CROSS-PARAGRAPH DISCOURSE STRUCTURE IN RHETORICAL STRUCTURE THEORY PARSING AND TREEBANKING FOR CHINESE AND ENGLISH

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Siyao Peng, M.S.

Washington, DC
December 1, 2022
Hierarchical discourse structures benefit Natural Language Understanding tasks, such as text summarization and sentiment analysis. Rhetorical Structure Theory (RST) is particularly significant at the macro levels, such as between paragraphs. Moreover, RST parsing at the macro level is more challenging than at the micro level, where intra-sentential relations are the easiest to identify.

Despite a dozen RST datasets available in multiple languages, a sizeable Chinese RST corpus still needs to be created. Moreover, there awaits an in-depth analysis regarding how much RST associates with macro-level structures and how much parsing performance deteriorates at the macro level and across genres. Using English and Chinese as examples, this dissertation examines how macro-level discourse relations are presented in RST and whether state-of-the-art RST parsers capture them properly.

Firstly, I create the largest Chinese RST Corpus, namely Georgetown Chinese Discourse Treebank (GCDT), an open-source treebank with 50 medium-to-long documents from five different genres. I present basic statistics and highlight annotation decisions for Mandarin Chinese. I believe this sizeable multi-genre RST corpus can promote discourse analysis and RST parsing in Chinese and across languages.

Secondly, I examine the association between paragraphs and RST trees from three aspects: a) studying how the lengths of EDU, sentence, and paragraph segments differ
across genres and corpora; b) assessing whether or not paragraphs are fully contained in RST subtrees and 3) analyzing the distribution of intra- versus inter-paragraph relations across corpora and genres.

Thirdly, I conduct parsing experiments on Chinese GCDT and English GUM using a state-of-the-art multilingual RST parser. I present both datasets’ benchmark monolingual and multilingual parsing scores and boost the performance by pretraining and automatic translation. Moreover, I show that SOTA parsers are unsatisfactory in some genres and the inter-paragraph scenario.

INDEX WORDS: Rhetorical Structure Theory, Discourse Parsing, Corpus, Paragraph, Genre, Multilingual, Chinese, English
ACKNOWLEDGMENTS

Life is a journey:

Step 1: Set a destination;
Step 2: Choose a favorite route;
Step 3: Feel free to reroute but make sure you can reach your goal in a tolerable time frame.

Looking back at my six years of doctoral journey, I cherish every moment I spent, passing by expected and unexpected scenery on the road and reaching another exhilarating starting point.

First and foremost, I would like to thank my two advisors, Prof. Amir Zeldes and Prof. Nathan Schneider, who not only mentored me through my Ph.D. studies but also empowered me to enjoy my doctoral life. From hands-on coding sessions to annotation adjudications, Prof. Zeldes tutored me with his brilliance and patience, not to mention the indispensable tools used in my dissertation, for example, rstweb and ANNIS. Prof. Nathan Schneider always answers my weird scenario questions in a timely manner and connects me with researchers from other institutes, including bridging me to a postdoc position in Germany. Prof. Nianwen (Bert) Xue – my external committee member – provided mighty theoretical support when I encountered tricky situations regarding Chinese annotations. Through my three years of coursework, I also embraced bits of knowledge learned from Prof. Paul Portner, Prof. Lourdes Ortega, Prof. Ruth Kramer, Prof. Grace Hui Yang, Dr. Hannah Sande, Dr. Corey Miller, etc.
While I usually meet my advisors for 2-3 hours a week, I spend hours with my peers daily. I shared an apartment with Yilun for 3+ years, and we listened to each other’s day-to-day accomplishments and struggles. I go to Janet if I want to double-check anything administrative or about discourse because she does the best record-keeping. Yiran is two years ahead of me, and if I hesitate about my career planning, I would ask her for stories. Undoubtedly, so many colleagues walked into my doctoral studies at Georgetown, and we always learn something from each other – Jakob, Austin, Emma, Yue, Luke, Michael, Shabnam, Shira, Lauren, Jessica, Tatsuya, Aryaman, etc. Besides my colleagues at GUCL, I adore my hangout buddies for food and fun – Yushi, Ivy, Linxi, Diana, Irene, Ullie, Yao, Fengdi, Weiting, Yarou, etc. Hoya Saxa!

It is now time to thank my best friends forever and their spare couches and bedrooms, where I spent my run-away nights – Rui in Philadelphia, Mengjia in Boulder, Hailey in Bay Area, Junze in Riverside, Dongyi in Monterey Park, Yuhang in Flagstaff, Siyu in Seattle, Erica in NYC, and Haochen in Munich. I also feel grateful for my unforgettable travel flashes – at White Sands National Park in New Mexico, Joshua Tree National Park in California, Acadia National Park in Maine, Banff National Park in Canada, Chaka Salt Lake and Daocheng Yading in China, and Fethiye in Turkey – that unconsciously reshaped my value of life.

Last but not least, ENDLESS THANKS to my dearest parents – Qinghua Peng and Zongyu Yang – who raised and financially supported me along the way and allowed me to live my life thousands of miles away from them. And closing thanks to the biggest surprise in my journey, my beloved partner Yue Yin, whom I met during a leisure trip in Munich, with whom I share thousands of romantic moments, for whom I am relocating to Germany for a postdoc at LMU with Prof. Barbara Plank.

致我生命中出现的可爱的人们，我爱你们！
TABLE OF CONTENTS

CHAPTER
1 Introduction ............................................. 1
  1.1 Background ........................................ 1
  1.2 Contributions .................................... 5
2 Related Work .......................................... 8
  2.1 English RST Corpora .............................. 10
  2.2 Discourse Datasets in Chinese ................... 12
  2.3 Genres in RST .................................... 15
  2.4 Paragraph Structures in RST ....................... 17
  2.5 Structure-Informed RST Parsing ................. 19
  2.6 Neural Network and Multilingual Discourse Parsing ... 22
  2.7 Interim Summary ................................ 26
3 GCDT: Multi-Genre Chinese RST Corpus ............... 27
  3.1 GCDT: Georgetown Chinese Discourse Treebank .... 28
  3.2 Corpus Annotation ................................. 29
    3.2.1 Document Selection .......................... 29
    3.2.2 Metadata and Structural Annotation ........... 30
    3.2.3 Tokenization ................................ 31
    3.2.4 EDU Segmentation ........................... 31
    3.2.5 Relation Annotation .......................... 38
  3.3 Data Management and Agreement .................... 45
    3.3.1 Data Types ................................ 45
    3.3.2 Data Split .................................. 46
    3.3.3 Inter-Annotator Agreement (IAA) ............. 47
  3.4 Statistics of Discourse Markers and Relations ....... 48
    3.4.1 Discourse Markers ............................ 48
    3.4.2 Discourse Relations .......................... 51
  3.5 Interim Summary ................................ 54
4 Analysis: Paragraphs in RST .......................... 55
  4.1 Aligning EDUs to Sentences and Paragraphs ........ 56
    4.1.1 English RST-DT .............................. 56
    4.1.2 English GUM ................................ 57
List of Figures

2.1 An example RST tree (wsj_1153) from RST-DT. .................... 8
2.2 Two inter-paragraph structures presented in Sporleder and Lascarides (2004). ................................................. 18
3.1 Sample relation annotations in gcdt_academic_dingzhen. ........ 40
3.2 An RST subtree with attribution-positive scoping over causal-cause from gcdt_academic_dingzhen. ......................... 42
3.3 Different attachment orders of mode-means and evaluation-comment satellites in two annotation versions of gcdt_academic_dingzhen. . . . 43
3.4 An RST subtree with two relative clauses annotated as elaboration-attribute and same-unit in gcdt_academic_dingzhen. .......... 44
4.1 The distribution of the number of tokens per document (#tokens/doc) within the three corpora. ................................. 61
4.2 The distribution of the number of paragraphs per document (#paras/doc) within the three corpora. .............................. 62
4.3 An example of properly-contained paragraphs in gum_whow_packing. 66
4.4 Four common containment violation types. ........................... 68
4.5 An example of paragraph containment violation with multiple outgoing edges pointing to the same external parent EDU from two versions of wsj_1123 annotations. ................................. 70
4.6 An example of paragraph containment violation with multiple outgoing edges pointing to the same paragraph from two versions of \textit{wsj\_1322} annotations. ............................... 72

4.7 An example of paragraph containment violation with root node in the main annotation of \textit{wsj\_1394}, and two outgoing edges in the double annotation. ............................... 74

4.8 An example of paragraph containment violation with outgoing edges in two directions in the main annotation of \textit{wsj\_1387}, and a double annotation without containment violation. ............................... 76

4.9 An example of paragraph containment violation with three outgoing edges from the second paragraph of \textit{gum\_fiction\_pixies}. ............................... 78

4.10 An example of paragraph containment violation with three outgoing edges from the second paragraph of \textit{gcdt\_whow\_hiking}. ............................... 80

5.1 The original and automatic zh\textarrowright en EDU-wise translated subtrees of \textit{gcdt\_academic\_dingzhen}. ............................... 101
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>GCDT corpus statistics.</td>
<td>28</td>
</tr>
<tr>
<td>3.2</td>
<td>GCDT’s 8-1-1 data splits, document names, and the number of tokens</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>and EDUs in the train, dev, and test sets.</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>GCDT’s genre-wise inter-annotator agreements on EDU segmentation</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>and relation annotation.</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Top 30 explicit multi-character connectives in CDTB and their</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>frequencies in GCDT across genres.</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>Distribution of 32 relations in GCDT, as well as in the same five</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>genres of GUM (GUM-5) and all twelve genres of GUM (GUM-12) as of</td>
<td></td>
</tr>
<tr>
<td></td>
<td>V8.0.0.</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>Precision, Recall, and F1 scores for high and low agreement relations</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>between the main and double annotation versions of the five GCDT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>test documents.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>An example of sentence and paragraph aligned Discourse Dependency</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Structure (DDS) for <code>gum_academic_chao.rsd</code>.</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Minimum, average and maximum numbers of tokens, EDUs, sentences,</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>and paragraphs per document in the three RST corpora: RST-DT, GUM,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and GCDT.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>The numbers of EDUs, sentences, and paragraphs per 1K tokens within</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>12 genres of GUM.</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>The numbers of EDUs, sentences, and paragraphs per 1,000 tokens</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>within five genres of GCDT.</td>
<td></td>
</tr>
</tbody>
</table>
5.6 Intra- versus inter-paragraph precision, recall and F1 on Span, Nuclearity and Relation for three models trained on GCDT, translation-augmented GCDT+GUM-5, and GCDT+GUM-12 with hfl/chinese-roberta-wwm-ext. .......................................................... 107

5.7 Intra- versus inter-paragraph precision, recall and F1 on Span, Nuclearity, and Relation for DMRST model trained on RST-DT with roberta-base and human performances when the test set is modified to be the 53 double-annotated documents. ........................................... 109
1.1 Background

In the past few years, the combination of neural architectures, language models, and unlimited access to raw text data led to success in understanding the forms of languages in computational linguistics. Alongside this, document-level Natural Language Understanding (NLU) has become a cutting-edge problem after recent dramatic improvements in word-level and sentence-level tasks (Kong et al., 2019; Kang et al., 2019). Since recognizing and generating grammatical sentences is not as challenging thanks to syntactic representations and deep neural architectures (Nivre et al., 2007; Nivre and Fang, 2017; Hewitt et al., 2020; Hewitt and Liang, 2019; Devlin et al., 2019; Conneau et al., 2020), understanding the discourse structure of a coherent text receives more attention in Natural Language Processing (NLP) research.

Among discourse frameworks, Rhetorical Structure Theory (RST, Mann and Thompson 1988) is a document-level discourse analysis formalism that assumes a single-rooted, labeled constituent tree for each document. Like the syntactic constituency structure over tokens that constructs a syntax tree, RST builds a discourse tree over Elementary Discourse Units (EDUs), mostly sentences or clauses within a document (Zeldes et al., 2019). Unlike the Penn Discourse Treebank (PDTB, Prasad et al. 2008), which mostly focuses on local discourse relations and for which more data exists in Chinese, RST builds a document tree using nested relations within and
across micro-level (e.g., sentences) and macro-level structures (e.g., paragraphs). The RST framework is linguistically meaningful and creates a hierarchical coherent structure (Mann and Thompson, 1988; Taboada and Mann, 2006b). Adapting RST into different languages and genres promotes cross-lingual and cross-genre comparisons and generalizations. For example, the realization of discourse relations (part-of-speech, syntax, word order, discourse markers, etc.) varies across genres (Liu, 2019). Taboada and Mann (2006b) argue for RST’s long-run goal of providing a conceptual coherence structure to understand texts better.

RST was initially designed to help Natural Language Generation (NLG) using its hierarchical discourse structure (Mann and Thompson, 1988). Early applications of the framework in computational linguistics include text generation and summarization, machine translation, and essay scoring (Taboada and Mann, 2006a). With growing research interests in document-level NLP tasks and the birth of neural language models in recent years, RST continues to show significant contributions in text summarization (Yoshida et al., 2014; Goyal and Eisenstein, 2016; Xu et al., 2020; Xiao et al., 2020; Huang and Kurohashi, 2021), controlled Natural Language Generation (Stevens-Guille et al., 2020; Maskharashvili et al., 2021), sentiment analysis (Bhatia et al., 2015; Markle-Hus et al., 2017; Sidarenka, 2019; Kraus and Feuerriegel, 2019; Huber and Carenini, 2020b), machine translation (Joty et al., 2014; Chen et al., 2020), fake news detection (Karimi and Tang, 2019), argument structure parsing (Musi et al., 2018; Hewett et al., 2019; Chakrabarty et al., 2019), etc. For example, in text summarization, RST’s structural importance can help content selection; in sentiment analysis, the nucleus would affect the final sentiment more than the satellite; in machine translation, connectives and discourse relations guide the arrangement of clauses within a sentence. RST incorporates both the micro and macro levels and its
conceptual design fits well with the goals of document-level natural language understanding tasks (Jia et al., 2018; Hou et al., 2020; Zhang et al., 2020).

Carlson et al. (2001) created the first RST corpus, the RST Discourse Treebank (RST-DT), by annotating 385 documents from the English Penn Treebank dataset (Marcus et al., 1993), following Mann and Thompson (1988)’s framework. In the past decades, RST-styled discourse annotations expanded to around ten languages (Zeldes et al., 2019; Liu et al., 2020), including: English (Carlson et al., 2001; Zeldes, 2017), German (Stede and Neumann, 2014), Basque (Iruskieta et al., 2013), Bangla (Das and Stede, 2018), Dutch (Redeker et al., 2012), Persian (Shahmohammadi et al., 2021), Portuguese (Cardoso et al., 2011a), Russian (Toldova et al., 2017), Spanish (da Cunha et al., 2011; Cao et al., 2018), and Chinese (Yue, 2008; Cao et al., 2018).

However, a substantial gap remains in the availability of RST datasets for non-European languages, particularly Chinese, of sufficient magnitude for training contemporary neural parsers. The Spanish-Chinese parallel corpus (Cao et al., 2018) is a small Chinese RST corpus (15K tokens) constructed for translation studies. Its EDUs are adjusted to align between Spanish and Chinese rather than staying faithful to the syntax of the individual languages. Its relation inventory is also distinct from inventories used for English corpora, as are the segmentation criteria used in the corpus, limiting its compatibility with other datasets. Another older Chinese RST corpus was reported in Yue (2008) with 97 news commentaries annotated. However, the dataset is no longer accessible or used in RST parsing or other tasks (Cao, 2018). Thus, after 30+ years of RST development, there remains a gap for a sizeable Chinese RST corpus for analysis and model training. Consequently, no RST constituent parsers are trained for Chinese and cannot yet benefit downstream tasks.

Other discourse frameworks focus on a subset of discourse features in a document (Demberg et al., 2019). The Penn Discourse Treebank (PDTB, Prasad et al. 2008)
addresses lexically-grounded discourse relations with no commitment to higher-level structures. Similarly, the Cognitive Approach to Coherence Relations (CCR, Sanders et al. 1992; Hoek et al. 2019) provides cognitively based coherence annotations but with a smaller set of discourse relations. Segmented Discourse Representation Theory (SDRT, Asher et al. 2003) is another hierarchical discourse framework, but it complicates annotations by allowing directed acyclic graphs (DAGs) and multiple parents for a node. Compared to the other discourse framework, RST’s complete analysis shows its advantage in including all segments and their relations within a document using a tree structure.

With the increasing amount of gold and silver data and the advancement of model architectures, RST parsing has received great attention in the past few years. Though lower-level discourse parsing achieved promising performance, higher-level discourse structuring is argued to be more challenging (Feng and Hirst, 2012; Jia et al., 2018). Previous literature in RST parsing showed significant improvements by featuring intra-sentential, inter-sentential, and inter-paragraph discourse structures (Feng and Hirst, 2014; Joty et al., 2013; Wang et al., 2017; Kobayashi et al., 2020). Yet, there lacks an in-depth analysis regarding how much the RST theory associates with natural paragraph boundaries and how much RST parsers vary across paragraph boundaries.

On the other hand, since RST datasets across languages share the same unlabeled constituent tree structure that more prominent units should serve as nuclei to less prominent satellite units, multilingual joint training has achieved SOTA results in RST parsing in several languages (Iruskieta and Braud, 2019; Braud et al., 2017; Cheng and Li, 2019; Liu et al., 2020, 2021b). Translating EDUs across languages (Cheng and Li, 2019; Liu et al., 2020, 2021b) and mapping word embeddings into the same space (Braud et al., 2017; Iruskieta and Braud, 2019; Liu et al., 2020, 2021b) are two common approaches to encoding EDUs across languages in joint training.
However, since there was no sizeable Chinese RST dataset, the Chinese language did not participate in the multilingual parsing regime. Moreover, combining multilingual datasets from different domains increases the generalizability of RST parsers to different genres (Iruskieta and Braud, 2019; Liu et al., 2021b). Yet, a detailed genre-wise performance analysis on the SOTA multilingual models is lacking for both English and Chinese.

1.2 Contributions

With the prosperity of neural networks and contextualized word representations, RST parsing has progressed in the past decade by differentiating intra-sentential, inter-sentential, and inter-paragraph relations in a bottom-up or top-down method. Models incorporate these sentence and paragraph boundaries as features or constraints in multi-stage parsing. Annotation schemas also use paragraph and topic structures to simplify the annotation procedure. This dissertation refers to intra-sentence and inter-sentence levels as micro-level and inter-paragraph as macro-level.

Discourse parsing at the document level (i.e., the text level) is still an unsolved problem. The larger theoretical question behind the dissertation lies in how macro-level discourse relations are presented in the RST structure and whether they are well represented in SOTA RST parsers. The long-term goal is to strengthen RST parsers’ capability in capturing document-level structures in a multi-lingual and multi-genre scenario and examine how genres and languages would affect RST annotation and parsing (Taboada and Lavid, 2003; Gruber and Muntigl, 2005; Gruber and Huemer, 2008). More specifically, achieving better generalizability across domains is as vital as scoring state-of-the-art on the widely-used but news-centric RST-DT corpus.
This dissertation aims to enhance the linguistic understanding of higher-level paragaph structures and genres in the RST framework, particularly in English and Chinese, to improve document-level RST parsing further. The dissertation is organized into six chapters. The current Chapter 1 serves as a general introduction that motivates the studies in the following chapters. Chapter 2 surveys literature in three relevant areas: Rhetorical Structure Theory and existing RST corpora, discussions regarding genres and paragraph structures in RST corpora, and previous structure-informed and multilingual RST parsing models.

Chapter 3 presents the largest Chinese RST Corpus, Georgetown Chinese Discourse Treebank (GCDT), an open-access treebank with 50 medium-to-long documents from five different genres. I present basic statistics of GCDT and highlight EDU segmentation and relation annotation decisions for Mandarin Chinese. Since designs of the GCDT corpus are motivated by the English Georgetown University Multilayer (GUM) corpus, I analyze the distribution of discourse relations in GCDT compared to GUM. This sizeable multi-genre RST corpus can promote discourse analysis and RST parsing in Chinese and across languages.

Paragraphs represent important structural information in longer documents. Chapter 4 examines the association between paragraphs and RST trees in three corpora: RST-DT, GUM, and GCDT. Alignment procedures between RST trees and paragraphs are presented to prepare for three types of paragraph analyses. Firstly, I study how the lengths of EDU, sentence, and paragraph segments differ across genres and corpora; Secondly, I conduct a containment analysis to assess whether or not paragraphs are fully contained in RST trees; Thirdly, I report the distribution of intra- versus inter-paragraph relations across corpora and genres. The quantitative and qualitative analyses benefit from double annotations in RST-DT and GCDT and the same relation inventories between GUM and GCDT.
Training with combined datasets from different languages achieved SOTA for RST parsing thanks to multilingual contextualized embeddings. Chapter 5 conducts multilingual and multi-genre parsing experiments using the newly created GCDT and similar GUM corpora. This chapter presents the benchmark parsing scores on GCDT and the updated GUM v8.0.0, experiments with pretraining and automatic translation, and shows that these techniques can further boost parsing performance. I also analyze how much more difficult macro-levels are than micro-level structures for humans and parsers and their different performances across genres.

Chapter 6 summarizes this dissertation and points out future research directions, such as expanding macro-level structures to topic segments and revising evaluation weights to simulate annotators’ disagreements. Current contributions and future directions serve the same long-term goal: to strengthen RST parsers’ capability to capture document-level structures and provide more effective representations for downstream document-level natural language understanding tasks.
Chapter 2

Related Work

Rhetorical Structure Theory (RST, Mann and Thompson 1988) presents a tree structure to capture the discourse relations of a document. The linguistically motivated RST framework is particularly beneficial to document-level Natural Language Processing (NLP) tasks, such as text summarization (Yoshida et al., 2014; Goyal and Eisenstein, 2016; Xu et al., 2020; Xiao et al., 2020; Huang and Kurohashi, 2021), controlled Natural Language Generation (NLG) (Stevens-Guille et al., 2020; Maskharashvili et al., 2021), sentiment analysis (Bhatia et al., 2015; Markle-Hus et al., 2017; Sidarenka, 2019; Kraus and Feuerriegel, 2019; Huber and Carennini, 2020b), fake news detection (Karimi and Tang, 2019), argument structure parsing (Musi et al., 2018; Hewett et al., 2019; Chakrabarty et al., 2019), etc. Thus, many RST annotation projects and parsing experiments have been conducted in the past two decades.

Figure 2.1: An example RST tree (wsj_1153) from RST-DT.
The RST schema creates a hierarchical discourse tree over Elementary Discourse Units (EDUs) within a document, as shown in Figure 2.1. EDUs (i.e., EDU_1 to EDU_10) are the minimum units in an RST tree that cannot extend across sentence boundaries, e.g., clauses or phrases. RST requires every word in a document to be included in one of the EDUs. EDUs are successively conjoined with neighbors to form larger Discourse Units (e.g., DU_1-3 or DU_1-10) while establishing discourse relations between the two joined units (e.g., attribution or Topic-Drift). RST distinguishes two types of relations: Nucleus-Satellite (NS) relations, where the Satellite (S) span (e.g., EDU_1) modifies the Nucleus (N) span (e.g., DU_2-3); and Multi-Nuclear (NN) relations, where the joined discourse units (DU_1-5 and DU_6-10) are equally prominent. The scheme gears toward the speaker’s communication goal by annotating NS relations such as Elaboration, Concession, and Cause; a small set of NN relations additionally concatenate parallel DUs in the structure, such as Joint and Contrast. Moreover, most RST trees align with document structures such as paragraphs (e.g., <P> marks paragraph boundaries in the RST-DT example), where all EDUs within each paragraph (DU_1-5, DU_6-8, and DU_9-10) form a sub-tree first and then join with EDUs or subtrees from other paragraphs. All DUs in a document accumulate to form a single-rooted discourse tree (e.g., DU_1-10).

