language Models

This study uses RoBERTa embedding for all tasks. For the monolingual training on JP, the monolingual embedding rinna/japanese-roberta-base (Zhao and Sawada, 2021) and the multilingual xlm-roberta-base model (Conneau et al., 2020) are tested. For multilingual trainings on JP+GCDT and JP+GUM, xlm-roberta-base is applied.

5 Results and Analyses

The results of monolingual and multilingual training tasks are presented in Table 2 and Table 3. Surprisingly, xlm-roberta-base slightly outperforms rinna/japanese-roberta-base in the monolingual training task. Although both models reach a P_Rel of 45.73, xlm-roberta-base achieves higher precision rates of Span and Nuc. Both multilingual joint training improve the precision rate of relation type, but do not predict on Span and Nuc better than the monolingual training. Besides, the combination of JP+GCDT completely outperforms JP+GUM on parsing Japanese data, and the predictions on GCDT are better than on GUM overall.

Considering that the corpora in this study are relatively small in size and might not be very well-balanced, it is difficult to draw a strong conclusion out of the results; however, a detailed analysis enables us to get valuable inferences.

As introduced in 2.4, many prior studies have highlighted distinctions between intra- and inter-sentential levels. Table 4 shows training results of all monolingual and multilingual experiments using xlm-roberta-base at the two different levels.

No doubt that intra-sentential precision rates are significantly higher than inter-sentential ones for every single training & test combination. Moreover, the table provides a clearer view of what

Corpus	JP	GCDT	GUM
#1	same-unit (22.73%)	joint-list (22.28%)	joint-list (17.86%)
#2	joint-list (16.63%)	same-unit (18.69%)	elaboration-additional (6.66%)
#3	elaboration-additional (9.14%)	elaboration-attribute (7.71%)	same-unit (6.53%)
#4	elaboration-attribute (9.05%)	joint-sequence (4.99%)	elaboration-attribute (6.37%)
#5	joint-sequence (8.01%)	joint-other (4.83%)	joint-other (6.20%)

Table 5: Top 5 relation types in the whole corpus

contributes to the differences between training & test combinations. At the inter-sentential level, both JP+GCDT and JP+GUM joint training experiments slightly improve the accuracy of Rel (21.64 compared to 20.74). But it is hard to tell whether any experiment improves the precision rate about Japanese data overall. At the intra-sentential level, however, multilingual experiments demonstrate prominently divergent performances on predicting Japanese data. Compared to the monolingual training results, JP+GCDT improves the precision rate of all three indexes, while JP+GUM fails to generate prediction of any index that outperforms the monolingual training. It is very likely that the joint training with Chinese could yield better results, while it is hard for the joint training with English to produce outcomes superior to the monolingual training results.

Based on the results above, this study is interested in exploring language differences related to the distribution of relation types at intra- and inter-sentential levels.

Corpus	Whole Dataset	Test Set
JP	68.86	63.09
GCDT	69.98	64.51
GUM	61.33	58.73

Table 6: Proportion of Intra-sentential Relations in different corpora

Table 6 illustrates the percentage of intra-sentential relations in three corpora. Although the proportions of all three corpora happen to be lower in the test sets than in the whole corpus, it is clear that JP and GCDT contain higher percentage of intra-sentential relations than GUM. This might contribute to the fact demonstraed in Table 3 that training on JP+GUM and testing on GUM produces the lowest precision rates, because intra-sentential relations are much easier to predict while GUM has the lowest percentage of intra-sentential relations in the text set.

I further look into relation type distributions in the three corpora. Table 5 shows the top 5 relation types in the three corpora. While joint-list takes up a high proportion in all three corpora, the percentage of same-unit is much higher in JP and GCDT than in GUM, which is the most prominent difference. The same-unit relation is the most frequent relation in JP and the second most in GCDT, making up about 20% of all relations in both corpora. This is not surprising for JP and GCDT, because the syntactic features of Japanese and Chinese, such as the topic-comment structure and the prenominal relative clause, require the same-unit label in annotation, as discussed in 3.2. This means that in terms of relation type distribution, GCDT is more similar to JP than GUM to JP. In particular, the same-unit relation, which is found only at the intra-sentential level, is frequent in both corpora. Considering that GCDT also has a higher proportion of intra-sentential relations, this very likely explains why GCDT performs better than GUM in improving intra-sentential prediction abilities on JP via joint training.