This dissertation contributes to the understanding of RST’s macro-level structure in Chinese and English by: 1) creating the largest and multi-genre Chinese RST corpus; 2) examining paragraph containment and intra- versus inter-paragraph relation distributions across corpora and genres; and 3) experimenting with monolingual and multilingual RST parsing on Chinese and English RST corpora.

To prepare for corpus construction, paragraph analysis, and parsing experiment, this chapter lays out previous related studies from six research areas. Section 2.1 introduces the benchmark English RST datasets that serve as the reference standards for
RST datasets from other languages. Section 2.2 discusses hierarchical corpora in Mandarin Chinese and why these corpora are unsatisfactory for RST studies. Section 2.3 considers how genres differ and how genre diversity adds difficulty to discourse analysis and parsing. Section 2.4 presents the concepts for paragraph structures and why they are important in an RST Theory. Section 2.5 investigates previous models where document structures, such as paragraph boundaries, help discourse parsing. Similarly, Section 2.6 talks about how incorporating multilingual data and using multilingual embeddings boost RST parsing performances. The lack of Chinese RST data and the importance of diverse genres and document structures necessitate the data, analyses, and experiments presented in the forthcoming chapters.

2.1 English RST Corpora

The English Rhetorical Structure Theory Discourse Treebank (RST-DT) corpus is the founding RST corpus (Carlson et al., 2001). The corpus includes 205K tokens and annotates 385 Wall Street Journal newspaper documents from the Penn Treebank (Marcus et al., 1993). There are 78 fine-grained relation types (76 relations plus 2 pseudo-relations: Textual-Organization and Same-Unit) in RST-DT, and they are divided into 18 classes (16 classes plus 2 pseudo-classes). Subsequent research, particularly parsing experiments, mostly focus on the coarse-grained classes in RST-DT rather than the fine-grained ontology for practical reasons.

Two other English RST corpora also reach a considerable size for RST parsing and downstream tasks. The Georgetown University Multilayer (GUM) corpus (Zeldes, 2017) is a yearly growing corpus. As of January 2022, GUM V8.0.0 reached a similar size to RST-DT. GUM is thus slightly smaller in the token count (180K tokens) but has a larger number of discourse relation instances due to a shorter average EDU
length in tokens. GUM V8.0.0 uses an inventory of 15 first-level relation classes and 32 second-level relation labels; for example, causal-cause and causal-result are two relation labels in the causal class. GUM presents a multi-layer corpus with 12 genres, including interviews, news, travel guides, how-to guides, academic writings, biographies, fiction, forum discussions, conversations, political speeches, video logs, and textbook sections. Besides RST annotations, the corpus includes other annotation layers such as Universal Dependencies, nested entities, and coreference. The corpus is designed as a multi-genre test bed to compensate for the status quo that most discourse parsers and applications are biased towards the predominant news genre in RST-DT (Zhu et al., 2020). Moreover, the dynamic aspect of GUM makes it different to set up benchmark scores compared to other RST corpora. Chapter 5 of this dissertation publishes the first set of RST parsing performances on GUM V8.0.0.

SciDTB (Yang and Li, 2018) is another RST English corpus slightly smaller than GUM. The corpus annotates 798 abstracts from the ACL anthology with 17 coarse-grained relation types (including Same-unit and ROOT). Different from RST-DT and GUM, SciDTB is annotated in the Discourse Dependency Structure (DDS) (Morey et al., 2018), similar to other linguistic annotation schemes, such as Universal Dependencies (Nivre et al., 2016a) for syntax. The dependency structure is a more general structure that bridges different existing corpora (Morey et al., 2018). The conversion from discourse constituents to discourse dependencies seconds that dependencies are conceptually and practically more friendly to parsing and downstream tasks. Li et al. (2014b); Hirao et al. (2013) proposed algorithms to transform RST constituent trees such as RST-DT to RST dependency trees. Along with this conversion, parsing performance measures also change from four metrics: unlabeled spans (S), spans labeled with nuclearity (N), spans labeled with relation (R), and fully labeled spans (F), to two: Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS).
In a similar design to SciDTB, COVID19-DTB (Nishida and Matsumoto, 2022) annotates 300 scientific abstracts of COVID-19-related scholarly papers using the Discourse Dependency Structure. The corpus functions as a smaller out-of-domain test bench with 6,005 EDUs. It employed a set of 14 relations and discovered that the proportions of *Elaboration* and *Same-Unit* are larger in COVID19-DTB than in SciDTB since COVID-DTB is composed of longer and more complex sentences. Though SciDTB and COVID-DTB are quite similar, Nishida and Matsumoto (2022) argue that such domain adaptions, or in other cases, language adaptions, are still difficult for their SOTA bootstrapping method.

2.2 Discourse Datasets in Chinese

Chinese is a language that is quite different from English. Not only are subjects frequently dropped (Kang et al., 2019), Chinese is a topic-prominent language whose syntax emphasizes the topic-comment structure rather than the semantic subject in a subject-prominent language such as English (Li and Thompson, 1984). Moreover, connectives in Chinese are more frequently implicit. For example, 82.0% of connectives are implicit in the Chinese CDTB corpus compared to the 54.5% in English PDTB (Zhou and Xue, 2012). Regarding identifying EDU boundaries, commas may function as a full stop marker due to the late introduction of punctuation marks into Chinese (Xue and Yang, 2011). Considering these differences between the two languages, adapting the RST scheme to Chinese is a challenging task (Kang et al., 2019; Kong et al., 2019).

Yue (2008) pioneers in accommodating RST structures into Chinese by annotating 97 News Commentaries from people.cn and building the CJPL (Cai-Jing PingLun, ‘news commentaries’) corpus. The corpus uses punctuation-based
approaches (periods, question marks, exclamation marks, colons, semi-colons, etc.) to segment EDUs and includes 47 relations in the scheme. However, the dataset is no longer accessible and fails to impact the RST community (Cao, 2018).

Among the Chinese discourse corpora, the Chinese portion of the RST Spanish-Chinese Treebank is the closest to RST-DT (Cao et al., 2017, 2018; Cao and Gete, 2018). The corpus annotates 50 Spanish-English parallel documents (15K tokens) from eight sources and spans four genres: research paper abstracts, news, advertisements, and announcements. They propose segmentation criteria based on the syntax and punctuations in Chinese and sample 26 commonly-used relations from RST-DT. The dataset has two limitations. Firstly, all 50 documents are short and are translated from the Spanish originals. Secondly, the corpus prioritizes EDU alignments between Spanish-Chinese translations. Thus, EDU segmentation in the corpus is not strictly faithful to the syntactic criteria of the individual languages.

The Connective Driven Treebank (CDT or CDT-CDTB, Li et al. 2014c) receives the most attention among hierarchical Chinese treebanks for annotation and parsing experiments. CDT includes 500 news documents from the Penn Chinese Treebank (CTB) (Xue et al., 2005a) and is widely used in CDT parsing (Kong and Zhou, 2017; Zhang et al., 2020; Hung et al., 2020; Zhu et al., 2022), zero pronoun resolution (Cheng et al., 2018), etc. However, CDT differs largely from RST in that all edges are labeled with explicit or inserted connectives rather than relation labels. One crucial limitation of CDT is that connectives are highly ambiguous and can anchor different discourse relations depending on context (Pitler and Nenkova, 2009; Zhou et al., 2012; Webber et al., 2019). Moreover, CDT only builds up within paragraphs resulting in small discourse trees (4.5 EDUs/tree on average). Though CDT is prosperous in its research realm, it differs substantially from the expected structure of an RST treebank, in which EDUs are clauses with functionally motivated relation labels (e.g.,
cause or background) and are successively conjoined to form a discourse tree for an entire document.

Motivated by the different distributions of discourse relations in intra-paragraph versus inter-paragraph scenarios, the Macro-level Chinese Discourse Treebank (MCDTB, Jiang et al. 2018b) utilizes a set of 15 discourse relations to connect between paragraphs within 720 documents from the Penn Chinese Treebank 8.0 (Xue et al., 2005a). This MCDTB dataset treats each paragraph as the elementary unit and builds a discourse structure over paragraph spans. The work follows their previous research on the distinction between primary versus secondary relations (Chu et al., 2017), as well as on Li et al. (2014c)’s Connective Driven Treebank (CDT) where EDU relations are annotated only within paragraphs. In other words, Jiang et al. (2018a) and Li et al. (2014c) decompose the mainstream RST structure annotation into two levels and creates an interesting distinction between macro versus micro relations (Sporleder and Lascarides, 2004; Wang et al., 2017). The former annotates the inter-paragraph structure using 15 relation types within three categories: Coordination, Causality and Elaboration, whereas the latter establishes relations among EDUs within paragraphs using explicit or implicit discourse connectives. Unfortunately, the design of CDT and MCDTB deviates from RST’s fundamental idea of constructing a single tree for an entire document, in which the same inventory of labels is used for all nodes.

Moving beyond constituency-based discourse trees, Cheng et al. (2021) devote time and effort to create a larger and unified Discourse Dependency Structure (DDS, Hirao et al. 2013; Morey et al. 2018) collection by sourcing three major Chinese discourse corpora: SciCDTB (Cheng and Li, 2019), HIT-CDTB (Zhang et al., 2014), and CDT (Li et al., 2014c). SciCDTB follows the DDS scheme of SciDTB (Yang and Li, 2018) and annotates 108 abstracts from a Chinese NLP journal, Journal of Chinese
Information Processing (JCIP), using 17 coarse-grained and 26 fine-grained relation types. More importantly, Cheng et al. (2021) transforms two large and non-RST corpora into the DDS scheme. The original HIT-CDB follows the PDTB style. To create a fully annotated DDS, they automatically predict relations based on discourse markers on positions where discourse relations are missing. CDT sub-divides EDUs to follow the RST scheme and translates the edge labels from connective markers to relation types. They manually correct all conversions and employ the same set of 17 coarse-grained labels for the three DDS Chinese corpora. The unified conversion reaches a similar size as RST-DT and would greatly help future DDS parsing in Chinese. Their approach to scheme conversion can also shed light on the connection between RST and other discourse frameworks.

Even though DDS simplifies parsing and is more similar to other linguistic annotation schemes, such as Universal Dependencies (Nivre et al., 2016a) for syntax, dependency-style discourse annotation loses significant information on the ordering or scope of satellite attachments. For example, whether a unit with cause and attribution satellites means that both the cause and the result are attributed to someone, or that something caused an attributed statement. In other words, when multiple discourse units modify the same nucleus, the relative importance of the satellites and their scopes are ignored in DDS. Thus, the slightly more complex constituency-styled RST framework maintains its value, and there still lacks a sizeable Chinese RST corpus, particularly with complete annotations for longer documents.

2.3 Genres in RST

Within Systemic Functional Linguistics, “genre” is defined as “a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture” (Martin,
Martin (1992) further argues that genres are realized through the register to achieve their social goals. However, Mann et al. (1992) assumes homogeneity in the RST framework and believes in a unified description of text structure regardless of genres. He argues that only label frequencies diverge across scale, genre, etc., but the relational organization stays the same.

Since genre and other factors were assumed to be outside the scope of RST, Taboada (2004); Taboada and Mann (2006b) present the first analysis that grounds RST relations in the dialogue genres in English and Spanish. Her distributional analyses have motivated researchers in the past two decades to study the divergence of RST annotations across genres. Gruber and Huemer (2008) find out that the explicitness of coherence relations differs from genre to genre. For example, academic genres are more argumentative, and thus more signaling devices can be expected. Popoola (2018) concludes that a series of Elaboration relations is more common in fake book reviews, whereas Contrast is more common in real ones. Liu (2019) points out that signaling information can also be genre-specific.

Besides differences in RST annotations, genre diversity also influences RST parsing (Taboada, 2004; Imaz and Iruskieta, 2017). The longstanding RST-DT benchmark is the largest English RST corpus but only contains Wall Street Journal news. Many previous experiments reveal that parsing performances vary across domains and deteriorate largely for out-of-domain (OOD) texts (Bourgonje and Schäfer, 2019; Iruskieta and Braud, 2019; Nishida and Matsumoto, 2022). Bourgonje and Schäfer (2019) state that their models’ F-scores range from 73.00 to 94.47 across languages and genres. Results from Iruskieta and Braud (2019) also show that parsers fail to label the central units (CU) if parsing on another genre. Nishida and Matsumoto (2022) argue that though bootstrapping and discourse parsing models are assumed to be language-independent, training and evaluation data in both source and target are still needed.
for domain adaption. In particular, one should find the closest labeled source data and ensure the same annotation framework, i.e., EDUs and discourse relations. Moreover, genre diversity further affects downstream tasks. Ji and Smith (2017) demonstrate that the effectiveness of incorporating discourse parser into their text classification model highly depends on the target genres of the classification task. In their experiments, news-trained discourse parsers benefit genres such as restaurant reviews and congressional debates but harm the classification accuracy on legislative bills, a distant genre with technical legal content and highly specialized conventions. It is worth noticing that the legislative bill corpus is larger than the others, and their performance drop is certainly not caused by the lack of data but rather by genre diverseness.

2.4 Paragraph Structures in RST

RST corpora provide extensive annotations over paragraphs in an entire document even though the original homogeneous design of the RST schema has no association with holistic structures, such as sections and paragraphs (Mann et al., 1992). This section surveys recent analyses of paragraph structures and how they contribute to a better understanding of the rhetorical structure for longer documents.

Paragraphs are explicit textual breaks in a document to facilitate reading. For example, Filippova and Strube (2006) show two important linguistic features for paragraph initials: firstly, humans prefer proper nouns to pronominal references at the beginning of a paragraph; and secondly, paragraph-initial sentences often start with the theme in German rather than an agentive subject (Stark, 1988).

Inter-paragraph discourse relations are exemplified to be different from intra-paragraph ones. For example, Sporleder and Lascarides (2004) find out that the most common one in RST-DT is Elaboration taking up 37% of inter-paragraph rela-
Wang et al. (2017) show that the majority of *Topic-Comment* and *Textual-Organization* occur across paragraphs even though their occurrences are generally rare in RST-DT. Similarly, Feng and Hirst (2014) point out that *Textual-Organization* predominately occurs in upper-level discourse spans. Inkova et al. (2022) assign paragraphs as the meso-level, positioned between global and local levels of text organizations. Polakova et al. (2021) also compare intra- and inter-paragraph relations in English and Czech PDTB-styled corpora and propose to assess their equivalences in the hierarchically structured RST framework in the future.

![Diagram](image)

**Figure 2.2:** Two inter-paragraph structures presented in Sporleder and Lascarides (2004).

It is generally assumed that discourse sub-trees are formed within the same textual chunk before crossing over to form a larger structure (Nishida and Nakayama, 2020). Sporleder and Lascarides (2004) report statistics over the RST-DT dataset and show that 79% of paragraphs correspond to discourse segments, i.e., the paragraphs share the same boundaries as discourse spans. The 21% misalignments can be analyzed as two scenarios: 13% of the paragraphs (as in Figure 2.2a) are situations where DUs in Paragraph B are successively linked up to Paragraph A, rather than as an entirety; the rest 8% (as in Figure 2.2b) occur when DU_4-6 and EDU_7 in Paragraph B are
linked to two different paragraphs A, and C. Sporleder and Lascarides (2004) manually modify these RST trees by using majority voting to determine the correlation between RST and paragraph structures.

In the rest of this dissertation, I refer to the phenomenon where “EDUs in a paragraph form a proper, single-rooted sub-tree before merging with discourse units from other paragraphs” as proper paragraph containment. Namely, paragraphs are assumed to be aligned and properly contained as sub-trees in a document-level RST structure. This paragraph containment assumption is vital in many RST datasets where the annotation procedure starts with intra-paragraph relations and is followed by inter-paragraph relations (Cao et al., 2018).

2.5 Structure-Informed RST Parsing

RST Parsing is a task that merges a sequence of gold or predicted EDUs and forms a labeled tree structure for the entire document. RST parsing is essential for building document-level representations and braces downstream NLU tasks. Sporleder and Lascarides (2004) present the earliest work of structural-informed discourse parsing. Their model uses diverse features, including word co-occurrence, lexical chains, cue phrases, punctuation, and tense. They argue that cue phrases are more useful for modeling discourse relations among local units. At the same time, word co-occurrence is the most crucial factor for higher-level structures, such as paragraphs or sections.

Ji and Eisenstein (2014) use latent spaces to represent lexical and positional features of EDUs in their shift-reduced DPLP (Discourse Parsing from Linear Projection) parser. Their features capture human intuitions about discourse connectives and benefit from the preprocessing step to binarize the tree structure to be right-branching. They use all vocabularies in the stack DU to model the long paragraph-level lexical
dependencies implicitly. Their model remains the state-of-the-art baseline as of Morey et al. (2017) and still prevails among the top-pick parsers for downstream applications.

Two other widely-used strong baselines share the commonality of parsing RST trees in a multi-stage fashion (Joty et al., 2015; Feng and Hirst, 2014). Joty et al. (2013) observe that the distribution and relevant features differ for intra-sentential versus inter-sentential relations. They first parse intra-sentential relations and then build up inter-sentential ones. Both sub-models use a non-greedy Conditional Random Field in their bottom-up parsing algorithm. Their later CODRA (a COmplete probabilistic Discriminative framework for performing Rhetorical Analysis) parser includes a segmenter to build an entire parsing pipeline (Joty et al., 2015).

On the other hand, Feng and Hirst (2014) employ a greedy bottom-up algorithm with two linear-chain CRFs as local probabilistic models following Feng and Hirst (2012). However, they acknowledge that the bottom-up approach is unaware of the upper-level constituents. They thus include a post-editing step to modify the fully-built discourse tree by considering upper-level spans to remedy this issue. The depth of the discourse span can be quite an indicative factor in modeling discourse trees. In addition to post-editing, Feng and Hirst (2014) include discourse-motivated features, such as DUs’ position within a sentence or a document, the size of DUs, the presence of cue phrases, etc.

Joty et al. (2013) and Feng and Hirst (2014) show that multi-stage systems are efficient and can achieve competitive performance for document-level RST parsing. Building upon these long-lasting baselines, a few research enhanced RST parsing using paragraph information (Liu and Lapata, 2017; Wang et al., 2017, 2019; Nishida and Nakayama, 2020; Kobayashi et al., 2020; Koto et al., 2021). While lower-level discourse relations can benefit from intra-sentential syntactic features (Braud et al., 2016, 2017; Zhao and Huang, 2017; Yu et al., 2018), higher-level discourse structuring turns to
more abstract document-level features or architecture designs. Liu and Lapata (2017) follow the two-stage parsing strategy and propose a neural transition-based constituent parser based on long-range dependencies. Their Contextually Informed Discourse Parser (CIDER) learns contextual information at the intra-sentential and inter-sentential levels and captures long-distance dependencies with an LSTM method.

Subsequent research extends this two-way structural distinction to a three-way one: intra-sentential, inter-sentential, and inter-paragraph. Wang et al. (2017) present two sub-models to construct the naked tree structure using transition-based algorithms and then label the relation arcs. In their second stage of relation labeling, they use the three-way distinction and train three separate classifiers for intra-sentential, inter-sentential, and inter-paragraph relations. Wang et al. (2019) present a tensor-based approach with a three-stage bottom-up strategy using the shift-reduce algorithm. Nishida and Nakayama (2020) present an unsupervised three-stage approach using the Viterbi expectation-maximization (EM) algorithm with a margin-based criterion. Their Combinational Incremental Parsers (CIPs) incrementally build discourse trees from sentence to document levels. Their unsupervised model achieves a similar unlabelled constituent score as Feng and Hirst (2014).

Similar sub-models based on levels of granularity are also designed in a top-down approach. The top-down view captures the basic structure better because it sees the depth of the tree (Feng and Hirst, 2014). Kobayashi et al. (2020) build a top-down RST parser by decomposing documents into paragraph sub-trees, paragraphs into sentence sub-trees, and sentences into EDU sub-trees, following their unsupervised model of similar three levels of granularity (Kobayashi et al., 2019). Their document-to-paragraph-to-sentence-to-EDU model achieves a promising score on discourse constituent and dependency parsing. Koto et al. (2021) provide a conceptually more straightforward top-down approach that eliminates the decoder and reduces the search
space. They frame the task as recursively splitting a sequence of discourse units into two sub-sequences using an LSTM and pre-trained BERT embeddings (Devlin et al., 2019). They also introduce a dynamic oracle to minimize the deviation by comparing the constructed and gold trees and achieve state-of-the-art performance among all paragraph-informed models. Shen et al. (2022) further experiment with a bottom-up variation of their previous model (Koto et al., 2021) and find out that the bottom-up variant performs better on Span and Nuclearity but worse on Relation and Full scores.

In addition to boosting document-level RST parsing, the paragraph-level hierarchical structure can also help relation classification tasks on the non-hierarchical Penn Discourse Treebank (PDTB, Prasad et al. 2008) corpus. For example, Dai and Huang (2018) improve implicit discourse relation classification by introducing paragraph-level neural networks that model inter-dependencies between discourse units within a paragraph. Even though the PDTB annotation framework does not propose any hierarchical structure, the results show that context sensitivity is also crucial for classifying local discourse relations.

2.6 NEURAL NETWORK AND MULTILINGUAL DISCOURSE PARSING

In addition to the multi-stage pipeline incorporating sentence and paragraph-level information, RST parsing also benefits from other external knowledge, particularly pretrained and cross-lingual language embeddings in neural models. Recent advancements in neural networks and language models raise state-of-the-art performance on document-level RST parsing. Jia et al. (2018) realize the difficulties in parsing long-distance EDUs and exemplify that the Label Attachment Score (LAS) drops to less than 5 for dependencies with a range of 6 EDUs in RST-DT. To assimilate lexical
chains, they employ a memory network in the traditional BiLSTM to group EDUs with similar topics into the same memory slot. Zhu et al. (2020) conduct a qualitative study to examine the amount of rhetorical information represented in contextualized neural language models. They derive 24 features based on three categories: discourse relations, tree properties, and EDU features, and show that EDU and tree features are easier to learn, but relation features are challenging to probe. Among the examined language models, the BERT-based language models win over GPT-2. They also show that the automatic parser (Feng and Hirst, 2014) trained on RST-DT degrades on the out-of-domain IMDB dataset. Guz and Carenini (2020) attempt to incorporate coreference resolver into discourse parsing at different levels of their shift-reduce parser. Though coreference features do not help, their inclusion of contextualized SpanBERT embeddings promotes their model. Guz et al. (2020) further combine structural features (Joty et al., 2015) with RoBERTa string-encoding (Liu et al., 2019) in a shift-reduce parser and show that pre-training on the large-scale silver MEGA-DT corpus (Huber and Carenini, 2020a) provides performance gain in RST parsing.