6 Conclusion

This paper presents a Japanese discourse treebank containing tokens across three writing genres, which is comparable to the existing English and Chinese RST corpora. In conclusion, this paper delved into several noteworthy aspects of Japanese RST annotation, leveraging the nuances of the language's grammatical structure, in hopes of contributing valuable insights toward the establishment of robust annotation guidelines. Through multilingual training on two corpora combinations, I observed that parser jointly trained with GCDT outperformed that with GUM when tested on JP data. Building on this, this study conducted a detailed analysis at the intra- and inter-sentential levels, revealing that JP+GCDT exhibited a comprehensive improvement in accuracy at the intra-sentential level. Given the close resemblance in

intra-sentential relation ratios and relation distributions between GCDT and JP, this finding suggests that linguistically similar languages might excel in enhancing intra-sentential predictions in joint training tasks. This study lays the groundwork for future advancements in Japanese RST annotation and underscores the importance of considering linguistic affinities in multilingual training models.

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Yi Cheng and Sujian Li. 2019. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Givon. 1976. Topic, pronoun and grammatical agreement. *Subject and Topic*, pages 149–188.

James Huang. 2016. The syntax and semantics of prenominals: Construction or composition? *Language and Linguistics*, 17(4):431–475.

Mikel Iruskieta and Chloé Braud. 2019. EusDisParser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 62–71, Minneapolis, MN. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Susumu Kuno. 1973. *The structure of the Japanese language*. MIT Press.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Siyao Peng. 2023. *Cross-Paragraph Discourse Structure in Rhetorical Structure Theory Parsing and Treebanking for Chinese and English*. Phd dissertation, Georgetown University.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022a. Chinese discourse annotation reference manual.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022b. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a Dutch text corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2820–2825, Istanbul, Turkey. European Language Resources Association (ELRA).

Masayoshi Shibatani. 1990. *The languages of Japan*. Cambridge University Press.

Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. 2003. Patent claim processing for readability: Structure analysis and term explanation. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*, PATENT '03, page 5665, USA. Association for Computational Linguistics.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics.

Natsuko Tsujimura. 2014. *An introduction to Japanese linguistics*. Wiley Blackwell.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for rhetorical structure theory and discourse relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, pages 1–5, San Diego, CA.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Tianyu Zhao and Kei Sawada. 2021. Release of pre-trained models for japanese natural language processing. *JSAI Technical Report, SIG-SLUD*, 93:169–170.

A Appendix: Label Distributions in the Japanese discourse treebank

Relation name	Proportion%
Nucleus-Satellite Relations	
elaboration-additional	9.14
elaboration-attribute	9.05
explanation-evidence	5.56
context-background	3.94
attribution-positive	3.05
adversative-concession	2.88
causal-cause	1.85
organization-heading	1.56
mode-means	1.56
context-circumstance	1.17
contingency-condition	0.77
purpose-goal	0.62
mode-manner	0.58
adversative-antithesis	0.47
evaluation-comment	0.40
organization-preparation	0.36
explanation-justify	0.32
causal-result	0.17
purpose-attribute	0.11
restatement-partial	0.08
topic-solutionhood	0.08
topic-question	0.08
organization-phatic	0.02
Multi-Nucleus Relations	
same-unit	22.73
joint-list	16.63
joint-sequence	8.01
joint-other	6.56
adversative-contrast	1.73
joint-disjunction	0.17
restatement-repetition	0.34

Table 7: Distribution of relations (30 of 32 labels were used)

B List of Documents

Genre	Academic	Bio	News
Train	Akutagawa	Beethoven	Amazon
	Iran	Friedrich	EV
	Kato	Fukuzawa	Hosoda
	R2P	Higuchi	Philippines
	Ranuki	Katsushika	Reuse
	ASchiller	Luxun	Takarazuka
	Tanzania	Murasaki	Thailand
	Translation	Sankara	Zoom
Dev	JinroBBS	Napoleon	Panasonic
Test	Maruyama	Sakamoto	DX

Table 8: Document list