In the past decades, more than ten RST datasets have emerged in multiple languages. Since RST datasets share the same unlabeled constituent tree structure based on the principle that more prominent units should serve as nuclei to less prominent satellite units, many recent types of research have focused on RST parsing for multiple languages. Combining RST datasets from multiple languages remedies the sparsity of annotated data (Liu et al., 2021b).

Zeldes et al. (2019, 2021a) organize the DISRPT 2019 and 2021 shared tasks for EDU segmentation, connective detection, and relation labeling. Muller et al. (2019) was the first to employ the contextualized BERT embedding in EDU segmentation and ranked first in the DISRPT 2019 shared task. Yu et al. (2019) design handcrafted features and implement model stacking from logistic regression to bi-LSTM-CRF to
account for large and small training data. Gessler et al. (2021)’s DisCoDisCo system wins the DISRPT 2021 shared task by inheriting linguistic features from Yu et al. (2019) and selecting the best-performing language- and task-dependent CWEs in a Transformer-based neural classifier. Ablation studies show that CWEs and instance-level features, such as position in the document, the distance between units, etc., increase model performance across datasets. They also report models’ bias caused by the unbalanced labels in the RST datasets and suggest more training data for the minority classes.

In addition to training the same model architecture but separately for different languages or corpora, combining multilingual data in training has been a promising approach in recent years and has achieved SOTA results in multilingual RST parsing in several languages. Braud et al. (2017); Iruskieta and Braud (2019) harmonize discourse treebanks across different languages, including preprocessing, tree binarization, and label set harmonization, enabling combined training on multilingual RST datasets. They include lexical, positional, and length features and use cross-lingual word representations (Levy et al., 2017; Artetxe et al., 2018) in their transition-based constituent parser borrowed from syntax. They also experiment with two data augmentation scenarios: with or without training data from the target corpus. The issue with bilingual word embedding mapping occurs in their parsing experiments, particularly between English and Basque (an isolated language more distant from English than languages such as German or Spanish), and hinders their cross-lingual approach. Even though their multilingual combined training fails to outperform monolingual models, they provide fundamentals in harmonizing multilingual discourse datasets for joint training.

Due to the limited size of labeled Chinese data, Cheng and Li (2019) apply a zero-shot method for Discourse Dependency Parsing by translating Chinese SciCDTB
into English and using an English SciDTB-trained model to parse the Chinese-to-
English translated SciCDTB texts. They consequently align the parsed tree to the
original Chinese texts. The zero-shot approach with automatic translation outper-
forms training on the small Chinese dataset, even though the results are far below
human performance. On the other hand, after the birth of multilingual contextual-
ized word embeddings, Kurfal and Östling (2021) demonstrate that the contextualized
multilingual XLM-RoBERTa models (Conneau et al., 2020) perform the best on RST
relation classification when training only on RST-DT and zero-shot testing on other
languages.

To date, Liu et al. (2020, 2021b) achieve SOTA in multilingual RST parsing
on six languages. Liu et al. (2020) demonstrate that using multilingual embeddings
and translating EDUs from all languages to English effectively maps RST datasets
across languages. Specifically, they choose XLM-RoBERTa (Conneau et al., 2020)
over BERT (Devlin et al., 2019) due to BERT’s limitations in encoding lengthy sequences.
Their improved DMRST parser (Liu et al., 2021b) further integrates EDU segmenta-
tion and applies cross-translations (i.e., converting samples from one language to all
other languages) to datasets, and increases training size and domain generalizability.

In sum, translating EDUs across languages and mapping word embeddings into the
same space are two common approaches in encoding EDUs across languages in joint
training. Models mentioned above rely heavily on the harmonized mappings (Braud
et al., 2017) due to the diverse annotation procedures in different RST datasets.
Moreover, due to the limited availability of Chinese RST data, SOTA multilingual
RST parsers have not experimented on Chinese. In this dissertation, the multilingual
training experiments detailed in Chapter 5 report SOTA parsing results on the Chi-
nese GCDT corpus (see Chapter 3), which utilizes the same set of relations as contem-
porary English RST benchmarks and avoids error propagation from label mappings.
2.7 **Interim Summary**

This chapter presents previous works relevant to the dissertation from three domains. Section 2.1 and Section 2.2 present benchmark RST datasets from English and RST-like discourse datasets in Chinese, revealing the lack of sizeable and comparable RST corpus in Mandarin Chinese. Section 2.3 and Section 2.4 illustrate two important aspects of the RST framework: genres and paragraph structures. Section 2.5 and Section 2.6 set forth baseline and SOTA RST parsers and substantiate the effectiveness of paragraph structures, contextualized word embeddings, and multilingual joint training. These gaps in previous studies lead the way to the three contributions of this dissertation: creating the largest Chinese RST dataset (Chapter 3), analyzing paragraph structures across datasets (Chapter 4), and establishing multilingual and multi-genre parsing results in English and Chinese (Chapter 5).
Chapter 3

GCDT: Multi-Genre Chinese RST Corpus

The lack of RST data in Mandarin Chinese has been a bottleneck for Chinese RST parsing for many years. Various hierarchical discourse corpora were annotated in the past decade (Cao and Gete, 2018; Li et al., 2014c; Jiang et al., 2018a; Cheng et al., 2021). Still, none build a single-rooted tree over medium to long documents (see Section 2.2 for details). As a result, Chinese is not included in recent multilingual SOTA RST parsing experiments (Braud et al., 2017; Iruskieta et al., 2019; Liu et al., 2021b, 2020).

This chapter introduces the Georgetown Chinese Discourse Treebank (GCDT, Peng et al. 2022b), the largest hierarchical discourse treebank for Mandarin Chinese in the framework of Rhetorical Structure Theory (RST). GCDT covers over 60K tokens across five genres of freely available text, using the same relation inventory as contemporary RST treebanks for English. The source texts and annotations are open-access and available at https://github.com/logan-siyao-peng/GCDT. I also provide an extensive Chinese RST annotation manual (Peng et al., 2022a) for future reference.

The GCDT dataset is presented from a few different perspectives. Section 3.1 presents basic statistics of GCDT. Section 3.2 highlights the annotation decisions in creating the GCDT corpus, including document selection, metadata annotation, tokenization, EDU segmentation, and relation annotation. Section 3.3 discusses how GCDT data is presented to the public and assesses the reliability of annotations by
inter-annotator agreements. Section 3.4 analyzes the distribution of discourse markers and relations in GCDT compared with the Penn Chinese Treebank (CDTB, Zhou and Xue 2015) and the English Georgetown University Multilayer (GUM, Zeldes 2017) corpora. Section 3.5 summarizes this chapter and promotes the GCDT dataset for multilingual and multi-genre discourse parsing and treebanking.

3.1 GCDT: Georgetown Chinese Discourse Treebank

GCDT is a new, freely available, multi-genre RST corpus for Mandarin Chinese (Peng et al., 2022b). Following the design of GUM (Zeldes, 2017), GCDT contains 50 documents, 10 from each of 5 genres that also appear in GUM: academic articles, biographies (bio), interview conversations, news, and how-to guides (whow). Table 3.1 presents the basic statistics for GCDT, including the number of documents, tokens, Elementary Discourse Units (EDUs), the source websites, and their licenses.

<table>
<thead>
<tr>
<th>Genre</th>
<th>#Docs</th>
<th>#Toks</th>
<th>#EDUs</th>
<th>Source</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>academic</td>
<td>10</td>
<td>14,168</td>
<td>2,033</td>
<td><a href="https://www.hanspub.org/">https://www.hanspub.org/</a></td>
<td>CC-BY or CC-BY-NC</td>
</tr>
<tr>
<td>bio</td>
<td>10</td>
<td>13,485</td>
<td>2,021</td>
<td><a href="https://zh.wikipedia.org/">https://zh.wikipedia.org/</a></td>
<td>CC-BY-SA</td>
</tr>
<tr>
<td>interview</td>
<td>10</td>
<td>11,464</td>
<td>1,812</td>
<td><a href="https://zh.wikinews.org/">https://zh.wikinews.org/</a></td>
<td>CC-BY</td>
</tr>
<tr>
<td>news</td>
<td>10</td>
<td>11,249</td>
<td>1,652</td>
<td><a href="https://zh.wikinews.org/">https://zh.wikinews.org/</a></td>
<td>CC-BY</td>
</tr>
<tr>
<td>whow</td>
<td>10</td>
<td>12,539</td>
<td>2,199</td>
<td><a href="https://zh.wikihow.com/">https://zh.wikihow.com/</a></td>
<td>CC-BY-NC-SA</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>62,905</td>
<td>9,717</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unlike existing Chinese discourse corpora, GCDT focuses on building larger discourse trees for medium-to-long documents since RST parsing is particularly challenging at the macro-level (Jia et al., 2018; Hou et al., 2020; Zhang et al., 2020). Documents with an average of 1K+ tokens are selected to provide more training data for learning higher-level discourse structures. The corpus covers over 60K tokens and 9K EDUs.
3.2 Corpus Annotation

This section exemplifies the annotation procedure of the GCDT corpus, including document selection (Section 3.2.1), metadata and structural annotation (Section 3.2.2), tokenization (Section 3.2.3), EDU segmentation (Section 3.2.4), and relation annotation (Section 3.2.5). For both EDU segmentation and relation annotation, a second annotation version is implemented by another doctoral student at Georgetown University who is also a native Mandarin Chinese speaker and an expert in RST.

The annotation process took six months, including one month for document selection, metadata, XML, and tokenization annotations, two for EDU annotation and adjudication, and three for relation annotation. The dataset is released on GitHub\(^1\), and will be periodically updated if annotation inconsistency occurs. The annotation manual is publicly available on arXiv (Peng et al., 2022a).

3.2.1 Document Selection

After deciding on the annotation framework, the first step in corpus creation is to select the target documents. Ten documents from each of the five genres in Table 3.1 are selected to create a multi-genre and balanced corpus. Based on experience from previous discourse datasets, only freely available texts (CC-BY, CC-BY-NC, CC-BY-SA, and CC-BY-NC-SA) are included to increase the accessibility of this corpus. Four of the five GCDT genres inherit the sources from GUM: bio, interview, news, and who. Since many Wikipedia articles include versions from different languages, Chinese versions of existing GUM documents are intentionally included in GCDT for future cross-lingual comparisons. For academic, I select documents from Hans 汉斯 (https://www.hanspub.org/), a website with a large collection of open-access

\(^{1}\text{https://github.com/logan-siyao-peng/GCDT.}
journal articles. Continuous section spans of ~2K Chinese characters from the beginning of a document are extracted from these web pages for the following annotations.

3.2.2 Metadata and Structural Annotation

To provide more textual and extra-linguistic backgrounds to the audience, XML tags are used to annotate document metadata and structural information in GCDT. The <text> tag includes all tokens in a document and provides document-level metadata information as attributes. Since many websites are version-controlled and updated over time, metadata information is included to facilitate future association with specific source data versions.

- **id**: a unique ID in the format of gcdt_{genre}_{shortTitle}, for example, gcdt_academic_aging;
- **genre**: the genre of the document, e.g., academic;
- **shortTitle**: a one-word English short title for each document, e.g., aging;
- **title**: the original document title in Chinese characters, e.g. 老龄化对言语感知的影响;
- **author**: the author(s) of the document; in many cases, it would be persons’ names, e.g., Meijuan Ning, while in others, it would be the name of the website, e.g., Wikinews;
- **sourceURL**: the source URL of the document, e.g., https://www.hanspub.org/journal/PaperInformation.aspx?paperID=28037; if the website is version controlled, version suffixes are included; for example, oldid denotes the version information in the Wikipedia URL: https://zh.wikipedia.org/w/index.php?title=%E8%B5%B5%E5%85%83%E4%BB%BB&oldid=68901642.
- **dateCollected**: the date when the document is collected, in the format of yyyy-mm-dd, e.g., 2022-01-02;
• `dateCreated`: the date when the document is first created on the website; if this information is not given, the filler value is `XXXX-XX-XX`;

• `dateModified`: the date when the document is most recently modified before the collection date;

• `speakerCount`: the number of speakers in the document; this is especially useful for `interview`, while in other genres, the number is always 0;

• `speakerList`: the list of speakers in the document, e.g., `#WN, #Falkvinge` for `gcdt_interview_falkvinge` and `none` for `gcdt_academic_aging`.

Moreover, gold section, paragraph, and sentence splits are annotated for macro versus micro-level RST analyses. Section and subsection breaks are marked by XML tags: `<section>`, `<subsection>`, and `<subsubsection>`; paragraph breaks are marked by two line feeds: `|n|n` and sentence breaks are marked by one line feed: `|n`.

### 3.2.3 Tokenization

Gold tokenization is fundamental to the boundary decisions in EDU segmentation and detecting discourse relation signals. Thus, manually annotating gold tokens in GCDT improves the quality of its discourse annotations and provides tokenized sources for future annotation projects. Tokenization in GCDT follows the guideline (Xia, 2000b) from the Penn Chinese Treebank (CTB, Xue 2005), one detailed and linguistically-motivated syntactic guideline with a large amount of consistently annotated data for reference.

### 3.2.4 EDU Segmentation

Elementary Discourse Unit (EDU) segmentation establishes the elementary discourse segments for building RST trees. This is the first step in RST annotations and end-
to-end RST parsing. GCDT deviates from previous Chinese corpora that predomin-
antly use potentially ambiguous punctuations to segment EDUs regardless of the
surrounding syntactic structures (Li et al., 2014a). Instead, GCDT’s EDU segmenta-
tion for Chinese mirrors the syntactic criteria established in the English RST-DT and
GUM corpora (Carlson and Marcu, 2001; Carlson et al., 2001; Zeldes, 2017). Specif-
ically, EDUs are equated with the propositional structure of clauses as proposed in
Mann and Thompson (1988). In practice, the Penn Chinese Treebank (Xue et al.,
2005b) is treated as syntactic guidelines (particularly the parts-of-speech guidelines
Xia 2000a) to identify clausal elements. This section exemplifies a few important EDU
segmentation criteria in GCDT.

Firstly, GCDT segments adverbial clauses from their main clauses. Different types
of adverbial clauses are marked in bold in the following examples, including but not
limited to purpose (1), means (2), cause (3), result (4), circumstance (5), condition
(6), and concession (7). A double-pipe symbol ‘||’ is used in the following examples
to indicate EDU boundaries.

### Purpose:

1. 他 于 1724 年 前往 圣彼得堡 || 出任 数学 教授 。
   3SG.M in 1724 go-to St-Petersburg || take-office math professor .
   ‘He went to St. Petersburg in 1724 to be a professor of mathematics.’

   source: gcdt_bio_bernoulli

### Means:

2. 往往 由 公安 机关 || 以 寻衅滋事 为 由 ||
   often by security department || taking trouble-provoking as reason ||
   处以 行政 拘留 。
   sentence administrative detention .
‘(They) were often sentenced to administrative detention by the security departments because of making trouble.’

source: gcdt_academic_supervision

Cause:

(3) just COP for help advertise, because 1PL NEG have winning DE illusion.

‘Just to help spread the word, because we don’t have the illusion of winning.’

source: gcdt_interview_graaf

Result:

(4) its fake news is often hidden or adapted in real and trending social events, and thus difficult to identify.

source: gcdt_academic_supervision

Circumstance:

(5) when position at one CL broadband or wireless network DE environment, NEG people will think-of any information-flow DE emergence.

‘When positioned in a broadband or wireless network environment, no one will think of any emergence of information flow.’

source: gcdt_interview_wimax

Condition:

(6) once disease occur, mortality nearly 100%.

33
‘Once the disease occurs, the mortality is nearly 100%.’

source: gcdt_academic_rabies

Concession:

(7) 虽然仍然是三镜头设计，但每颗镜头都有明显改进。

‘Although it is still a three-lens design, each lens is a marked improvement.’

source: gcdt_news_apple

Secondly, the corpus also segments coordinated (8) and attribution (9) clauses following the English RST-DT and GUM guidelines. Attribution predicates (e.g., speech and cognitive predicates) indicate the source of information. Even though the reported message is in a syntactically embedded clause, it is more prominent than the source of information and thus segmented and treated as the nucleus.

Coordination:

(8) 他、欧拉的同现代人——也是密友。

‘He was Euler’s contemporary and a close friend.’

source: gcdt_bio_bernoulli

Attribution:

(9) 他 自己说：‘在应用文方面，英文、德文、法文没有问题。’

‘He said: “as for formal writing, (he) has no problem with English, German, and French.”’

source: gcdt_bio_chao
Thirdly, GCDT follows both English datasets to segment EDUs separated by certain punctuation marks, including parenthesis (10), dash (11), and colon (12), as well as EDUs separated by strong discourse markers (13). The annotation guideline enumerates a list of strong discourse markers for Chinese similar to the ones from RST-DT and GUM (see Peng et al. 2022a for details). These include 虽然 (although), 但是 (but), 因为 (because), 所以 (as a result), 按照 (according to), 如果 (if), 无论 (regardless of), 不仅 (not only), 而且 (but also), etc.

Parenthesis:

(10) 约翰还曾试图盗窃丹尼尔的著作 《Hydrodynamica》 || (流体 mechanics) || 也 BA 3SG.IN anew name to-be 《Hydraulica》.

‘John also tried to steal Daniel’s book Hydrodynamica (Fluid Mechanics) and rename it Hydraulica.’

source: gcdt_bio_bernoulli

Dash:

(11) 德沃夏克在纽约遇到了他后来的学生哈里·布雷 Dvorak at New-York meet PERF 3SG.M future DE student Harry-Bray || —— 最早的美国黑人作曲家之一。 || —— most early DE American black composers one-of.

‘Dvorak met his future student Harry Bray in New York, one of the first African-American composers.’

source: gcdt_bio_dvorak

Colon:

(12) 英语中主要分为三个“态”：|| 主动态， English in mainly divide-into three CL “voices”：|| active-voice， 中动态和被动态。 middle-voice and passive-voice.
‘English is mainly divided into three “voices”: active, middle, and passive.’

*source: gcdt_academic_iconicity*

**Strong Discourse Marker:**

(13) 虽然 安全， || 但 不 方便
although safe , || but NEG convenient

‘Although it is safe, it is inconvenient.’

*source: gcdt_interview_wimax*

In addition to the segmentation decisions above that are parallel to English, two criteria below are distinctive to Chinese syntax. The first one concerns predicative adjectives and nominals in Mandarin Chinese. Unlike English, predicative sentences in Chinese do not require an overt copula (Tu and Zhang, 2013). Thus, a predicative clause can be formed in Chinese by constructions such as *Subject + PredAdj* (14) or *Subject + PredNoun* (15).

**Predicative Adjective:**

(14) 拜伦 先天性 的 跛足， || 而 他 的 母亲 性情 乖戾
Byron congenital DE lame , || but 3SG.M POSS mother temper grumpy

, 喜怒 无常
happy-sad unstable

‘Byron is lame by birth, and his mother was surly and moody.’

*source: gcdt_bio_byron*

**Predicative Noun:**

(15) 通口一叶 原名 通口奈津 或 通口夏子 ， || 是
Higuchi-Ichiyo original-name Higuchi-Najin or Higuchi-Natsuko , || COP

日本 明治 初期 主要 的 女性 小说家 。
Japan Meiji early-period leading DE female novelist .

‘Higuchi Ichiyo’s original name is Higuchi Najin or Higuchi Natsuko, and she was Japan’s leading female novelist in the early Meiji period.’

*source: gcdt_bio_higuchi*
GCDT also segments relative clauses following the practice in some RST corpora (Carlson et al., 2001; Zeldes, 2017; Das and Stede, 2018; Cardoso et al., 2011b; Redeker et al., 2012; Toldova et al., 2017). Chinese relative clauses present a unique feature in the existing RST treebanks. GCDT is the first RST corpus in any language where prenominal relative clauses are annotated for discourse relations. Cross-referencing Dryer (2013a,b) with languages of existing RST corpora suggests that only Basque exhibits the Relative-Noun order similar to Chinese. Yet, neither restrictive nor non-restrictive relative clauses are segmented in the Basque RST dataset (Iruskieta et al., 2015).

Example (16) below presents a sentence with two embedded relative clauses: the workers that destroyed machines and the bill that states machine-destroying workers must be given death penalty. Both relative clauses are relativized by a 的 relativizer and occur before the head noun.

Relative Clause:

(16) `In February, the House of Lords passed the bill announcing that workers who destroy machines must be sentenced to death.’

source: gcdt_bio_byron

Since relative clauses intervene between Verb-Object in Chinese, the pseudo-relation same-unit is used to express discontinuous EDUs. Segmenting and annotating discourse relations for relative clauses is one of the reasons that GCDT has relatively short EDUs, on average 6.5 tokens/EDU. Other reasons for shorter EDUs in Chinese include compound tokens, the lack of articles, less use of auxiliaries, etc. (Webster and Kit, 1992; Jiang and Liu, 2015).
Examples (1)-(16) above exemplify the major EDU segmentation criteria in GCDT. Despite these propositional phrases, many other phrases are not segmented in GCDT, for example, complement clauses, coordinated nominals or verbs, existential clauses, and prepositional phrases. For detailed segmentation guidelines, please refer to the Chinese Discourse Annotation Reference Manual (Peng et al., 2022a).

3.2.5 Relation Annotation

GCDT builds constituency-style discourse trees from gold-segmented EDUs by annotating on the rstWeb interface (Zeldes, 2016). The corpus uses the enhanced two-level relation labels from GUM V8.0.0 with 15 coarse and 32 fine-grained relations (Zeldes, 2017). The revised GUM relation inventory is named by hyphen-joining the first-level classes with second-level relation names. For example, causal-result belongs to the first-level causal class, and it is a second-level result relation.

Each relation in the inventory concerns the reader (R) and/or the writer (W). The definition also specifies the direction between the nucleus (N) and the satellite (S) discourse units. Among the 32 relations listed below, 25 are nucleus-satellites, and 7 are multinuclear relations. The relations are enumerated alphabetically below such that relations belonging to the same class and having the same nuclearity are presented next to each other (see Peng et al. 2022a for more details).

**Nucleus-Satellite Relation:**

1. adversative-antithesis: R finds N more credible than S;
2. adversative-concession: W admits S but still claims N;
3. attribution-negative: S does not assert the information in N;
4. attribution-positive: S provides a positive source of information to N;
5. causal-cause: S causes N;
6. causal-result: S is a result of N;
context-background: R needs to know S to understand N;
context-circumstance: S gives circumstances, e.g., time, place, of N;
contingency-condition: S is a condition for N to happen;
elaboration-additional: S provides more information about N;
elaboration-attribute: S provides more information about some non-propositional phrases in N;
evaluation-comment: S gives an opinion about N;
explanation-evidence: S gives evidence that N is true;
explanation-justify: S justifies why W can say N;
explanation-motivation: S motivates R to act based on N;
mode-manner: S describes how (in which way) N happened;
mode-means: S indicates how (the means by which) N happened;
organization-heading: S is graphically arranged to prepare for N;
organization-phatic: S holds the floor for N, with no semantic value;
organization-preparation: S prepares R for N;
purpose-attribute: only a non-propositional phrase in N occurs for S to happen;
purpose-goal: N occurs for S to happen;
restatement-partial: S reiterates part of N;
topic-question: S requests the information in N;
topic-solutionhood: N is the answer to a problem in S.

Multinuclear Relation:
adversative-contrast: W presents similar units with contrast;
joint-disjunction: W presents a set of alternatives;
joint-list: W presents coordinated and similar units;
joint-sequence: W presents multiple units of chronological sequence;
joint-other: W presents unlike units with no other relation;
restatement-repetition: W presents equivalent or redundant units;
same-unit: this is a technical device for interrupted EDUs.

Figure 3.1 presents a sample RST sub-tree with relation annotations from an academic article in GCDT. The snippet includes six EDUs from four sentences. The first sentence consists of three EDUs due to the pre-nominal relative clause in EDU_44 (see Section 3.2.4 for EDU segmentation of pre-nominal relative clauses in Chinese). EDU_44 ‘that is named Smiling Collector Brother Bo’ provides more detail about ‘an online user’ in EDU_45, and thus results in an elaboration-attribute relation and a same-unit technical device. The ‘going viral overnight’ in EDU_46 chronologically follows ‘the posting of a video about Dingzhen’ in DU_43-45, and they jointly form a multi-nuclear sequence relation. These first two sentences provide background for the most central discourse unit of the snippet, ‘Dingzhen launching social platforms’ in EDU_47. EDU_47 is further elaborated by EDU_48, adding that ‘Douyin and Weibo platforms are the most popular and contain the most representative content’

Figure 3.1: Sample relation annotations in gcdt_academic_dingzhen.
Among the 32 relation labels in GCDT, there are two last-resort labels for nucleus-satellite and multinuclear relations (i.e., *elaboration-additional* and *joint-other*), especially at the higher structural levels. Section 3.4.2 in this chapter will present the distribution of 32 relations in comparison with GUM. Section 4.5.4 further delves into the different percentages of discourse relations between macro- versus micro-levels and across genres.

Another crucial feature of the RST structure is the order of satellite attachment. The constituency-styled discourse framework first associates the most relevant satellite to the nucleus (i.e., at the lowest level) and then successively connects the less relevant satellites. Figure 3.2 presents a relation hierarchy of *attribution-positive* scoping over *causal-cause.*\(^2\) In this example, the cause and the result are attributed to *理论* (the theory). Alternatively, the interpretation would change if *causal-cause* scopes over *attribution-positive.* Though the order of satellite attachments signifies the importance of different discourse units, it also triggers disagreements among annotators. Figure 3.3 presents two annotation versions of a subtree in *gcdt_academic_dingzhen* where the satellites modify the same nucleus but are attached to it in different orders. It is ambiguous whether *mainstream media making full use of their resources* should be scoped under *earlier interviews paving the way for subsequent agenda setting* as in Figure 3.3a, or vice versa as in Figure 3.3b. In these scenarios, annotators’ choices to alter the order of satellite attachments negatively impact the inter-annotator agreement in RST’s hierarchical discourse structure.

\(^2\)In this and the following RST examples, Google Translate is employed to automatically translate Chinese EDUs to English and append them after the source Chinese EDUs using a double-slash ‘//’ separator. Though these translations are imperfect, they help present RST trees to non-Chinese audiences. These translation-augmented EDUs also boost model performance in multilingual RST parsing experiments shown in Chapter 5.
The unbalanced distribution of relations and the disagreements in the structural hierarchy discussed above are universal across RST corpora in different languages (Zeldes et al., 2021b). However, the frequent combinations of *elaboration-attribute* and *same-unit* labels are dominant in Mandarin Chinese. Because relative clauses occur prenominally, an object noun phrase would be separated from its predicate if a relative clause modifies it. The relative clause is most commonly annotated as *elaboration-attribute* because it only elaborates the object noun phrase part of the clause. In the meantime, a *same-unit* relation is required to bridge the two discontinuous units, i.e., the predicate and the dislocated object noun phrase. Figure 3.4 presents two typical RST structures associated with relative clauses in Mandarin Chinese. In DU_161-163, the object 小鲜肉 (young hunk) is modified by the relative clause 被美颜和化妆品包装的 (decorated by the beauty and cosmetic products) and thus separated from its predicate 充斥着 (be full of). In DU_164-165, 审美价值 (aesthetic value) is
Figure 3.3: Different attachment orders of *mode-means* and *evaluation-comment* satellites in two annotation versions of *gcdt_academic_dingzhen*.
the subject that occurs before the predicate 日趋统一 (increasingly unified), and its relative clause in EDU_164 does not intervene in the main clause. Thus, a same-unit relation is required for DU_161-163 but not necessary for DU_164-165.

Figure 3.4: An RST subtree with two relative clauses annotated as elaboration-attribute and same-unit in gcdt_academic_dingzhen.

Relative clauses present substantial discourse information at the clausal and phrasal levels (Cornish, 2018; Roland et al., 2012). Since RST is a function-based formalism, one should admit propositional relative clauses as Elementary Discourse Units because they could be paraphrased as non-relative propositional structures (Inui and Nogami, 2001). GCDT is not only the first corpus that annotates pre-nominal relative clauses across RST datasets but also the first that annotates relative clauses in any discourse framework for Mandarin Chinese. The rich annotations of pre-nominal Chinese relative clauses in GCDT can take discourse annotation and modeling to the next level.

In summary, Section 3.2 highlights the document selection, metadata, XML, tokenization annotation, EDU segmentation, and relation annotation decisions in the
Georgetown Chinese Discourse Treebank (GCDT). Peng et al. (2022a) provide a detailed Chinese RST annotation guideline with 80+ pages and 150+ examples, particularly for EDU segmentation and relation annotation. Similar to the examples in this dissertation, the guideline provides EDU-level automatic Chinese-to-English translations for relation labeling and token-level glossing for EDU segmentation.

3.3 Data Management and Agreement

3.3.1 Data Types

All text sources and annotations of GCDT are freely available on GitHub.\footnote{https://github.com/logan-siyao-peng/GCDT} The following data sources and formats are available under their respective sub-directories within the data/ directory.

- \texttt{xml/}: raw texts are extracted from the source websites, followed by metadata, sentence, paragraph, and section annotations as described in Section 3.2.2;
- \texttt{tokenized/}: manually tokenized texts with one sentence per line (see Section 3.2.3);
- \texttt{conllu/}: automatically UD-parsed (Universal Dependencies) gold-tokenized texts using stanza (Qi et al., 2020) to facilitate future syntax-aware tasks;
- \texttt{rs3/}: original constituency-styled RST annotations of 50 documents and five additional double-annotated test documents (see Sections 3.2.4 and 3.2.5);
- \texttt{rs3\_extracted\_edus/}: EDUs (one per line) automatically extracted from annotated rs3 files to facilitate parsing experiments;
- \texttt{autotrans\_rs3/}: constituency-styled rs3 files with automatic EDU-level Chinese-to-English Google translations appended to the end of each EDU in the format...
ZH // EN; annotation examples presented in this dissertation are graphical displays of this translation-appended rs3 format;

- autotrans_extracted_edus/: EDUs (one per line) automatically extracted from translation-appended rs3 files;
- rsd/: automatically converted rs3 files to the rsd format using https://github.com/amir-zeldes/gum/blob/master/_build/utils/rst2dep.py;
- dis/: automatically converted rs3 files to the dis format using https://github.com/amir-zeldes/gum/blob/master/_build/utils/rst2dis.py.

3.3.2 Data Split

GCDT provides an 8-1-1 train-dev-test split per genre in Table 3.2 to facilitate future RST parsing experiments. Both human inter-annotator agreements and parsing results are assessed on the five test documents, one from each of the five genres.

Table 3.2: GCDT’s 8-1-1 data splits, document names, and the number of tokens and EDUs in the train, dev, and test sets.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Train Documents</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>academic</td>
<td>aging, governance, iconicity, rabies, socialized, supervision, taoist, yun</td>
<td>peoples</td>
<td>dingzhen</td>
</tr>
<tr>
<td></td>
<td>#tokens: 10,870; #EDUs: 1,616</td>
<td>1,519; 207</td>
<td>1,779; 210</td>
</tr>
<tr>
<td>bio</td>
<td>bernoulli, chao, emperor, galois, hadid, higuchi, jerome, marbles</td>
<td>byron</td>
<td>dvorak</td>
</tr>
<tr>
<td></td>
<td>#tokens: 9,722; #EDUs: 1,486</td>
<td>1,901; 279</td>
<td>1,862; 256</td>
</tr>
<tr>
<td>interview</td>
<td>cycle, exiderdome, falkvinge, game, graaf, ideal, keyman, stardust</td>
<td>ward</td>
<td>wimax</td>
</tr>
<tr>
<td></td>
<td>#tokens: 9,142; #EDUs: 1,485</td>
<td>1,280; 174</td>
<td>1,042; 153</td>
</tr>
<tr>
<td>news</td>
<td>apple, bubble, estate, five, hubei, kangle, tiktok, unemployment</td>
<td>famine</td>
<td>simplified</td>
</tr>
<tr>
<td></td>
<td>#tokens: 8,769; #EDUs: 1,299</td>
<td>1,175; 168</td>
<td>1,305; 185</td>
</tr>
<tr>
<td>whow</td>
<td>basil, flirt, glowstick, mice, pool, procrastinating, quinoa, skittles</td>
<td>hiking</td>
<td>thanksgiving</td>
</tr>
<tr>
<td></td>
<td>#tokens: 9,136; #EDUs: 1,594</td>
<td>1,744; 317</td>
<td>1,659; 288</td>
</tr>
</tbody>
</table>

46
3.3.3 Inter-Annotator Agreement (IAA)

This section evaluates the inter-annotator agreement on the five test documents to set up the human ceiling scores for future comparisons with parsers’ performances. I drafted the annotation guidelines and annotated the entire corpus. Another Chinese native-speaker linguist read the guidelines and conducted independent EDU segmentation. The two annotators measured segmentation agreement, adjudicated segmentation, and then separately annotated relation trees on gold EDUs to measure relation agreement. Table 3.3 presents the overall and per-genre inter-annotator agreements (IAA) on EDU segmentation and relation annotation.

Table 3.3: GCDT’s genre-wise inter-annotator agreements on EDU segmentation and relation annotation.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Segmentation</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Cohen’s κ</td>
</tr>
<tr>
<td>academic</td>
<td>0.97</td>
<td>0.89</td>
</tr>
<tr>
<td>bio</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>interview</td>
<td>0.97</td>
<td>0.88</td>
</tr>
<tr>
<td>news</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>whow</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td>Overall</td>
<td>0.97</td>
<td>0.89</td>
</tr>
</tbody>
</table>

On segmentation, GCDT achieves a micro-averaged token-wise accuracy of 0.97 and Cohen’s κ of 0.89. The variation in EDU segmentation is trivial among genres. In RST relation annotations, the overall agreement scores on micro-averaged original Parseval F1 of Span, Nuclearity, and Relation are 84.2, 66.1, and 57.7. This IAA of GCDT is similar to that of the English RST-DT benchmark when evaluated using the original Parseval metrics, i.e., 78.7, 66.8, and 57.1 (Morey et al., 2017). However, genre-wise relation agreements differ broadly. While biographies, interviews, and news have similar IAAs and are close to the average, the F1_Relation of how-to guides (whow) is almost 20 points higher than that of academic articles. Namely, all three relation metrics for whow are remarkably higher than the other four genres, and
Nuclearity and Relation scores for academic are a few points lower than bio, interview, and news. Whow and academic exhibit diverse writing styles, particularly regarding how the supporting evidence and claims construe the central ideas. Still, both genres exhibit strong signals for discourse labeling (cf. Zeldes 2018 for similar findings in English).

In short, results show that GCDT’s annotation agreement is highly satisfactory compared to the English RST-DT benchmark, even though the documents are much longer and exhibit more genre diversity in GCDT. The data is freely available on GitHub in various formats and provides gold train/dev/test splits for future RST parsing evaluations, as well as double annotations on EDU segmentation and relation structuring for disagreement studies between annotators and parsers.

3.4 Statistics of Discourse Markers and Relations

Discourse marker recognition and relation classification are essential NLP tasks for interpreting discourse representations (Song and Liu, 2020; Li et al., 2022). To quantify what kinds of discourse information the newly annotated GCDT corpus can bring to the field, this section presents frequency statistics on discourse markers (§3.4.1) and relation labels (§3.4.2).

3.4.1 Discourse Markers

Discourse markers (e.g., connectives) are important to document structuring (Li et al., 2014c; Prasad et al., 2008, 2018; Das and Taboada, 2018). Though Rhetorical Structure Theory (RST) does not explicitly annotate discourse markers in its primary

GCDT does not include discourse signal annotation in its current release. This section explores the distribution of discourse connectives by extracting them from the PDTB-styled Chinese Discourse Treebank corpus (CDTB, Zhou and Xue 2015) and comparing their frequencies in CDTB and GCDT and across GCDT genres. CDTB contains 73K tokens from 164 news documents, slightly larger than GCDT’s 62K tokens. Table 3.4 selects the top 30 multi-character connectives from CDTB and presents their frequencies per 100K tokens in CDTB, GCDT, as well as in each of GCDT’s five genres.

Firstly, the frequencies of most connectives differ widely between CDTB’s Xinhua news and GCDT’s multiple genres. The connectives that occur at least twice more frequently in CDTB than GCDT are marked in blue; and vice versa, red for GCDT’s more frequent connectives. Many discourse markers are more frequent in the CDTB’s traditional and authoritative news. These include connectives to signal enumerations, such as 其中 (among them), 此外 (besides that), 二是 (secondly), as well as ones that mark temporal relations, such as 的同时 (the same time of), and 与此同时 (at the same time). On the other hand, since GCDT has a higher genre diversity and many of the genres incorporate deeper logical reasoning than fact-reporting news, more argumentative connectives are observed, such as 如果 (if), 通过 (by means of), 但是 (but), 因此 (because of that), 之后 (after so), 虽然 (even though), 因为 (because of), 甚至 (even), 以及 (as well as), 不过 (nevertheless), 所以 (therefore), 那么 (in that case), etc.

Secondly, Table 3.4 also compares the frequencies of connectives in five individual genres against GCDT’s overall frequency. Discourse markers that signal an accompa-
Table 3.4: Top 30 explicit multi-character connectives in CDTB and their frequencies in GCDT across genres. All frequency counts are per 100K tokens within the specific corpus or genre.

<table>
<thead>
<tr>
<th>Connective</th>
<th>Translation</th>
<th>CDTB</th>
<th>GCDT</th>
<th>GCDT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>all</td>
<td>all</td>
<td>aca</td>
</tr>
<tr>
<td>其中</td>
<td>among them</td>
<td>111</td>
<td>37</td>
<td>71</td>
</tr>
<tr>
<td>随着</td>
<td>along with</td>
<td>49</td>
<td>26</td>
<td>79</td>
</tr>
<tr>
<td>此外</td>
<td>besides that</td>
<td>47</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td>同时</td>
<td>at the same time</td>
<td>40</td>
<td>71</td>
<td>86</td>
</tr>
<tr>
<td>由于</td>
<td>because of</td>
<td>37</td>
<td>39</td>
<td>36</td>
</tr>
<tr>
<td>以来</td>
<td>since then</td>
<td>32</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>而且</td>
<td>but also</td>
<td>23</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>为了</td>
<td>for the purpose of</td>
<td>23</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td>如果</td>
<td>if</td>
<td>23</td>
<td>182</td>
<td>14</td>
</tr>
<tr>
<td>尽管</td>
<td>even though</td>
<td>22</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>的同时</td>
<td>the same time of</td>
<td>22</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>通过</td>
<td>by means of</td>
<td>21</td>
<td>66</td>
<td>164</td>
</tr>
<tr>
<td>不仅</td>
<td>not only</td>
<td>19</td>
<td>31</td>
<td>57</td>
</tr>
<tr>
<td>但是</td>
<td>but</td>
<td>18</td>
<td>95</td>
<td>86</td>
</tr>
<tr>
<td>因此</td>
<td>because of that</td>
<td>16</td>
<td>89</td>
<td>143</td>
</tr>
<tr>
<td>经过</td>
<td>by way of</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>与此同时</td>
<td>at the same time</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>然而</td>
<td>however</td>
<td>12</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>特别是</td>
<td>especially</td>
<td>12</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>之后</td>
<td>after so</td>
<td>11</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>虽然</td>
<td>even though</td>
<td>11</td>
<td>68</td>
<td>21</td>
</tr>
<tr>
<td>因为</td>
<td>because of</td>
<td>11</td>
<td>118</td>
<td>36</td>
</tr>
<tr>
<td>甚至</td>
<td>even</td>
<td>8</td>
<td>39</td>
<td>21</td>
</tr>
<tr>
<td>从而</td>
<td>accordingly</td>
<td>8</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>以及</td>
<td>as well as</td>
<td>8</td>
<td>82</td>
<td>71</td>
</tr>
<tr>
<td>不过</td>
<td>nevertheless</td>
<td>8</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>二是</td>
<td>secondly</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>所以</td>
<td>therefore</td>
<td>7</td>
<td>63</td>
<td>29</td>
</tr>
<tr>
<td>那么</td>
<td>in that case</td>
<td>7</td>
<td>39</td>
<td>7</td>
</tr>
<tr>
<td>首先</td>
<td>at first</td>
<td>7</td>
<td>8</td>
<td>14</td>
</tr>
</tbody>
</table>
nying event (in brown) are the most frequent in the academic genre, including 随着 (along with), 通过 (through), and 从而 (accordingly). Purpose and causal markers (in green), for example, 为了 (for the purpose of), 因此 (because of that), 因为 (because of), have a higher appearance in interview. Temporal markers (in orange) such as 同时 (at the same time), 以来 (since then), and 首先 (at first) are frequent in GCDT news, the same genre as the CDTB dataset. Conditional markers (in cyan) are dominant in how-to guides, such as 如果 (if) and 那么 (in that case). In summary, the top-ranking CDTB connectives exhibit substantially different distributions in GCDT and across GCDT’s multiple genres.

The current analysis has its limitations. Since GCDT does not yet include discourse marker annotation, this dissertation only string-matches connectives to existing CDTB connective annotations. Since string match is error-prone when matching mono-character connectives, this analysis only includes multi-character connectives in Table 3.4. Moreover, connectives that occur exclusively outside of the news domain remain unknown. Beyond this dissertation work, I envision including signal annotations (e.g., discourse connectives) to GCDT to discover new discourse markers in non-news genres, analyze their distributions across genres and corpora, and facilitate discourse parsing.

3.4.2 Discourse Relations

In addition to discourse markers, discourse relations are fundamental to discourse representations. However, many discourse datasets within the same or across frameworks employ different sets of relations labels in their annotation scheme. For example, relation categories in GUM are moderately divergent from RST-DT or PDTB. Since GCDT was initiated to be directly comparable with the English GUM benchmark, GCDT uses the same relation inventory and covers a subset of GUM genres. Table 3.5
lays out descriptive statistics regarding the distribution of relations in GCDT and percentages for comparison from the GUM V8.0.0 corpus. Distributional statistics from two GUM sub-samples are included: GUM-12 (193 English documents from 12 genres) and GUM-5 (99 GUM documents from the same five genres in GCDT).

Table 3.5: Distribution of 32 relations in GCDT, as well as in the same five genres of GUM (GUM-5) and all twelve genres of GUM (GUM-12) as of V8.0.0.

<table>
<thead>
<tr>
<th>Relation Name</th>
<th>GCDT</th>
<th>GUM-5</th>
<th>GUM-12</th>
<th>Relation Name</th>
<th>GCDT</th>
<th>GUM-5</th>
<th>GUM-12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nucleus-Satellite Relations</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Nucleus-Satellite Relations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elaboration-attribute</td>
<td>7.83%</td>
<td>8.66%</td>
<td>7.81%</td>
<td>evaluation-comment</td>
<td>1.07%</td>
<td>1.64%</td>
<td>3.04%</td>
</tr>
<tr>
<td>attribution-positive</td>
<td>4.35%</td>
<td>4.27%</td>
<td>4.56%</td>
<td>mode-means</td>
<td>1.04%</td>
<td>0.89%</td>
<td>0.56%</td>
</tr>
<tr>
<td>elaboration-additional</td>
<td>4.35%</td>
<td>10.65%</td>
<td>9.01%</td>
<td>causal-result</td>
<td>0.97%</td>
<td>1.52%</td>
<td>1.55%</td>
</tr>
<tr>
<td>explanation-evidence</td>
<td>4.12%</td>
<td>3.88%</td>
<td>2.70%</td>
<td>explanation-justify</td>
<td>0.95%</td>
<td>1.17%</td>
<td>1.60%</td>
</tr>
<tr>
<td>context-background</td>
<td>2.76%</td>
<td>3.72%</td>
<td>3.80%</td>
<td>adversative-antithesis</td>
<td>0.62%</td>
<td>1.31%</td>
<td>1.48%</td>
</tr>
<tr>
<td>context-circumstance</td>
<td>2.76%</td>
<td>3.22%</td>
<td>3.26%</td>
<td>mode-manner</td>
<td>0.59%</td>
<td>0.63%</td>
<td>0.89%</td>
</tr>
<tr>
<td>organization-preparation</td>
<td>2.27%</td>
<td>2.37%</td>
<td>2.28%</td>
<td>topic-question</td>
<td>0.45%</td>
<td>0.89%</td>
<td>1.11%</td>
</tr>
<tr>
<td>causal-cause</td>
<td>1.88%</td>
<td>1.51%</td>
<td>1.95%</td>
<td>explanation-motivation</td>
<td>0.28%</td>
<td>0.79%</td>
<td>0.71%</td>
</tr>
<tr>
<td>contingency-condition</td>
<td>1.80%</td>
<td>1.81%</td>
<td>1.68%</td>
<td>organization-phatic</td>
<td>0.26%</td>
<td>0.07%</td>
<td>1.38%</td>
</tr>
<tr>
<td>organization-heading</td>
<td>1.78%</td>
<td>1.91%</td>
<td>1.50%</td>
<td>attribution-negative</td>
<td>0.23%</td>
<td>0.13%</td>
<td>0.31%</td>
</tr>
<tr>
<td>adversative-concession</td>
<td>1.64%</td>
<td>2.15%</td>
<td>2.59%</td>
<td>purpose-attribute</td>
<td>0.21%</td>
<td>1.08%</td>
<td>0.88%</td>
</tr>
<tr>
<td>purpose-goal</td>
<td>1.52%</td>
<td>2.41%</td>
<td>1.92%</td>
<td>topic-solutionhood</td>
<td>0.01%</td>
<td>0.17%</td>
<td>0.20%</td>
</tr>
<tr>
<td>restatement-partial</td>
<td>1.29%</td>
<td>1.33%</td>
<td>1.14%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multinuclear Relations</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>Multinuclear Relations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>joint-list</td>
<td>22.16%</td>
<td>11.87%</td>
<td>12.90%</td>
<td>adversative-contrast</td>
<td>3.25%</td>
<td>3.00%</td>
<td>3.33%</td>
</tr>
<tr>
<td>same-unit</td>
<td>18.74%</td>
<td>9.53%</td>
<td>8.40%</td>
<td>joint-disjunction</td>
<td>0.73%</td>
<td>1.03%</td>
<td>1.14%</td>
</tr>
<tr>
<td>joint-sequence</td>
<td>5.10%</td>
<td>7.62%</td>
<td>6.82%</td>
<td>restatement-repetition</td>
<td>0.32%</td>
<td>0.86%</td>
<td>2.23%</td>
</tr>
<tr>
<td>joint-other</td>
<td>4.69%</td>
<td>7.91%</td>
<td>7.25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5 demonstrates that the less frequent relations have very close distributions between GCDT and GUM. The most significant variations between GCDT and GUM lie in elaboration-additional (in blue), joint-list, and same-unit (in red). The higher percentage of same-unit in GCDT is a direct result of the Chinese syntax, namely the order that relative clauses occur before their noun heads in Chinese (see Section 3.2.4). On the other hand, besides other language-dependent distinctions, annotators’ choices could contribute to the different distributions of elaboration-additional and joint-list. Table 3.6 presents the relation-wise agreements on the GCDT test set between the two annotators. This agreement is evaluated using the Dis-
course Dependency Structure and examines the relation labels of the discourse units that share the same dependency head between the two annotation versions. The primary annotation is treated as the gold version and the double annotation as the predicted version. The main annotator prefers relations such as joint-list and joint-sequence, resulting in low recall in these two relations. In contrast, the second annotator is biased towards joint-other and elaboration-additional, resulting in low precision. In contrast, other frequent relations such as same-unit, elaboration-attribute, and attribution-positive are highly agreeing between the two annotators. Annotation ambiguity and disagreement remain unresolved in the complex RST framework, especially for higher-level discourse units (Marcu et al., 1999; Stede, 2008).

Table 3.6: Precision, Recall, and F1 scores for high and low agreement relations between the main and double annotation versions of the five GCDT test documents.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Support</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent relations with high agreements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>same-unit</td>
<td>102</td>
<td>93.6</td>
<td>98.1</td>
<td>95.8</td>
</tr>
<tr>
<td>elaboration-attribute</td>
<td>85</td>
<td>94.4</td>
<td>75.9</td>
<td>84.2</td>
</tr>
<tr>
<td>attribution-positive</td>
<td>30</td>
<td>88.2</td>
<td>93.8</td>
<td>90.9</td>
</tr>
<tr>
<td>Frequent relations with low agreements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>joint-list</td>
<td>63</td>
<td>86.3</td>
<td>52.5</td>
<td>65.3</td>
</tr>
<tr>
<td>elaboration-additional</td>
<td>27</td>
<td>30.7</td>
<td>79.4</td>
<td>44.3</td>
</tr>
<tr>
<td>joint-other</td>
<td>9</td>
<td>15.0</td>
<td>81.8</td>
<td>25.4</td>
</tr>
<tr>
<td>joint-sequence</td>
<td>7</td>
<td>87.5</td>
<td>25.9</td>
<td>40.0</td>
</tr>
</tbody>
</table>

The diverse genre composition also contributes to differences in relation distributions in Table 3.5. When comparing GCDT with two samples of GUM: GUM-5 versus GUM-12, both GCDT, and GUM-5 have higher distributions of explanation-evidence and same-unit (in brown) due to the more argumentative nature and the more formal writing style of these five genres. Moreover, since GCDT and GUM-5 exclude informal or spoken genres such as reddit, conversation and vlog, they have lower distributions for evaluation-comment, organization-phatic, and restatement-repetition (in cyan).
This section demonstrates that most discourse distribution differences between corpora and languages lie in the more frequent relation labels. Language syntax, genre diversity, and inter-annotator disagreement contribute to the different relation distributions. The complexity of the RST tree structure necessitates future in-depth quantitative and qualitative analyses to ablate multifaceted factors for diversified discourse relation distributions across datasets.

3.5 **Interim Summary**

This chapter presents the Georgetown Chinese Discourse Treebank (GCDT), the largest multi-genre RST corpus for Mandarin Chinese. The corpus not only provides an extensive annotation guideline for Chinese EDU segmentation and RST relation annotation, it also focuses on providing more annotated data for longer documents and diverse genres, and both are bottlenecks in RST parsing experiments. To promote multilingual and multi-genre discourse structural analyses and parsing experiments in Chapters 4 and 5, this chapter also compares GCDT to existing Chinese and English discourse corpora regarding two critical components in a discourse representation: discourse markers and relations. Results show that genre heterogeneity, annotator bias, and language syntax trigger the distribution diversity of discourse elements in different corpora. Yet, there is always a demand for larger datasets, and hopefully, GCDT and its Chinese RST annotation guidelines can benefit future discourse corpora and studies.
Chapter 4

Analysis: Paragraphs in RST

Paragraphs are important macro-level structures in longer documents. Even though the original design of the Rhetorical Structure Theory has no association with holistic structures, such as sections and paragraphs (Mann et al., 1992), paragraphs play an important role in RST annotations, and multi-stage RST parsing (Sporleder and Lapata, 2004; Cao et al., 2018; Wang et al., 2019). For example, Cao et al. (2018) start with annotating RST relations within a paragraph and then continue with cross-paragraph relations. Wang et al. (2019); Nishida and Nakayama (2020); Kobayashi et al. (2020); Koto et al. (2021) decompose RST parsing into three stages, i.e., within a sentence, between sentences but within a paragraph, and between paragraphs, using either top-down or bottom-up algorithms. Even though paragraph structure plays an essential role in constructing and interpreting RST, there is a gap for a focused study associating paragraph boundaries with the RST theory across genres and corpora.

This chapter analyzes the associations between RST corpora and paragraph structures quantitatively and qualitatively. Section 4.1 aligns sentences and paragraphs to RST trees to create an extended discourse dependency format with sentence and paragraph segmentation information as described in Section 4.2. Section 4.3 presents such structural segment statistics in three RST corpora: RST-DT (Carlson et al., 2001) and GUM (Zeldes, 2017) for English, and GCDT (Peng et al., 2022b) for Chinese. These are the only three RST corpora with preserved document structure information within English and Chinese. The preprocessing steps and basic statistics lead to two
main analyses: paragraph containment and relation distribution. The containment analysis in Section 4.4 correlates RST tree structures with paragraph structures and looks into whether or not RST trees are properly contained in paragraphs. Section 4.5 inspects the distributional differences between intra-paragraph and inter-paragraph relations. Both analyses are conducted across the three corpora and the included genres. Section 4.6 concludes this chapter and draws further attention to document structures, such as paragraphs, in RST-style hierarchical discourse frameworks.

4.1 Aligning EDUs to Sentences and Paragraphs

A structure-informed RST analysis starts with aligning sentence and paragraph boundaries to the RST structure. Since the main annotations of the benchmark RST datasets do not contain sentence- and paragraph-level information, this section aligns each Elementary Discourse Unit (EDU) to unique sentence and paragraph indexes. Three benchmark RST corpora are the focus of this study: RST-DT (Carlson et al., 2001) and GUM V8.0.0 (Zeldes, 2017) for English, and GCDT (Peng et al., 2022b) for Chinese. The three corpora are the largest within the two languages and preserve relevant sentence and paragraph information in other annotation layers or corpora. Sections 4.1.1, 4.1.2 and 4.1.3 exemplify how sentences and paragraphs are aligned in RST-DT, GUM, and GCDT since the three datasets are released in moderately different formats.

4.1.1 English RST-DT

The English RST-DT corpus (Carlson et al., 2001) annotates a subset of 385 Wall Street Journal documents from the Penn Treebank (PTB) corpus (Marcus et al.,
1993). The LDC release (LDC2002T07) of RST-DT\(^1\) includes a \textit{wsj\_xxx.out} file\(^2\), for each document. In the \textit{wsj\_xxx.out} file, paragraphs are separated by two line feeds and sentences by one line feed. However, texts are not tokenized, and there are errors, such as multiple sentences with periods occurring in one line. To further rectify document-level structures in RST-DT, tokenization and sentence-splitting of RST-DT are aligned to the PTB dataset (LDC95T7, Marcus et al. 1993)\(^3\), where gold token, sentence, and paragraph splits are preserved in different layers of annotations.

4.1.2 **English GUM**

The English GUM corpus (Zeldes, 2017) is a yearly growing corpus and reached 180K tokens as of V8.0.0 (Jan 2022). The corpus comprises multiple layers of annotations on 12 genres and 193 documents. These layers include 1) document structure in TEI XML, such as sentences, paragraphs, headings, and figures; 2) syntactic annotations, such as tokenization, POS tags, and Universal Dependencies; 3) discourse annotations, such as RST, nested entities, and coreference. Thus, by combining different layers of XML, syntactic, and discourse annotations, I aligned EDUs in GUM to corresponding document structures. Specifically, the following XML tags are treated as annotations at the paragraph level: \textit{<p>} (document paragraphs), \textit{<head>} (document headings, e.g., document, section, and subsection titles), \textit{<table>}, \textit{<figure>}, \textit{<caption>}, \textit{<list>}, and \textit{<lg>} (poetry line groups).\(^4\) Since GUM requires that any token must be included in an \textit{<s>} span, \textit{<s>} is used as the only label at the sentence level.

\(^1\)https://catalog.ldc.upenn.edu/LDC2002T07
\(^2\)xxx stands for the four-digit integer that matches the main RST annotation file.
\(^3\)https://catalog.ldc.upenn.edu/LDC95T7
\(^4\)See https://wiki.gucorpling.org/gum/tei_markup_in_gum for XML definitions and examples.
4.1.3 Chinese GCDT

The GCDT corpus (Peng et al., 2022b) includes five genres and 50 documents from Mandarin Chinese for RST annotations. Paragraphs and sentences in GCDT are presented in a raw file with two new line feeds for paragraph boundaries and one for sentence boundaries. Like the RST-DT corpus, sentences and paragraphs are aligned to the EDU level in GCDT.

4.2 Creating a Unified Format

The alignment from EDUs to sentences or paragraphs is a surjective mapping. Namely, each sentence or paragraph can be linked to at least one EDU. Since the analysis focuses on structural information at the EDU level, the Discourse Dependency Structure (DDS) of RST (Morey et al., 2018) that assimilates other dependency structures, such as the Universal Dependencies (Nivre et al., 2016b) for syntax is the most convenient format. DDS can be converted from constituency-styled discourse trees and preserves the order of span attachments in a head-ordered dependency tree using a script provided in the GUM repo.\(^5\) The format with ten tab-separated columns and one EDU per line is a common way of visualizing the DDS framework that provides the possibility to include additional EDU-level information.

This chapter employs an extension of the DDS schema to include sentence-level and paragraph-level segmentation information. Specifically, the empty columns in a 10-column DDS file are used to record EDU-aligned sentence and paragraph information, as shown in Table 4.1. The EDU number (column 1), tokens within an EDU (column 2), order of head attachment (column 3), head index (column 7), and discourse relation (column 8) are standard entries in a DDS annotation. On top of that,\(^5\) See https://github.com/amir-zeldes/gum/blob/dev/_build/utils/rst2dep.py.

\(^5\)
the table adds the following information to the DDS format for structure-informed RST analyses while leaving columns 9 and 10 empty for future annotations:

- Column 4: the start and end token indexes of each EDU within the document;
- Column 5: the sentence index $i$ within the paragraph index $j$;
- Column 6: the sentence index $k$ within the entire document.

Table 4.1: An example of sentence and paragraph aligned Discourse Dependency Structure (DDS) for *gum_academic_chao.rsd*.

<table>
<thead>
<tr>
<th>EDU</th>
<th>Tokens</th>
<th>Order</th>
<th>Token Span</th>
<th>i-th S in j-th P</th>
<th>k-th S in doc</th>
<th>Head</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Early life</td>
<td>4</td>
<td>STARTTOK=1 ENDTOK=2</td>
<td>HEAD-1-S-1</td>
<td>S-1</td>
<td>3</td>
<td>organization-heading_r</td>
</tr>
<tr>
<td>2</td>
<td>Born in Tianjin with ancestry in Changzhou, Jiangsu province.</td>
<td>1</td>
<td>STARTTOK=3 ENDTOK=13</td>
<td>P-2-S-1</td>
<td>S-2</td>
<td>3</td>
<td>context-background_r</td>
</tr>
<tr>
<td>3</td>
<td>Chao went to the United States with a Boxer Indemnity Scholarship in 1910.</td>
<td>0</td>
<td>STARTTOK=14 ENDTOK=26</td>
<td>P-2-S-1</td>
<td>S-2</td>
<td>0</td>
<td>ROOT</td>
</tr>
<tr>
<td>4</td>
<td>to study mathematics and physics at Cornell University.</td>
<td>0</td>
<td>STARTTOK=27 ENDTOK=35</td>
<td>P-2-S-1</td>
<td>S-2</td>
<td>3</td>
<td>purpose-goal_r</td>
</tr>
<tr>
<td>5</td>
<td>where he was a classmate and lifelong friend of Hu Shih, the leader of the New Culture Movement.</td>
<td>0</td>
<td>STARTTOK=36 ENDTOK=55</td>
<td>P-2-S-1</td>
<td>S-2</td>
<td>4</td>
<td>elaboration-attribute_r</td>
</tr>
</tbody>
</table>

For example, EDUs 2-5 in Table 4.1 belong to the first sentence of the second paragraph (P-2-S-1) in *gum_academic_chao*. Structural information included in columns 5 and 6 facilitates the concatenation of EDUs and the categorization of discourse relation arcs within or across sentences and paragraphs. After the alignment procedures described in Section 4.1, three targeted corpora, i.e., RST-DT, GUM, and GCDT, are converted into this structurally augmented DDS format. Structurally augmented GUM and GCDT datasets are released on Github. However, due to license restrictions, the repository does not include structural-aligned RST-DT data.

4.3 **Structural Segment Statistics**

Previous studies in Sections 2.4 and 2.5 demonstrate that impactful structural segments are not limited to EDUs but also include sentences and paragraphs. With

---

6https://github.com/logan-siyao-peng/paragraph_analysis/data/
the extended DDS conversions available, this section presents the overall frequency statistics and dataset-specific characteristics of these structural segments in the three corpora: RST-DT, GUM, and GCDT.

4.3.1 Overall Frequency

RST-DT, GUM, and GCDT are from two languages and have different genre distributions: RST-DT is monotonically English news, whereas GUM and GCDT are balanced across 12 English and 5 Chinese genres. Consequently, the distributions of structural segments vary across these three datasets. Table 4.2 presents the per-document frequencies of tokens, EDUs, sentences, and paragraphs in the three corpora.

Table 4.2: Minimum, average and maximum numbers of tokens, EDUs, sentences, and paragraphs per document in the three RST corpora: RST-DT, GUM, and GCDT.

<table>
<thead>
<tr>
<th>corpus</th>
<th>#docs</th>
<th>#tokens/doc</th>
<th>#EDUs/doc</th>
<th>#sents/doc</th>
<th>#paras/doc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>min avg max</td>
<td>min avg max</td>
<td>min avg max</td>
<td>min avg max</td>
</tr>
<tr>
<td>RST-DT</td>
<td>385</td>
<td>32 548.2 2621</td>
<td>2 56.6 304</td>
<td>1 22.6 187</td>
<td>1 10.2 46</td>
</tr>
<tr>
<td>GUM</td>
<td>193</td>
<td>167 936.6 1878</td>
<td>14 119.7 250</td>
<td>9 51.9 216</td>
<td>1 14.8 61</td>
</tr>
<tr>
<td>GCDT</td>
<td>50</td>
<td>529 1257.1 1997</td>
<td>99 194.2 358</td>
<td>20 53.8 120</td>
<td>7 22.0 49</td>
</tr>
</tbody>
</table>

Generally speaking, GUM and GCDT annotate longer documents on average than RST-DT in all frequency counts: tokens, EDUs, sentences, and paragraphs. However, since all three datasets have different design principles, Sections 4.3.2 and 4.3.3 closely examine two aspects (namely document length and genre distribution) to understand further the frequency statistics presented in Table 4.2.

4.3.2 Short Documents in RST-DT

Unlike GUM and GCDT, which require documents to be around or over 1,000 tokens, news documents in RST-DT can be much shorter. Figure 4.1 shows the per-document token frequency distributions within the three corpora. While distributions for GUM
and GCDT in Figures 4.1b and 4.1c are centered around 1K tokens per document, the distribution of RST-DT in Figure 4.1a leans heavily towards the lower end.

![Figure 4.1](image)

(a) RST-DT

(b) GUM

(c) GCDT

Figure 4.1: The distribution of the number of tokens per document (\#tokens/doc) within the three corpora. The blue dashed lines indicate the medians of the corpora.

The numbers of paragraphs per document in the three corpora are similar to the number of tokens but not identical. In Figure 4.2, RST-DT’s per-document paragraph frequency still leans towards the left. However, GUM and GCDT are no more symmetrically centered. Statistically, Pearson’s correlation between the number of tokens and the number of paragraphs for RST-DT is high, with an $r$-value of 0.9142.
and a $p$-value of $2.42\times10^{-152}$. Whereas, GCDT has much lower correlation: $r=0.4754$ and $p=4.84\times10^{-04}$; and the lowest correlation for GUM: $r=0.1838$ and $p=1.05\times10^{-02}$.

Figure 4.2: The distribution of the number of paragraphs per document (# paras/doc) within the three corpora. The blue dashed lines indicate the medians of the corpora.

In multi-genre corpora such as GUM and GCDT, token length is no more the only factor that influences the number of paragraphs in a document. The properties of different genres and the consequent annotation decisions result in diverse characteristics in genre-wise statistics.
4.3.3 Genre Differences in GUM and GCDT

GUM V8.0.0 includes 193 documents from 12 genres, written or spoken. Table 4.3 presents the numbers of EDUs, sentences, and paragraphs per 1K tokens in the 12 genres with two types of ratios. At the document level, the table presents the lowest, the average (with standard deviation), and the highest ratios per 1K tokens among individual documents within the genre. At the genre level, the table lays out the overall averages across all documents within the specific genres.

Table 4.3: The numbers of EDUs, sentences, and paragraphs per 1K tokens within 12 genres of GUM. The minimum, average±standard deviation, maximum on document-level ratios, and genre-level average are included. Outlying genres on the higher and lower ends are colored in red and blue.

<table>
<thead>
<tr>
<th>genre</th>
<th>#docs</th>
<th>#EDUs/1K tokens</th>
<th>#sents/1K tokens</th>
<th>#paras/1K tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>document-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>min</td>
<td>mean±std</td>
<td>max</td>
</tr>
<tr>
<td>academic</td>
<td>18</td>
<td>93.0</td>
<td>114.8±11.9</td>
<td>130.8</td>
</tr>
<tr>
<td>bio</td>
<td>20</td>
<td>57.0</td>
<td>112.5±14.7</td>
<td>134.8</td>
</tr>
<tr>
<td>conversation</td>
<td>9</td>
<td>131.6</td>
<td>181.9±25.5</td>
<td>215.8</td>
</tr>
<tr>
<td>fiction</td>
<td>19</td>
<td>111.6</td>
<td>139.5±15.4</td>
<td>166.4</td>
</tr>
<tr>
<td>interview</td>
<td>19</td>
<td>131.1</td>
<td>131.0±12.3</td>
<td>161.8</td>
</tr>
<tr>
<td>news</td>
<td>23</td>
<td>80.9</td>
<td>109.5±14.2</td>
<td>133.7</td>
</tr>
<tr>
<td>reddit</td>
<td>18</td>
<td>102.2</td>
<td>135.7±13.3</td>
<td>153.1</td>
</tr>
<tr>
<td>speech</td>
<td>10</td>
<td>83.1</td>
<td>115.2±21.2</td>
<td>156.3</td>
</tr>
<tr>
<td>textbook</td>
<td>10</td>
<td>116.6</td>
<td>124.7±5.4</td>
<td>130.9</td>
</tr>
<tr>
<td>vlog</td>
<td>10</td>
<td>114.3</td>
<td>137.5±10.5</td>
<td>150.8</td>
</tr>
<tr>
<td>voyage</td>
<td>18</td>
<td>75.7</td>
<td>107.8±15.6</td>
<td>135.3</td>
</tr>
<tr>
<td>whow</td>
<td>19</td>
<td>114.0</td>
<td>130.3±16.1</td>
<td>171.1</td>
</tr>
<tr>
<td>all</td>
<td>193</td>
<td>75.7</td>
<td>126.5±22.6</td>
<td>215.8</td>
</tr>
</tbody>
</table>

Two outlying genres in the dataset are conversation and vlog. These are among the most spoken genres in GUM. The EDUs and sentences in conversation are much shorter than in other genres. Namely, the numbers of EDUs and sentences over 1K tokens are much larger when compared to other genres. Additionally, since texts in both genres are transcribed from speech productions, no natural paragraph boundaries are imposed, resulting in only one paragraph per document. Consequently, paragraph containment analysis is less interesting for these two genres. Moreover, other genres that are more on the spoken or informal end also show a larger quantity of
EDUs, sentences, and paragraphs per 1K tokens, for example, fiction, reddit, vlog, whow.

The Chinese GCDT corpus includes five of the twelve genres from GUM but focuses on longer documents, mostly in written genres. A similar per-1K-token statistic is presented in Table 4.4. Due to its design, GCDT omits extreme spoken outliers such as conversation or vlog from GUM. However, similar to GUM, whow texts in GCDT have much shorter EDUs, sentences, and paragraphs when compared to the other genres. One important factor is the abundance of ordered and unordered listings in how-to guide documents that shorten the length of this genre’s structural segments.

Table 4.4: The numbers of EDUs, sentences, and paragraphs per 1,000 tokens within five genres of GCDT. The minimum, average±standard deviation, and maximum on document-level ratios, as well as a genre-level average, are provided.

<table>
<thead>
<tr>
<th>corpus</th>
<th>#docs</th>
<th>#EDUs/1K tokens</th>
<th>#sents/1K tokens</th>
<th>#paras/1K tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>document-level</td>
<td>document-level</td>
<td>document-level</td>
</tr>
<tr>
<td></td>
<td></td>
<td>min</td>
<td>mean±std</td>
<td>max</td>
</tr>
<tr>
<td>academic</td>
<td>10</td>
<td>115.1</td>
<td>145.0±23.4</td>
<td>174.8</td>
</tr>
<tr>
<td>bio</td>
<td>10</td>
<td>122.2</td>
<td>151.1±15.5</td>
<td>181.2</td>
</tr>
<tr>
<td>interview</td>
<td>10</td>
<td>136.0</td>
<td>157.1±15.9</td>
<td>182.2</td>
</tr>
<tr>
<td>news</td>
<td>10</td>
<td>86.2</td>
<td>147.5±25.7</td>
<td>182.0</td>
</tr>
<tr>
<td>whow</td>
<td>10</td>
<td>150.2</td>
<td>177.7±17.0</td>
<td>209.8</td>
</tr>
<tr>
<td>all</td>
<td>50</td>
<td>86.2</td>
<td>155.7±22.5</td>
<td>209.8</td>
</tr>
</tbody>
</table>

To summarize, Section 4.3 rolls out the overall frequency of structural segments and delves into two impacting factors: the distribution of document length in RST-DT and the genre diversity in GUM and GCDT. The following Sections 4.4 and 4.5 set tokens and sentences aside and focus keenly on associating paragraph structures with RST. Section 4.4 presents a containment analysis that examines the correlation between paragraph boundaries and the boundaries of larger discourse units. Section 4.5 differentiates inter-paragraph discourse relations from intra-paragraph ones and evaluates the structural preferences for various relation labels. Both analyses profit from anno-
tations in multiple languages, corpora, and genres and double annotations in RST-DT and GCDT.

4.4 Paragraph Containment

Paragraphs are explicit textual breaks in a document to facilitate reading. Previous studies summarize a few peculiarities regarding the beginning of a paragraph, including rhetorically motivated sentences, discourse signals, and proper nouns (Filippova and Strube, 2006; McGee, 2014). It is also claimed that paragraph breaks are not necessarily associated with topic changes (McGee, 2014). To better understand the correlations between discourse units and paragraphs, this section conducts a paragraph containment analysis to examine how often paragraphs are properly contained in an RST structure and what are the common types of containment violations. Specifically, Section 4.4.1 brings out the definition of paragraph containment and Section 4.4.2 categories four major violation types. Sections 4.4.3-4.4.6 exemplify major types of paragraph containment violations using RST-DT’s double annotations. Section 4.4.7 conducts genre-wise evaluations of the containment analysis and concludes the section.

4.4.1 Definition and Procedure

Proper paragraph containment in an RST tree is defined as the following:

\[ \text{A paragraph is properly contained in an RST tree if the paragraph contains at most one outgoing dependency edge.} \]

The core of proper paragraph containment is that EDUs in a paragraph form a proper, single-rooted subtree before merging with discourse units from other paragraphs. In other words, a paragraph is not properly contained if two or more nodes in the
paragraph form external dependencies. Figure 4.3 presents an example of a properly paragraph-contained RST tree. The five boxes are considered five paragraphs appearing at the beginning of a wikihow document:

- Box 1 is the title of the document, i.e., “How to pack your possession when moving”;
- Box 2 is an introductory paragraph;
- Box 3 is the header of the first section, i.e., “Steps”;
- Boxes 4 and 5 are two parallel paragraphs in a multi-nucleus relation within the first section.

Figure 4.3: An example of properly-contained paragraphs in gum_whow_packing.

For nucleus-satellite relations in an RST tree, the definition of proper containment is quite clear: proper containment does not allow two satellites from the same paragraph relating to the same nucleus or different nuclei outside of the paragraph. For example, Boxes 1-3 in Figure 4.3 are properly contained since they each have only one outgoing dependency edge. On the other hand, multiple sisters of the same parent are permitted for multi-nucleus relations, such as the two parallel paragraphs in Boxes 4-5.
4.4.2 Outgoing Dependencies and Violation Types

The following subsections conduct the paragraph containment analysis using the augmented DDS data described in Section 4.2. The number of outgoing edges (i.e., the number of discourse units that point to an external node) is computed for each paragraph. Paragraphs with two or more outgoing discourse edges are treated as violations of proper RST containment. Table 4.5 below presents the types and frequencies of paragraph containment violations in three corpora: RST-DT, GUM, and GCDT.

**Table 4.5: Types of violated paragraphs in RST-DT, GUM, and GCDT.**

<table>
<thead>
<tr>
<th>Containment violation types</th>
<th>RST-DT</th>
<th>GUM</th>
<th>GCDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>same-parent-EDU</td>
<td>220 (5.60%)</td>
<td>109 (3.74%)</td>
<td>22 (2.03%)</td>
</tr>
<tr>
<td>same-parent-para and diff-parent-EDU</td>
<td>34 (0.87%)</td>
<td>21 (0.72%)</td>
<td>6 (0.55%)</td>
</tr>
<tr>
<td>has-root</td>
<td>3 (0.08%)</td>
<td>12 (0.41%)</td>
<td>3 (0.28%)</td>
</tr>
<tr>
<td>two-way</td>
<td>55 (1.40%)</td>
<td>68 (2.33%)</td>
<td>14 (1.29%)</td>
</tr>
<tr>
<td>other violations</td>
<td>53 (1.35%)</td>
<td>58 (1.99%)</td>
<td>4 (0.37%)</td>
</tr>
<tr>
<td>Total number of paragraphs with outgoing violation</td>
<td>365 (9.29%)</td>
<td>268 (9.20%)</td>
<td>49 (4.52%)</td>
</tr>
<tr>
<td>Total number of paragraphs in the corpus</td>
<td>3930</td>
<td>2914</td>
<td>1083</td>
</tr>
</tbody>
</table>

The last two rows of Table 4.5 show low percentages of containment-violated paragraphs in the three corpora: 9.29% for RST-DT (365 out of 3930), 9.20% for GUM (268 out of 2914), and 4.52% for GCDT (49 out of 1083). In other words, over 90% paragraphs are properly contained in all three corpora. This result aligns with motivations in previous theories and experiments to treat inter-paragraph and intra-paragraph relations separately. Nevertheless, these containment-violated percentages guarantee a substantial failure rate for multi-stage RST parsing models discussed in Section 2.5 since these models are based on the underlying assumption that 100% paragraphs are properly contained in an RST tree.

Table 4.5 further categorizes the outgoing edges of violated paragraphs to understand why such violations, though not preferred, still appear in an RST structure. Four
major violation types are listed below by examining the position(s) of the external parent node(s) to which the multiple outgoing edges are attached. Figure 4.4 shows some simplified examples for these four violation types, with $A$, $B$, $C$ representing paragraphs and $A1$, $A2$ for sentences.

![Figure 4.4: Four common containment violation types.](image)

- **same-parent-EDU**: all outgoing edges of the paragraph attach to the same external parent EDU; this implies that they are also attached to the same paragraph; for example, $B1-B3$ all attach to $A1$ in Figure 4.4a;
- **same-parent-para**: all outgoing edges of the paragraph attach to the same external paragraph; for example, $B1$, $B2$ attach to $A2$ and $B3$ to $A1$ in Figure 4.4b where $A1$ and $A2$ are two different sentences in the same paragraph;
- **has-root**: any outgoing edge of the paragraph is the root of the document; for example, $A1$ is the root in Figure 4.4c;
- **two-way**: some outgoing edges of the paragraph are attached to the external left parent(s), and others are attached to the right; for example, $B1$ attaches to $A1$ on the left but $B3$ to $C1$ on the right in Figure 4.4d.
Table 4.5 also presents frequencies of the four violation categories in three corpora. The category where multiple outgoing edges attach to the same external parent EDU (same-parent-EDU) accounts for almost half of the violations in the three corpora. In other words, instead of combining to modify an external parent, they each modify the same parent with possibly different relation labels.

To better understand these four categories, Sections 4.4.3 to 4.4.6 below present double-annotated examples from RST-DT to demonstrate different annotation decisions regarding discourse coherence and structural integrity. Sections 4.4.3 and 4.4.4 bring examples with multiple outgoing edges pointing to the same external EDU or paragraph, and Sections 4.4.5 and 4.4.6 show containment violations with root nodes and outgoing edges in two opposite directions.

4.4.3 Same-Parent-EDU

Many instances of same-parent-EDU violations occur in RST-DT’s 53 double-annotated documents. Among them, Figure 4.5a presents a notable example where four outgoing edges (marked by red arrows) from the second paragraph (marked by a blue box) modify the discourse span DU_22-24 in the first paragraph. The four satellite spans serve different functions when modifying the same nucleus: elaboration (DU_25-26), evaluation (EDU_27), consequence (EDU_28), and elaboration (DU_29-33). Figure 4.5b presents an alternative analysis from RST-DT’s double annotation where only two outgoing edges occur in the second paragraph: two levels of elaboration from DU_25-28 and DU_29-33. Since “niche-itis” is mentioned in EDU_23 and EDU_25, annotation 4.5b chooses to interpret EDU_27 and EDU_28 as modifications of the second occurrence and thus conforms better to the paragraph containment principle. Annotation 4.5b can be resolved by making DU_25-28 and
DU_29-33 a multinuclear relation and jointly elaborating DU_22-24, and would thus avoid paragraph containment violation.

Figure 4.5: An example of paragraph containment violation with multiple outgoing edges pointing to the same external parent EDU from two versions of *wsj_1123* annotations.

4.4.4 Same-Parent-Para

Besides attaching to the same external EDU, outgoing edges can also attach to different levels of discourse units in the same paragraph. Figure 4.6a presents an
example of containment violation with three outgoing edges from the second paragraph attached to different discourse units in the first paragraph. EDU_154 attaches directly to EDU_153, explaining the dollar rose. DU_155-156 joins with the entire first paragraph, elaborating on Japanese government bonds ending lower. DU_157-161 finally forms a multinuclear relation with the previous units. In sum, two satellites from the second paragraph modify the first paragraph, and one modifies only EDU_153. Figure 4.6b presents an alternative annotation with only two outgoing edges from the second paragraph: DU_156-158 and DU_159-163. In this annotation, two parts of the second paragraph merge with the first paragraph, one as an explanation satellite and another as a multinuclear list.

The two sets of double annotations in Figures 4.5 and 4.6 illustrate a dilemma between information attribution and paragraph integrity. The main annotations tend to attach the satellite information to the topic’s first or most explicit occurrence. In contrast, the second annotations converge better with paragraph integrity and prefer attaching to relevant DUs within the same paragraph. Though the two annotators value annotation constraints differently, both annotation versions are valid and intelligible in the RST framework.

4.4.5 Has-Root

In a discourse structure, the most central unit (or the first item in the highest multinuclear relation) is the root of the tree and is modified by the rest of the document. Ideally, the root edge should be the only outgoing edge in a paragraph. However, a few violations are observed in the three datasets where a paragraph includes one root unit and other outgoing discourse dependencies. Figure 4.7a presents a paragraph that leads the first item in the highest multinuclear contrast relation and thus includes an outgoing root edge. An intriguing containment violation appears on the left edge of
(a) Main annotation with three outgoing edges from the second paragraph.

(b) Double annotation with two outgoing edges from the second paragraph.

Figure 4.6: An example of paragraph containment violation with multiple outgoing edges pointing to the same paragraph from two versions of wsj_1322 annotations.
the paragraph. The leftmost EDU_19 of the paragraph modifies an external discourse unit to its left, and the resulting larger discourse unit modifies DU_20-22 in return. The back-and-forth relations of EDU_19 and its parent units violate the proper containment of the paragraph it resides in. However, if such a structural constraint does not exist or such paragraph segmentation is absent to the annotator, relating EDU_19 to the previous paragraph is rhetorically motivated. Another annotation version of the same paragraph is shown in Figure 4.7b. The left edge does not violate paragraph containment anymore. However, such a back-and-forth relation hierarchy now occurs on the right edge, where DU_19-27 attaches to DU_28-102, and the joined bigger unit reversely modifies EDU_18 in the paragraph. The two annotations reveal that enforcing strict paragraph boundaries could intervene in the information flow between paragraphs and motivate future containment analyses based on structural units larger than paragraphs, such as topic segments (Xing et al., 2022).

4.4.6 Two-Way

The last property in Table 4.5 concerns the directions of the multiple outgoing edges in a containment-violated paragraph. Figure 4.8a presents a difficult situation where the three outgoing edges point in different directions. EDU_82 forms a multinuclear statement relation to its sisters on the left; EDU_83 creates a multinuclear contrast relation to the right; and DU_84-86 modifies an external right discourse unit as a satellite consequence relation. A few levels up, the multinuclear DU_74-134 spans over the target paragraph and the left and right neighbors associated with those outgoing edges. Thus, better structural contingency and containment would be achieved if one executes a containment analysis at a higher structural level. Alternatively, Figure 4.8b shows a second annotation version without containment violation. However, multinuclear relations, such as list, contrast, and sequence dominate the discourse tree of
Figure 4.7: An example of paragraph containment violation with root node in the main annotation of *wsj_1394*, and two outgoing edges in the double annotation.

(a) Main annotation with a root node and another outgoing edge.

(b) Double annotation with two outgoing edges.
the paragraph and loses the nuclearity and order of attachment information that is preferred in the first annotation version.

The current RST annotation guidelines allow some freedom in annotators’ preference between structural integrity and discourse coherence, as shown by the double-annotated RST-DT examples in Sections 4.4.3 to 4.4.6. Two future areas of studies are motivated: discourse integrity on larger structural spans such as topic segments and ambiguity and agreement analyses regarding double-annotated RST trees. More specifically, a future hypothetical question could be: are containment violations in one annotation version correlated with or predictive of inter-annotator disagreements between different annotation versions? Further insights into these questions can facilitate understanding discourse structures and improve RST parsing.

4.4.7 Genre-wise Violations in GUM and GCDT

In addition to the intriguing annotation decisions between structural integrity and discourse coherence in RST-DT’s double annotations, the genre-wise distribution of paragraph containment violations is also worth noticing. Since RST-DT only contains news documents, this section reports genre-wise paragraph containment violations in GUM and GCDT. Following previous discussions, violations with multiple outgoing edges are divided into five types: 1) all outgoing edges pointing to the same external parent EDU \( \textit{same-parent-EDU} \); 2) all outgoing edges pointing to different parent EDUs in the same paragraph \( \textit{same-parent-para} \); 3) root edge is contained in the outgoing edges \( \textit{has-root} \); 4) the outgoing edges point in more than one direction \( \textit{two-way} \); 5) other violations.

Table 4.6 presents the distribution of the five violation types in the twelve GUM genres. Firstly, since \textit{conversation} and \textit{vlog} (in blue) are transcriptions of spoken texts that do not exhibit paragraph boundaries, paragraph violation analysis does not apply.
(a) Main annotation with three outgoing edges, one to the left and two to the right.

(b) Double annotation without containment violation.

Figure 4.8: An example of paragraph containment violation with outgoing edges in two directions in the main annotation of \textit{wsj\_1387}, and a double annotation without containment violation.
to the two genres. On the other hand, though genres such as fiction, reddit, and speech (in red) contain a large number of conversational texts, they are usually written in multiple short paragraphs (see Table 4.3). Added that many entities and topics flow continuously throughout conversations, paragraph boundaries have a higher risk of intervening discourse coherence in these genres. As a result, these three genres are among the ones that have the highest paragraph containment violations. Another highly violated genre is academic (in red), which has a larger ratio of has-root and two-way violations. Contrary to the previous three genres, academic is on the other end of Table 4.3 with the longest paragraphs\(^7\) on average among GUM genres. Due to the lengthy paragraphs, there is a higher risk in academic where two different portions of a paragraph modify different nuclear units as schematized in Figure 4.4d.

Table 4.6: Percentages of paragraph containment violation types per genre in GUM.

<table>
<thead>
<tr>
<th>genre</th>
<th>same-parent-EDU</th>
<th>same-para &amp; diff-parent-EDU</th>
<th>has-root</th>
<th>two-way</th>
<th>others</th>
<th>all violations</th>
<th>#paras</th>
</tr>
</thead>
<tbody>
<tr>
<td>academic</td>
<td>3.61%</td>
<td>0%</td>
<td>1.20%</td>
<td>4.22%</td>
<td>2.41%</td>
<td>11.45%</td>
<td>166</td>
</tr>
<tr>
<td>bio</td>
<td>2.87%</td>
<td>0.08%</td>
<td>0%</td>
<td>2.87%</td>
<td>1.08%</td>
<td>7.89%</td>
<td>279</td>
</tr>
<tr>
<td>conversation</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>9</td>
</tr>
<tr>
<td>fiction</td>
<td>1.50%</td>
<td>0.30%</td>
<td>0.30%</td>
<td>2.99%</td>
<td>2.40%</td>
<td>7.49%</td>
<td>334</td>
</tr>
<tr>
<td>news</td>
<td>5.33%</td>
<td>0.94%</td>
<td>0%</td>
<td>0.31%</td>
<td>1.57%</td>
<td>8.15%</td>
<td>319</td>
</tr>
<tr>
<td>reddit</td>
<td>3.57%</td>
<td>0.60%</td>
<td>0.89%</td>
<td>4.46%</td>
<td>2.38%</td>
<td>11.90%</td>
<td>336</td>
</tr>
<tr>
<td>speech</td>
<td>5.84%</td>
<td>1.95%</td>
<td>0.65%</td>
<td>0.00%</td>
<td>0.65%</td>
<td>9.09%</td>
<td>154</td>
</tr>
<tr>
<td>textbook</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>10</td>
</tr>
<tr>
<td>vlog</td>
<td>1.47%</td>
<td>0.88%</td>
<td>0%</td>
<td>0.29%</td>
<td>0.29%</td>
<td>2.94%</td>
<td>340</td>
</tr>
<tr>
<td>voyage</td>
<td>3.03%</td>
<td>0%</td>
<td>0.47%</td>
<td>0.93%</td>
<td>1.40%</td>
<td>5.83%</td>
<td>429</td>
</tr>
<tr>
<td>whow</td>
<td>3.72%</td>
<td>0.72%</td>
<td>0.41%</td>
<td>2.32%</td>
<td>1.98%</td>
<td>9.14%</td>
<td>2933</td>
</tr>
</tbody>
</table>

Figure 4.9 presents an example of multiple outgoing edges in the second paragraph of a fiction document. The fragment is a two-turn dialogue between two speakers: Jenna in the first paragraph and Robert in the second. Robert responds to Jenna’s one-sentence paragraph with three different discourse intentions: 1) restating that this

\(^7\)Except for the conversation and vlog genres in which a document is a single paragraph.
does not happen in San Francisco; 2) contrasting that earthquakes and wood sprites exist in SF; 3) and evaluating Jenna’s arrival in New York. A different annotation version of the dialogue could avoid a paragraph containment violation. However, it would eliminate the three direct associations with different discourse functions and reduce the explicitness of the current discourse analysis.

![Figure 4.9: An example of paragraph containment violation with three outgoing edges from the second paragraph of gum_fiction_pixies.](image)

In sum, paragraph-level containment analysis is not as compatible with conversational data. In one scenario, conversation and vlog do not include paragraph segmentation, and paragraph-level containment would not apply. In another, fiction and reddit are multiple short paragraphs that might not be a large enough structural unit in discourse analysis. Moreover, since conversational genres are robust in switching between discussion topics and cross-referencing entities and events, differentiating inter-paragraph versus intra-paragraph discourse structures would be less linguistically motivated.
Table 4.7 presents similar genre-wise statistics for the Chinese GCDT corpus. By design, GCDT focuses on written genres and longer documents and only includes five of the twelve genres from GUM. Among the selected genres, whow (in red) creates the most paragraph containment violations. Most whow violations are cases of multiple outgoing edges pointing to the same external EDU. One large contributing factor is the paragraph status of listed items in these how-to guides. The current preprocessing steps introduced in Section 4.1 include <list> as an equivalence of a paragraph. Figure 4.10 presents an example where the second paragraph lists warning items. Alternatively, one can argue that each item should be treated as a separate paragraph rather than the whole list as one paragraph. Regardless of the annotation decision, itemization represents how-to guides and challenges the paragraph containment analysis for RST structures.

Table 4.7: Percentages of paragraph containment violations by genre in GCDT.

<table>
<thead>
<tr>
<th>genre</th>
<th>same-parent-EDU</th>
<th>same-para &amp; diff-parent-EDU</th>
<th>has-root</th>
<th>two-way</th>
<th>others</th>
<th>all violations</th>
<th>#paras</th>
</tr>
</thead>
<tbody>
<tr>
<td>academic</td>
<td>1.12%</td>
<td>0.56%</td>
<td>0.56%</td>
<td>2.25%</td>
<td>0.56%</td>
<td>5.06%</td>
<td>178</td>
</tr>
<tr>
<td>bio</td>
<td>1.99%</td>
<td>1.00%</td>
<td>0.00%</td>
<td>1.49%</td>
<td>0.50%</td>
<td>4.98%</td>
<td>201</td>
</tr>
<tr>
<td>interview</td>
<td>0.48%</td>
<td>0.00%</td>
<td>0.48%</td>
<td>1.90%</td>
<td>0.00%</td>
<td>2.86%</td>
<td>210</td>
</tr>
<tr>
<td>news</td>
<td>1.38%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.46%</td>
<td>1.84%</td>
<td>217</td>
</tr>
<tr>
<td>whow</td>
<td>4.33%</td>
<td>1.08%</td>
<td>0.36%</td>
<td>1.08%</td>
<td>0.36%</td>
<td>7.22%</td>
<td>277</td>
</tr>
<tr>
<td>all</td>
<td>2.03%</td>
<td>0.55%</td>
<td>0.28%</td>
<td>1.29%</td>
<td>0.37%</td>
<td>4.52%</td>
<td>1083</td>
</tr>
</tbody>
</table>

When comparing English GUM and Chinese GCDT, all genres in Chinese have a lower violation rate except for whow. Nevertheless, both Table 4.6 and Table 4.7 show that interview and news are below average violation rates while academic above average. A few factors might contribute to the overall low violation rate in GCDT. Firstly, all GCDT documents are annotated by the author of this dissertation, who might be biased towards fewer paragraph violations when annotating GCDT and preparing for his dissertation. Secondly, GUM is a collection of classroom annotations, and each
Figure 4.10: An example of paragraph containment violation with three outgoing edges from the second paragraph of *gcdt_whow_hiking*.

student annotator only annotates one document and thus might not be as familiar with the RST guidelines. Lastly, GCDT is a relatively small dataset, and documents are, on average, longer than GUM. In the future, conducting English and Chinese RST annotations on parallel documents would be a promising research direction to analyze language-dependent discourse phenomena regarding paragraph containment.

Nevertheless, GCDT's high paragraph containment rate is an interesting corpus phenomenon. It sheds light on the parsing results in Section 5.2 and Section 5.4. Namely, even though GUM is larger and more diverse in genres, parsing performances on GCDT consistently win GUM by 3-5 points when using the same model and under the same training scenario. The low containment violation rate in GCDT is undoubtedly one factor that facilitates RST parsing. At the end of the day, I hope the GCDT corpus can extend beyond this dissertation and include more annotations.
from multiple annotators to continue benefiting RST parsing for Chinese and across languages.

4.5 Intra- versus Inter-paragraph Relation Distributions

In addition to the overall high paragraph containment rate in Section 4.4, particularly in more traditionally written genres, the different distributions of discourse relation labels in intra-paragraph and inter-paragraph scenarios motivate structurally informed discourse analysis. This section examines intra- versus inter-paragraph relation distributions in three corpora: RST-DT, GUM, and GCDT, and in multiple genres.

4.5.1 Relation Distributions in RST-DT

RST-DT is the most widely experimented RST dataset with uniquely news articles. Table 4.8 presents the frequencies and percentages of relations in the Discourse Dependency Structure (DDS) in the three scenarios: intra-paragraph relations, inter-paragraph relations, and the subset of inter-paragraph relations that caused paragraph containment violations. The table sorts the 18 relation classes (16 + same-unit + textual-organization) plus root in RST-DT by their intra-paragraph frequencies. Due to the tree structure of an RST annotation, the number of intra-paragraph discourse relations is four times the size of inter-paragraph relations. The less annotated relations on the higher level result in data sparsity when training for macro-level discourse parsing tasks (Jia et al., 2018; Hou et al., 2020; Zhang et al., 2020).

Four key observations can be drawn from Table 4.8. Firstly, many relations are limited to one scenario or another by definition (with individual exceptions). Attribution (in blue) annotates reported speech, and the source of information occurs...
Table 4.8: Frequency and percentage of RST-DT relation classes in intra-paragraph, inter-paragraph, and containment-violated inter-paragraph scenarios.

<table>
<thead>
<tr>
<th>relation</th>
<th>nucularity</th>
<th>intra-paragraph</th>
<th>inter-paragraph</th>
<th>containment-violated inter-paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>elaboration</td>
<td>mono</td>
<td>6213 35.7%</td>
<td>1689 38.4%</td>
<td>382 49.9%</td>
</tr>
<tr>
<td>attribution</td>
<td>mono</td>
<td>3068 17.6%</td>
<td>2 0.0%</td>
<td>1 0.1%</td>
</tr>
<tr>
<td>same-unit</td>
<td>multi</td>
<td>1403 8.1%</td>
<td>1 0.0%</td>
<td>1 0.1%</td>
</tr>
<tr>
<td>joint</td>
<td>multi</td>
<td>1344 7.7%</td>
<td>643 14.6%</td>
<td>35 4.6%</td>
</tr>
<tr>
<td>contrast</td>
<td>both</td>
<td>893 5.1%</td>
<td>237 5.4%</td>
<td>62 8.1%</td>
</tr>
<tr>
<td>explanation</td>
<td>mono</td>
<td>751 4.3%</td>
<td>235 5.3%</td>
<td>39 5.1%</td>
</tr>
<tr>
<td>background</td>
<td>mono</td>
<td>730 4.2%</td>
<td>207 4.7%</td>
<td>54 7.1%</td>
</tr>
<tr>
<td>cause</td>
<td>both</td>
<td>588 3.4%</td>
<td>105 2.4%</td>
<td>33 4.3%</td>
</tr>
<tr>
<td>enablement</td>
<td>mono</td>
<td>556 3.2%</td>
<td>12 0.3%</td>
<td>5 0.7%</td>
</tr>
<tr>
<td>temporal</td>
<td>both</td>
<td>458 2.6%</td>
<td>72 1.6%</td>
<td>19 2.5%</td>
</tr>
<tr>
<td>evaluation</td>
<td>both</td>
<td>343 2.0%</td>
<td>256 5.8%</td>
<td>66 8.6%</td>
</tr>
<tr>
<td>condition</td>
<td>both</td>
<td>308 1.8%</td>
<td>20 0.5%</td>
<td>6 0.8%</td>
</tr>
<tr>
<td>comparison</td>
<td>both</td>
<td>263 1.5%</td>
<td>40 0.9%</td>
<td>10 1.3%</td>
</tr>
<tr>
<td>manner-means</td>
<td>mono</td>
<td>213 1.2%</td>
<td>13 0.3%</td>
<td>7 0.9%</td>
</tr>
<tr>
<td>summary</td>
<td>mono</td>
<td>141 0.8%</td>
<td>82 1.9%</td>
<td>12 1.6%</td>
</tr>
<tr>
<td>topic-comment</td>
<td>both</td>
<td>78 0.4%</td>
<td>78 1.8%</td>
<td>17 2.2%</td>
</tr>
<tr>
<td>textual-organization</td>
<td>multi</td>
<td>38 0.2%</td>
<td>119 2.7%</td>
<td>6 0.8%</td>
</tr>
<tr>
<td>topic-change</td>
<td>both</td>
<td>6 0.0%</td>
<td>199 4.5%</td>
<td>7 0.9%</td>
</tr>
<tr>
<td>root</td>
<td>N/A</td>
<td>0 0.0%</td>
<td>385 8.8%</td>
<td>3 0.4%</td>
</tr>
<tr>
<td>Total</td>
<td>N/A</td>
<td>17394 100.0%</td>
<td>4395 100.0%</td>
<td>765 100.0%</td>
</tr>
</tbody>
</table>
together with the reported content within a paragraph. *Same-unit* (in blue) links discontinuous text fragments within a sentence and thus can be only within a paragraph (and more specifically within a sentence). In contrast, *topic-change* (in red) links larger discourse units when a topic shifts or drifts from one area to another and only appears across paragraphs. Secondly, many top relation classes are frequent both between and across paragraphs. These include *elaboration, joint, contrast, explanation, background, cause, temporal* and *evaluation*. Thirdly, most of the least frequent relation classes are prone to occur in one environment or the other. Most discourse relations in the third box have a much higher frequency within a paragraph than across paragraphs, e.g., *enablement, condition, comparison* and *manner-means* (in blue). Document-level structuring relations appear at least twice more frequently in the inter-paragraph scenario, e.g., *summary, topic-comment, textual-organization* in the bottom box (in red). Lastly, Table 4.8 also compares the rate of containment-violated inter-paragraph relations against all inter-paragraph relations. Results show that mononuclear containment-violated inter-paragraph relations ratios are higher than their general ratios of all inter-paragraph cases, e.g., *elaboration, background, enablement, manner-means*. Section 4.4 illustrated many examples when multiple satellites contribute to the same nucleus with different discourse functions. In addition, relation classes on the more multinuclear side have a lower ratio of containment-violated relations, e.g., *joint, textual-organization, topic-change*.

In sum, the ratios of intra- versus inter-paragraph relations are different, especially in the second half of less frequent discourse relation classes. Containment violations of inter-paragraph relations are more prone to mononuclear relations, with the annotation goal to associate the satellite with the most relevant nucleus. Unavoidably, the unbalanced relation classes in RST-DT create a long-tail problem for differentiating
the less frequent relations that are sparse in terms of the number of annotations and
diverse in terms of their appearance in the macro- versus micro-level structures.

4.5.2 Overall Relation Distributions in GUM

Though converging on the same underlying RST tree structure, GUM and GCDT
differ from RST-DT in two aspects. The relation types decrease from RST-DT’s 78
to 32, and the number of classes to 15 to ease annotation decisions. Moreover, GUM
and GCDT include more genres than news and offer the opportunity to conduct
in-depth comparisons of discourse structures over text domains.

Table 4.9 presents the percentages of the 15 relation classes in 12 genres of GUM
V8.0.0, with separate statistics on intra-paragraph versus inter-paragraph relations.
The right-most numeric column in the upper and lower boxes demonstrates GUM’s
overall intra-paragraph and inter-paragraph relation percentages in all genres. The
percentages color-coded in cyan indicate an at least twice higher percentage than the
opposite scenario. For example, root is by design unique to inter-paragraph (occupy-
ing 5.5% inter-paragraph relations in all genres) and same-unit is unique to intra-
paragraph (6.1% inter-paragraph relations in all genres). Relation classes such as
contingency, mode, and purpose also barely have any appearance in inter-paragraph
situations. Two relations for higher-order text organization, joint and organization,
occur much (about 35%) more frequently in inter-paragraph circumstances. Most
other relation classes have a higher rate in intra-paragraph scenarios, including the
frequently-used adversative, attribution, causal, elaboration, and explanation. Never-
theless, context and evaluation are more versatile and frequent as intra- and inter-
paragraph relations.
Table 4.9: Percentages of intra-paragraph and inter-paragraph relation classes by genre in GUM.

### Intra-paragraph relations

<table>
<thead>
<tr>
<th>relation</th>
<th>nuclearity</th>
<th>academic</th>
<th>bio</th>
<th>fiction</th>
<th>interview</th>
<th>news</th>
<th>reddit</th>
<th>speech</th>
<th>textbook</th>
<th>voyage</th>
<th>how</th>
<th>all</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>adversative</td>
<td>both</td>
<td>6.2%</td>
<td>5.6%</td>
<td>8.1%</td>
<td>7.4%</td>
<td>5.2%</td>
<td>11.4%</td>
<td>5.5%</td>
<td>8.5%</td>
<td>8.5%</td>
<td>8.7%</td>
<td>7.6%</td>
<td>adversative</td>
</tr>
<tr>
<td>attribution</td>
<td>mono</td>
<td>3.3%</td>
<td>1.8%</td>
<td>11.4%</td>
<td>10.0%</td>
<td>16.1%</td>
<td>9.1%</td>
<td>7.1%</td>
<td>6.8%</td>
<td>1.1%</td>
<td>1.6%</td>
<td>6.8%</td>
<td>attribution</td>
</tr>
<tr>
<td>causal</td>
<td>mono</td>
<td>3.0%</td>
<td>3.8%</td>
<td>5.6%</td>
<td>4.9%</td>
<td>5.7%</td>
<td>5.1%</td>
<td>3.5%</td>
<td>4.1%</td>
<td>5.5%</td>
<td>3.2%</td>
<td>4.6%</td>
<td>causal</td>
</tr>
<tr>
<td>context</td>
<td>mono</td>
<td>5.6%</td>
<td>7.4%</td>
<td>12.4%</td>
<td>9.7%</td>
<td>7.7%</td>
<td>8.8%</td>
<td>8.7%</td>
<td>6.9%</td>
<td>6.9%</td>
<td>6.6%</td>
<td>8.0%</td>
<td>context</td>
</tr>
<tr>
<td>contingency</td>
<td>mono</td>
<td>1.1%</td>
<td>0.2%</td>
<td>0.9%</td>
<td>1.1%</td>
<td>1.6%</td>
<td>4.3%</td>
<td>1.0%</td>
<td>2.5%</td>
<td>3.1%</td>
<td>8.6%</td>
<td>2.4%</td>
<td>contingency</td>
</tr>
<tr>
<td>elaboration</td>
<td>mono</td>
<td>28.5%</td>
<td>29.9%</td>
<td>17.4%</td>
<td>24.8%</td>
<td>26.8%</td>
<td>19.1%</td>
<td>26.5%</td>
<td>28.4%</td>
<td>27.7%</td>
<td>19.1%</td>
<td>22.3%</td>
<td>elaboration</td>
</tr>
<tr>
<td>evaluation</td>
<td>mono</td>
<td>1.6%</td>
<td>1.2%</td>
<td>3.6%</td>
<td>3.9%</td>
<td>1.7%</td>
<td>4.3%</td>
<td>2.7%</td>
<td>1.1%</td>
<td>3.6%</td>
<td>1.6%</td>
<td>3.5%</td>
<td>evaluation</td>
</tr>
<tr>
<td>explanation</td>
<td>mono</td>
<td>14.1%</td>
<td>11.2%</td>
<td>4.0%</td>
<td>4.6%</td>
<td>2.9%</td>
<td>7.7%</td>
<td>5.8%</td>
<td>6.3%</td>
<td>5.7%</td>
<td>6.9%</td>
<td>6.6%</td>
<td>explanation</td>
</tr>
<tr>
<td>joint</td>
<td>multi</td>
<td>12.6%</td>
<td>22.2%</td>
<td>21.1%</td>
<td>15.5%</td>
<td>15.0%</td>
<td>14.5%</td>
<td>16.7%</td>
<td>14.3%</td>
<td>21.8%</td>
<td>19.5%</td>
<td>17.2%</td>
<td>joint</td>
</tr>
<tr>
<td>mode</td>
<td>mono</td>
<td>2.8%</td>
<td>1.5%</td>
<td>3.7%</td>
<td>1.9%</td>
<td>1.4%</td>
<td>1.4%</td>
<td>3.2%</td>
<td>2.4%</td>
<td>0.9%</td>
<td>2.9%</td>
<td>2.6%</td>
<td>mode</td>
</tr>
<tr>
<td>organization</td>
<td>mono</td>
<td>2.3%</td>
<td>1.8%</td>
<td>2.2%</td>
<td>3.7%</td>
<td>2.3%</td>
<td>3.0%</td>
<td>2.6%</td>
<td>3.7%</td>
<td>1.5%</td>
<td>4.7%</td>
<td>4.7%</td>
<td>organization</td>
</tr>
<tr>
<td>purpose</td>
<td>mono</td>
<td>5.5%</td>
<td>3.2%</td>
<td>2.9%</td>
<td>3.9%</td>
<td>4.9%</td>
<td>3.9%</td>
<td>5.1%</td>
<td>4.4%</td>
<td>3.4%</td>
<td>7.4%</td>
<td>4.0%</td>
<td>purpose</td>
</tr>
<tr>
<td>restatement</td>
<td>both</td>
<td>3.0%</td>
<td>1.1%</td>
<td>1.7%</td>
<td>2.2%</td>
<td>1.7%</td>
<td>1.6%</td>
<td>2.9%</td>
<td>1.0%</td>
<td>2.9%</td>
<td>3.4%</td>
<td>3.0%</td>
<td>restatement</td>
</tr>
<tr>
<td>root</td>
<td>N/A</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>root</td>
</tr>
<tr>
<td>same-unit</td>
<td>multi</td>
<td>10.2%</td>
<td>9.1%</td>
<td>4.7%</td>
<td>4.6%</td>
<td>7.0%</td>
<td>5.2%</td>
<td>8.2%</td>
<td>9.1%</td>
<td>7.3%</td>
<td>5.2%</td>
<td>6.1%</td>
<td>same-unit</td>
</tr>
<tr>
<td>topic</td>
<td>mono</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>1.6%</td>
<td>0.1%</td>
<td>0.8%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>0.1%</td>
<td>0.6%</td>
<td>1.2%</td>
<td>topic</td>
</tr>
</tbody>
</table>

### Inter-paragraph relations

<table>
<thead>
<tr>
<th>relation</th>
<th>nuclearity</th>
<th>academic</th>
<th>bio</th>
<th>fiction</th>
<th>interview</th>
<th>news</th>
<th>reddit</th>
<th>speech</th>
<th>textbook</th>
<th>voyage</th>
<th>how</th>
<th>all</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>adversative</td>
<td>both</td>
<td>2.6%</td>
<td>1.1%</td>
<td>3.8%</td>
<td>2.1%</td>
<td>2.2%</td>
<td>7.4%</td>
<td>6.1%</td>
<td>3.5%</td>
<td>1.0%</td>
<td>2.4%</td>
<td>3.1%</td>
<td>adversative</td>
</tr>
<tr>
<td>attribution</td>
<td>mono</td>
<td>3.1%</td>
<td>0.8%</td>
<td>0.6%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>1.3%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.5%</td>
<td>attribution</td>
</tr>
<tr>
<td>causal</td>
<td>mono</td>
<td>1.0%</td>
<td>1.4%</td>
<td>7.8%</td>
<td>1.9%</td>
<td>1.9%</td>
<td>2.8%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>0.6%</td>
<td>2.2%</td>
<td>causal</td>
</tr>
<tr>
<td>context</td>
<td>mono</td>
<td>16.1%</td>
<td>7.4%</td>
<td>8.8%</td>
<td>15.6%</td>
<td>23.0%</td>
<td>6.4%</td>
<td>12.6%</td>
<td>14.1%</td>
<td>9.8%</td>
<td>4.4%</td>
<td>10.9%</td>
<td>context</td>
</tr>
<tr>
<td>contingency</td>
<td>mono</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>contingency</td>
</tr>
<tr>
<td>elaboration</td>
<td>mono</td>
<td>16.6%</td>
<td>9.0%</td>
<td>5.3%</td>
<td>5.0%</td>
<td>16.9%</td>
<td>7.4%</td>
<td>6.6%</td>
<td>5.9%</td>
<td>4.9%</td>
<td>7.2%</td>
<td>8.1%</td>
<td>elaboration</td>
</tr>
<tr>
<td>evaluation</td>
<td>mono</td>
<td>1.6%</td>
<td>0.5%</td>
<td>11.3%</td>
<td>2.4%</td>
<td>2.2%</td>
<td>14.0%</td>
<td>8.6%</td>
<td>3.5%</td>
<td>1.0%</td>
<td>0.8%</td>
<td>4.8%</td>
<td>evaluation</td>
</tr>
<tr>
<td>explanation</td>
<td>mono</td>
<td>2.1%</td>
<td>1.1%</td>
<td>1.3%</td>
<td>2.1%</td>
<td>0.8%</td>
<td>7.4%</td>
<td>8.1%</td>
<td>6.5%</td>
<td>0.5%</td>
<td>2.0%</td>
<td>2.7%</td>
<td>explanation</td>
</tr>
<tr>
<td>joint</td>
<td>multi</td>
<td>23.3%</td>
<td>53.8%</td>
<td>45.8%</td>
<td>31.0%</td>
<td>33.9%</td>
<td>34.2%</td>
<td>43.9%</td>
<td>25.3%</td>
<td>51.7%</td>
<td>60.4%</td>
<td>42.6%</td>
<td>joint</td>
</tr>
<tr>
<td>mode</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.6%</td>
<td>0.1%</td>
<td>mode</td>
</tr>
<tr>
<td>organization</td>
<td>mono</td>
<td>21.8%</td>
<td>19.4%</td>
<td>5.2%</td>
<td>12.2%</td>
<td>10.7%</td>
<td>6.6%</td>
<td>6.6%</td>
<td>30.0%</td>
<td>26.1%</td>
<td>17.1%</td>
<td>14.4%</td>
<td>organization</td>
</tr>
<tr>
<td>purpose</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>purpose</td>
</tr>
<tr>
<td>restatement</td>
<td>both</td>
<td>1.0%</td>
<td>0.0%</td>
<td>1.5%</td>
<td>1.9%</td>
<td>1.9%</td>
<td>2.3%</td>
<td>2.0%</td>
<td>0.6%</td>
<td>0.0%</td>
<td>0.6%</td>
<td>1.2%</td>
<td>restatement</td>
</tr>
<tr>
<td>root</td>
<td>N/A</td>
<td>9.3%</td>
<td>5.5%</td>
<td>3.6%</td>
<td>5.0%</td>
<td>6.3%</td>
<td>4.6%</td>
<td>5.1%</td>
<td>5.9%</td>
<td>4.7%</td>
<td>3.8%</td>
<td>5.5%</td>
<td>root</td>
</tr>
<tr>
<td>same-unit</td>
<td>multi</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>same-unit</td>
</tr>
<tr>
<td>topic</td>
<td>mono</td>
<td>1.0%</td>
<td>0.0%</td>
<td>5.0%</td>
<td>20.4%</td>
<td>0.3%</td>
<td>5.4%</td>
<td>0.0%</td>
<td>4.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>3.8%</td>
<td>topic</td>
</tr>
</tbody>
</table>

### Table 4.9: Percentages of intra-paragraph and inter-paragraph relation classes by genre in GUM.
4.5.3 Genre-wise Relation Distributions in GUM

More importantly, GUM provides deeper insights into genre-wise relation frequencies at intra-paragraph and inter-paragraph levels. The list below compares the horizontal rows for each relation in Table 4.9, and exceptional high and low percentages in the respective genres are marked in red and blue.

- The adversative class (including relations such as antithesis, concession and contrast) occurs most frequently both intra- and inter-paragraph in reddit due to the back-and-forth rebuttals among bloggers in the genre.

- Attribution dominates in intra-paragraph news since the source of information is prototypically mentioned for credibility concerns. In contrast, academic articles can cite or rephrase long texts and thus is the most frequent inter-paragraph genre. In intra- and inter-paragraph cases, voyage and whow rarely state the source of information. Moreover, many genres do not contain any inter-paragraph attribution, including news and textbook.

- The causal class (incl. cause and result) occurs quite evenly across genres in intra-paragraph and most frequently in inter-paragraph fiction stories. Fiction has one of the shortest paragraphs among genres and thus is more common in inter-paragraph. However, many genres do not have inter-paragraph causal relations, including speech and textbook.

- Context (incl. background and circumstance) is high in intra-paragraph fiction and inter-paragraph news. Both genres put forward background information and events to promote the main event in the document.

- Contingency (incl. condition) is high in intra-paragraph whow and low in bio because hypothetical events occur more frequently in instructional texts than
factual biographies. The relation is usually local and barely occurs between paragraphs.

- *Elaboration* is more frequently used in formal written genres within and across paragraphs, e.g., *academic, bio, news* and *textbooks*.

- Contrary to *elaboration*, *evaluation* is more prone to *fiction* and *reddit* than written genres especially when appearing across paragraphs.

- *Explanation* (incl. *evidence, justify* and *motivation*) occurs quite often in intra-paragraph *academic* when claims are supported by evidences. On the other hand, *news* is relatively factual and has the least *explanation* within and across paragraphs. Moreover, *reddit* and *speech* are the highest for inter-paragraph because of their long-distance dependencies of information flow.

- *Joint* dominates intra-paragraph *bio, fiction, and voyage* where multinuclear topics are often of equal importance to each other. Oppositely, *academic* has the least *joint* since argumentative evidence and claims support each other progressively. As for inter-paragraph, *whow* ranks first due to the parallel importance of (sub)sections and listings.

- *Mode* (incl. *means* and *manner*) describes how an action was executed or promoted and thus occurs frequently in intra-paragraph *fiction* and *speech* where storytelling is an important component of the genre. The relation is low in *voyage* where storylines are absent. *Mode* also barely occurs between paragraphs.

- *Organization* has similar intra-paragraph distributions across genres. As for inter-paragraph, it is the second most frequent relation overall, especially in *textbook*. 
• **Purpose** only appears within a paragraph with the highest in *whow* where the instruction steps aim to achieve a goal.

• **Restatement** is even within paragraphs. It is the most frequent in inter-paragraph *reddit* where bloggers frequently refer to each other’s posts.

• **Root** only occurs inter-paragraph.

• **Same-unit** only occurs intra-paragraph and is particularly frequent in written genres, e.g., *academic*, *bio*, and *textbook*.

• **Topic** (incl. *question* and *solution*) is overall not frequent and distributed evenly within paragraphs. The dialogue between the interviewer and the interviewee brings the amount of *topic* relations in *interview* remarkably high across paragraphs.

To summarize, many of these relation frequencies highly depend on the genres they are most usually tied to and whether they tend to occur in more local or distant environments. GUM’s highly diverse genre distribution benefits such in-depth genre-wise relation analysis. For example, in the news genre (the only genre shared between RST-DT and GUM), the two corpora share the same three most frequent relations: *elaboration*, *attribution*, and *joint*, even though the relation classes between the two corpora do not map directly to each. On the other end, *topic*, *contingency/condition* and *organization* are infrequent in the two datasets. However, whether one can directly equate these relation classes between RST-DT and GUM is debatable. One crucial problem is that the *topic-comment*, *topic-change*, *condition* classes include both multi-nuclear or nucleus-satellite relations in RST-DT. However, in GUM, the equivalent *topic*, *contingency* can only be mono-nuclear. Fortunately, this label mapping issue
does not occur between GUM and GCDT since GCDT employs the same set of relations as GUM, and comparisons can be conducted directly in Section 4.5.4.

4.5.4 Relation Distributions in GCDT

GCDT uses the same relation inventory and includes five written genres from GUM’s twelve genres. Table 4.10 presents the percentages of intra- and inter-paragraph classes by genre for the Chinese GCDT corpus. Many of the observations in GCDT are similar to GUM, even though they are RST datasets from different languages. Firstly, joint, organization, and topic are noticeably higher inter-paragraph, and the rest tend to be the same or higher intra-paragraph. Secondly, many relations are completely or almost absent in the inter-paragraph scenario, including contingency, mode, purpose, and same-unit. Lastly, the genre-wise highlights of GCDT echo the detailed observations enumerated for GUM. For example, attribution is high in intra-paragraph news, contingency in intra-paragraph whow, and topic in inter-paragraph interview in both GUM and GCDT. The homogeneous distributions of discourse relations between English GUM and Chinese GCDT consolidate the compatibility of the RST framework across languages and evince the benefits of promoting the same annotation guidelines (particularly the same relation inventories) across languages and datasets.

4.6 Interim Summary

This chapter examines the association between natural paragraphs and discourse structures in the three corpora: RST-DT, GUM, and GCDT. Sections 4.1 and 4.2 prepare sentence and paragraph-aligned RST data for analyses in the chapter. Section 4.3 presents basic structural statistics of the three corpora. Results show that RST-DT has a non-normal distribution of document length, whereas the structural
Table 4.10: Percentages of intra-paragraph and inter-paragraph relation classes by genre in GCDT.

### Intra-paragraph relations

<table>
<thead>
<tr>
<th>relation</th>
<th>nuclearity</th>
<th>academic</th>
<th>bio</th>
<th>interview</th>
<th>news</th>
<th>whow</th>
<th>all</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>adversative</td>
<td>both</td>
<td>4.1%</td>
<td>5.1%</td>
<td>7.7%</td>
<td>4.6%</td>
<td>6.3%</td>
<td>5.6%</td>
<td>adversative</td>
</tr>
<tr>
<td>attribution</td>
<td>mono</td>
<td>3.6%</td>
<td>5.0%</td>
<td>8.7%</td>
<td>14.7%</td>
<td>3.9%</td>
<td>6.7%</td>
<td>attribution</td>
</tr>
<tr>
<td>causal</td>
<td>mono</td>
<td>4.3%</td>
<td>4.7%</td>
<td>4.5%</td>
<td>3.7%</td>
<td>3.6%</td>
<td>4.2%</td>
<td>causal</td>
</tr>
<tr>
<td>context</td>
<td>mono</td>
<td>7.2%</td>
<td>7.6%</td>
<td>9.4%</td>
<td>5.9%</td>
<td>6.7%</td>
<td>7.3%</td>
<td>context</td>
</tr>
<tr>
<td>contingency</td>
<td>mono</td>
<td>1.4%</td>
<td>0.5%</td>
<td>2.5%</td>
<td>1.8%</td>
<td>6.8%</td>
<td>2.7%</td>
<td>contingency</td>
</tr>
<tr>
<td>elaboration</td>
<td>mono</td>
<td>18.7%</td>
<td>15.2%</td>
<td>18.9%</td>
<td>17.7%</td>
<td>17.9%</td>
<td>17.7%</td>
<td>elaboration</td>
</tr>
<tr>
<td>evaluation</td>
<td>mono</td>
<td>1.8%</td>
<td>1.5%</td>
<td>1.6%</td>
<td>1.2%</td>
<td>1.1%</td>
<td>1.4%</td>
<td>evaluation</td>
</tr>
<tr>
<td>explanation</td>
<td>mono</td>
<td>7.8%</td>
<td>12.0%</td>
<td>4.0%</td>
<td>3.5%</td>
<td>7.9%</td>
<td>7.3%</td>
<td>explanation</td>
</tr>
<tr>
<td>joint</td>
<td>multi</td>
<td>24.3%</td>
<td>26.8%</td>
<td>15.3%</td>
<td>24.0%</td>
<td>18.4%</td>
<td>21.8%</td>
<td>joint</td>
</tr>
<tr>
<td>mode</td>
<td>mono</td>
<td>3.1%</td>
<td>1.7%</td>
<td>2.2%</td>
<td>1.8%</td>
<td>3.1%</td>
<td>2.4%</td>
<td>mode</td>
</tr>
<tr>
<td>organization</td>
<td>mono</td>
<td>3.5%</td>
<td>1.3%</td>
<td>4.0%</td>
<td>2.2%</td>
<td>5.8%</td>
<td>3.4%</td>
<td>organization</td>
</tr>
<tr>
<td>purpose</td>
<td>mono</td>
<td>2.5%</td>
<td>2.0%</td>
<td>2.5%</td>
<td>2.2%</td>
<td>3.5%</td>
<td>2.6%</td>
<td>purpose</td>
</tr>
<tr>
<td>restatement</td>
<td>both</td>
<td>2.7%</td>
<td>2.8%</td>
<td>2.8%</td>
<td>1.6%</td>
<td>0.9%</td>
<td>2.2%</td>
<td>restatement</td>
</tr>
<tr>
<td>root</td>
<td>N/A</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>root</td>
</tr>
<tr>
<td>same-unit</td>
<td>multi</td>
<td>15.0%</td>
<td>13.7%</td>
<td>16.1%</td>
<td>15.0%</td>
<td>14.1%</td>
<td>14.7%</td>
<td>same-unit</td>
</tr>
<tr>
<td>topic</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>topic</td>
</tr>
</tbody>
</table>

**Number of all relations:** 1841 1758 1591 1423 1893 8506

### Inter-paragraph relations

<table>
<thead>
<tr>
<th>relation</th>
<th>nuclearity</th>
<th>academic</th>
<th>bio</th>
<th>interview</th>
<th>news</th>
<th>whow</th>
<th>all</th>
<th>relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>adversative</td>
<td>both</td>
<td>0.0%</td>
<td>0.8%</td>
<td>1.4%</td>
<td>1.7%</td>
<td>2.6%</td>
<td>1.4%</td>
<td>adversative</td>
</tr>
<tr>
<td>attribution</td>
<td>mono</td>
<td>0.0%</td>
<td>1.5%</td>
<td>0.5%</td>
<td>0.4%</td>
<td>0.0%</td>
<td>0.5%</td>
<td>attribution</td>
</tr>
<tr>
<td>causal</td>
<td>mono</td>
<td>0.5%</td>
<td>1.9%</td>
<td>0.0%</td>
<td>0.4%</td>
<td>0.3%</td>
<td>0.7%</td>
<td>causal</td>
</tr>
<tr>
<td>context</td>
<td>mono</td>
<td>1.6%</td>
<td>4.6%</td>
<td>14.6%</td>
<td>9.2%</td>
<td>1.6%</td>
<td>6.1%</td>
<td>context</td>
</tr>
<tr>
<td>contingency</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>contingency</td>
</tr>
<tr>
<td>elaboration</td>
<td>mono</td>
<td>3.6%</td>
<td>0.8%</td>
<td>1.8%</td>
<td>4.8%</td>
<td>5.2%</td>
<td>3.3%</td>
<td>elaboration</td>
</tr>
<tr>
<td>evaluation</td>
<td>mono</td>
<td>1.0%</td>
<td>1.5%</td>
<td>1.8%</td>
<td>1.3%</td>
<td>0.0%</td>
<td>1.1%</td>
<td>evaluation</td>
</tr>
<tr>
<td>explanation</td>
<td>mono</td>
<td>8.9%</td>
<td>3.8%</td>
<td>1.8%</td>
<td>4.8%</td>
<td>5.9%</td>
<td>5.0%</td>
<td>explanation</td>
</tr>
<tr>
<td>joint</td>
<td>multi</td>
<td>41.7%</td>
<td>59.2%</td>
<td>36.5%</td>
<td>55.0%</td>
<td>62.1%</td>
<td>52.2%</td>
<td>joint</td>
</tr>
<tr>
<td>mode</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>mode</td>
</tr>
<tr>
<td>organization</td>
<td>mono</td>
<td>37.5%</td>
<td>21.9%</td>
<td>11.9%</td>
<td>17.9%</td>
<td>18.6%</td>
<td>21.0%</td>
<td>organization</td>
</tr>
<tr>
<td>purpose</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>purpose</td>
</tr>
<tr>
<td>restatement</td>
<td>both</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>restatement</td>
</tr>
<tr>
<td>root</td>
<td>N/A</td>
<td>5.2%</td>
<td>3.8%</td>
<td>4.6%</td>
<td>4.4%</td>
<td>3.3%</td>
<td>4.1%</td>
<td>root</td>
</tr>
<tr>
<td>same-unit</td>
<td>multi</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>same-unit</td>
</tr>
<tr>
<td>topic</td>
<td>mono</td>
<td>0.0%</td>
<td>0.0%</td>
<td>25.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>4.6%</td>
<td>topic</td>
</tr>
</tbody>
</table>

**Number of all relations:** 192 260 219 229 306 1206

90
statistics of GUM differ largely across genres, particularly the average paragraph and EDU lengths. Section 4.4 examines how often paragraphs are properly contained in RST trees, particularly through double-annotated documents in RST-DT and genre-wise analyses in GUM and GCDT. Even though on average more than 90% of paragraphs are properly contained in an RST tree, some types of containment violations are frequent across corpora, for example, two outgoing edges from the same paragraph attaching to the same or different discourse units in another paragraph. Analyses on RST-DT signify the value of double annotations when associating paragraph containment rate with the inter-annotator agreement. On the other hand, the diversity of genres in GUM provides insight that such a paragraph containment analysis might not be as robust in conversational genres, such as fiction, reddit, and speech in GUM. Lastly, Section 4.5 discusses the different relations distributions within and across paragraphs among the three corpora and across diverse genres. Even though elaboration, joint are overall the most frequent relations, relation distributions vary largely across genres. For example, in all three corpora, attribution gets a much higher intra-paragraph distribution in news than in other genres of GUM or GCDT. The analyses illustrate the cross-genre particularities of the RST framework but also exemplify the homogeneity of RST, and adaptability to different languages and datasets. The results in this chapter encourage developing future discourse annotations and experiments using structurally informed approaches.
Chapte 5

Experiment: Multilingual and Multi-Genre RST Parsing

Discourse annotations from diverse genres and languages aim to enhance model performances on parsing and downstream NLU tasks. Thanks to the recent emergence of multilingual contextualized word representations such as the multilingual RoBERTa embedding (Conneau et al., 2020), researchers can gather more annotated resources from different languages to jointly train a better RST parser. As a result, multilingual RST parsers achieve SOTA performances (Braud et al., 2017; Iruskieta et al., 2019; Liu et al., 2021b). However, due to the previous absence of a sizeable Chinese RST dataset, there lacks a multilingual training experiment incorporating and testing on Chinese.

This chapter presents multilingual and multi-genre parsing experiments using the newly created Chinese GCDT corpus in Chapter 3. Section 5.1 demonstrates the setups, including the SOTA RST parser, datasets, language models, and evaluation metrics used in the experiments. Section 5.2 presents the benchmark parsing scores on monolingual Chinese GCDT and English GUM datasets. Since GCDT uses the same set of discourse relations and contains a subset of genres from GUM (Zeldes, 2017), Section 5.3 continues with multilingual experiments to jointly train with both GCDT and GUM datasets by applying different data manipulation techniques, such as pretraining and automatic translation. Section 5.4 evinces that these techniques can further boost parsing performance on top of multilingual training. Moreover, Sections 5.5 and 5.6 evaluate these monolingual and multilingual models regarding different
genres and intra-paragraph versus inter-paragraph discourse units. Results show a wide variance across genres and large performance deterioration for inter-paragraph relations and encourage more RST annotations of longer documents from various languages to enhance RST parsing further, particularly to tackle the bottleneck of parsing macro-level structures.

5.1 Experiment Setups

This section explains the basic setups for multiple sets of experiments throughout the chapter. First and foremost, monolingual training experiments are conducted using only one dataset, i.e., training and testing on the same dataset. With the release of the new Chinese GCDT corpus, this section lays out the training specifications, and Section 5.2 reports its benchmark parsing performance. Additionally, since GUM was recently updated to V8.0.0 and shares the same set of relation labels as GCDT, both sections investigate SOTA monolingual parsing performances on GUM V8.0.0.

5.1.1 DMRST Parser

Benchmark results in this chapter are produced using the SOTA multilingual DMRST parser (Liu et al., 2020, 2021b). DMRST is an end-to-end parser, including EDU segmentation and RST discourse parsing. The model’s decoder is based on a pointer network, and its stack is maintained by top-down depth-first span splitting. DMRST uses xlm-roberta-base and achieves SOTA performances in six languages by fusing multilingual resources. Experiments in this chapter use the same set of hyperparameters as reported in Liu et al. (2021b).

However, a few minor changes are made to their code (https://github.com/seq-to-mind/DMRST_Parser) to adjust to experiments conducted in this chapter.
The adapted DMRST code can be found in the following GitHub directory (https://github.com/logan-siyao-peng/GCDT/tree/main/adapted_DMRST). This supplementary code release provides a working sample for future customizations of the DMRST parser.

- Experiments in this chapter use GUM and GCDT’s 15 relation classes instead of RST-DT mappings;
- Experiments in this chapter assess multiple language models in addition to xlm-roberta-base;
- Experiments in this chapter choose the best epochs based on dev performance and continuously save the best-performing parameters, whereas Liu et al. (2021b)’s original code determines the best epoch directly based on performances on the test set;
- Experiments in this chapter split GCDT and GUM data by the predefined train/dev/test sets rather than random sampling to facilitate equal comparisons with future models;
- Experiments in this chapter enable featuring the model to pre-train on one corpus combination and then continue training on another.

5.1.2 Datasets

This study involves two multi-genre RST benchmarks: the newly released Chinese GCDT (Peng et al., 2022b) and the recently updated English GUM V8.0.0 (Zeldes, 2017) corpora. Previous research (Nishida and Matsumoto, 2022; Atwell et al., 2021) and Sections 4.4-4.5 of this dissertation exemplify that discourse structures vary in paragraph containment and intra- versus inter-paragraph relation distributions across different genres and languages.
The experiments investigate different samples of genre compositions listed below to isolate cross-lingual versus cross-genre influences. For example, GCDT and GUM-5 share the same five genres but differ in language. On the other hand, GUM-5 and GUM-12 are both English annotations, but the former subset includes fewer genres. Results in the following sections show performance degradation when fewer genres and languages are included in model training.

- **GCDT**: 50 Chinese documents from 5 genres: *academic, bio, interview, news,* and *whow*;
- **GUM-12**: 193 English documents from 12 genres: GCDT’s five genres plus *conversation, fiction, reddit, speech, textbook, vlog,* and *voyage*;
- **GUM-5**: a subset of 99 GUM documents from the same five genres in GCDT.

### 5.1.3 Language Models

The advancement of pre-trained language models in the past few years largely improve a wide range of NLP tasks. While the DMRST parser (Liu et al., 2021b) only uses the *xlm-roberta-base* model (Conneau et al., 2020), this dissertation assesses different monolingual and multilingual BERT (Devlin et al., 2019) and RoBERTa (Conneau et al., 2020) embeddings for Chinese, English, and multilingual, as shown in Table 5.1. Results in the following sections demonstrate that RoBERTa embeddings typically outperform BERT, and language-specific embeddings perform better than multilingual ones in applicable settings.

### 5.1.4 Metrics

RST parsing experiments are conventionally conducted and evaluated on the coarse-grained relation classes rather than relation labels. For example, previous models
Table 5.1: An overview of pretrained BERT and RoBERTa language models used in the experiments.

<table>
<thead>
<tr>
<th>Type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>Chinese: bert-base-chinese (Devlin et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>English: bert-base-cased (Devlin et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>Multilingual: bert-base-multilingual-cased (Devlin et al., 2019)</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>Chinese: hfl/chinese-roberta-wwm-ext (Cui et al., 2021)</td>
</tr>
<tr>
<td></td>
<td>English: roberta-base (Liu et al., 2019)</td>
</tr>
<tr>
<td></td>
<td>Multilingual: xlm-roberta-base (Conneau et al., 2020)</td>
</tr>
</tbody>
</table>

trained on the founding RST-DT corpus concern the 18 (including same-unit) coarse-grained relation classes rather than the 70+ fine-grained labels. Thus, evaluations in this chapter use the 15 coarse relation classes shared between GCDT and GUM (see Section 3.2.5 for the detailed set of relation labels and classes in GUM and GCDT).

This dissertation follows SOTA RST parsing papers to use the micro-averaged original Parseval F1 for Span, Nuclearity (Nuc), and Relation (Rel) scores. Morey et al. (2017) recommend the original (or standard) Parseval scores over Marcu (2000)’s RST-Parseval metrics because the latter artificially raises the agreement between RST trees. The three metrics, i.e., Span, Nuclearity, and Relation, emphasize different aspects of RST structural agreements. Span only compares the unlabeled skeleton tree structure between the gold and predicted RST trees. Nuclearity evaluates the tree structure with nuclearity labels, nucleus or satellite. The relation score assesses the structure with relation labels, for example, one of the 15 relation classes in these experiments. There is as well a Full score that requires correct nuclearity and relation labels. However, since the Full metrics can be decomposed into Nuclearity and Relation scores, it is not frequently reported in RST parsing papers.
To summarize, Section 5.1 explains essential experiment setups for this chapter, including the deployed multilingual parser, the assessed RST datasets, the experimented neural language models, and the evaluation metrics. The following sections will report models’ performances in monolingual and multilingual regimes and genre and paragraph-wise results.

5.2 Monolingual Results

Previous studies (Staliūnaitė and Iacobacci, 2020; Naseer et al., 2021; Tarunesh et al., 2021; Liu et al., 2021a) show that RoBERTa outperforms BERT on different natural language understanding tasks. Particularly in an adjacent task, i.e., implicit discourse relation classification for Penn Discourse Treebank (PDTB), Liu et al. (2021a)’s Bilateral Matching and Gated Fusion model with RoBERTa (BMGF-RoBERTa) outperforms Shi and Demberg (2019)’s BERT model trained on the Next Sentence Prediction (NSP) task.

Table 5.2 observes similar results for Chinese, English, and multilingual BERT and RoBERTa embeddings when training on GUM-5, GUM-12, and GCDT. All RoBERTa variations out-win their BERT counterparts. For example, Chinese-specific RoBERTa hfl/chinese-roberta-wwm-ext outperforms Chinese-specific BERT bert-base-chinese when training and testing on GCDT (Rel scores 51.76 > 50.81). Similarly, multilingual RoBERTa xlm-roberta-base outperforms multilingual BERT bert-base-multilingual-cased on GUM-12 (Rel scores 45.06 > 43.25).

Table 5.2 also demonstrates that language-specific neural embeddings consistently perform better than multilingual embeddings. For example, monolingual RoBERTa models outperform the multilingual ones: Rel score reaches 46.29 on GUM-12 with roberta-base but only 45.06 with xlm-roberta-base. Thus, monolin-
Table 5.2: Monolingual parsing results on the test sets of GCDT, GUM-5, and GUM-12 with Chinese, English, and multilingual BERT and RoBERTa embeddings (mean±std over five runs).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language Model</th>
<th>Span (±std)</th>
<th>Nuc (±std)</th>
<th>Rel (±std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCDT</td>
<td>bert-base-chinese</td>
<td>73.15±0.53</td>
<td>55.71±0.66</td>
<td>50.81±0.65</td>
</tr>
<tr>
<td></td>
<td>bert-base-multilingual-cased</td>
<td>67.34±1.32</td>
<td>47.66±0.73</td>
<td>43.97±0.93</td>
</tr>
<tr>
<td></td>
<td>hfl/chinese-roberta-wwm-ext</td>
<td>75.51±0.68</td>
<td>57.08±0.81</td>
<td>51.76±0.97</td>
</tr>
<tr>
<td></td>
<td>xlm-roberta-base</td>
<td>74.35±0.54</td>
<td>54.17±1.20</td>
<td>50.45±1.09</td>
</tr>
<tr>
<td>GUM-5</td>
<td>bert-base-cased</td>
<td>64.61±1.42</td>
<td>49.58±1.51</td>
<td>40.43±1.56</td>
</tr>
<tr>
<td></td>
<td>bert-base-multilingual-cased</td>
<td>64.52±2.68</td>
<td>51.63±2.07</td>
<td>44.96±1.46</td>
</tr>
<tr>
<td></td>
<td>roberta-base</td>
<td>73.85±0.70</td>
<td>58.95±0.79</td>
<td>50.35±1.18</td>
</tr>
<tr>
<td></td>
<td>xlm-roberta-base</td>
<td>72.45±0.97</td>
<td>56.78±0.80</td>
<td>47.69±0.88</td>
</tr>
<tr>
<td>GUM-12</td>
<td>bert-base-cased</td>
<td>60.93±0.63</td>
<td>47.92±0.62</td>
<td>40.20±0.40</td>
</tr>
<tr>
<td></td>
<td>bert-base-multilingual-cased</td>
<td>64.47±0.50</td>
<td>50.69±0.32</td>
<td>43.25±0.35</td>
</tr>
<tr>
<td></td>
<td>roberta-base</td>
<td>68.59±0.58</td>
<td>55.32±0.27</td>
<td>46.29±0.46</td>
</tr>
<tr>
<td></td>
<td>xlm-roberta-base</td>
<td>66.12±0.59</td>
<td>52.58±0.52</td>
<td>45.06±0.45</td>
</tr>
</tbody>
</table>

gual RoBERTa embeddings perform best when training with monolingual data. For example, hfl/chinese-roberta-wwm-ext (Cui et al., 2021) obtains Span, Nuclearity, and Relation scores of 75.51, 57.08, and 51.76, the highest monolingual results on GCDT. Lastly, the overall result of GCDT is slightly higher than GUM-5 but much better than GUM-12. The hypothesis is that genre diversity creates difficulties. The following sections present more parsing experiments on multilingual training and assess genre and structure diversity in RST parsing.

5.3 Multilingual Experiment Setups

The multilingual experiments henceforth follow the basic experiment setups explained in Section 5.1, including using the multilingual DMRST parser and the original Parseval metrics. As for language models, Section 5.2 evinces that RoBERTa models consistently outperform BERT models on all three corpora. Thus, the following studies
only report multilingual training results with language-dependent and multilingual RoBERTa embeddings. Unlike Liu et al. (2021b)’s experiments with six languages, experiments in this study can efficiently train on GCDT and GUM without label mapping. Sections 5.3.1 and 5.3.2 discuss two further implementation details specific to multilingual training: data combination strategies and data augmentation techniques.

5.3.1 Multilingual Data Combinations

In addition to monolingual training on GCDT, GUM-5, and GUM-12, multilingual experiments in the following sections train on combinations of these three corpora. Table 5.3 presents the train/dev/test splits when jointly training with GCDT and GUM in multilingual experiments. To summarize, when jointly training on both datasets and testing on one, the model includes the pre-defined training partitions from both datasets and the dev partition from the other dataset for training. Only the designated dev and test partitions of the testing dataset are included for development and testing. The test partition of the other dataset is never presented to the model to prevent over-fitting.

Table 5.3: An overview of the train/dev/test splits of GCDT and GUM used for training in the multilingual experiments.

<table>
<thead>
<tr>
<th></th>
<th>train: GCDT+GUM</th>
<th>train: GCDT+GUM</th>
<th>dev/test: GUM</th>
<th>dev/test: GCDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>GUM-train</td>
<td>GCDT-train</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ GCDT-train</td>
<td>+ GUM-train</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ GCDT-dev</td>
<td>+ GUM-dev</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev</td>
<td>GUM-dev</td>
<td>GCDT-dev</td>
<td></td>
<td></td>
</tr>
<tr>
<td>test</td>
<td>GUM-test</td>
<td>GCDT-test</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3.2 Pretraining and Automatic EDU-wise Translation

The essence of multilingual training is to benefit from the larger amount and more diverse annotations from multiple languages. Though multilingual embeddings can directly encode RST annotations from various languages, previous studies reveal that data manipulation techniques, such as pretraining and automatic translation, can further boost performance (Liu et al., 2021b; Kobayashi et al., 2021). This section experiments with these two techniques while training with combinations of the original GCDT and GUM datasets using multilingual embeddings. For example, to improve parsing performance on the Chinese GCDT corpus, the following pretraining and automatic EDU-wise translation procedures are involved:

1) **Pretraining**: first train models with both English and Chinese data and then continue training only on the training partition of the target dataset (i.e., GCDT);

2) **Automatic EDU-wise Translation**: GoogleTranslator\(^1\) is used to automatically translate EDUs from the other dataset to the target language (i.e., EDU-wise en→zh translations of GUM), and then train on the original GCDT and translated GUM data. The advantage of the translation approach is that higher-performing monolingual embeddings can replace lower-performing multilingual embeddings.

Figure 5.1 shows an example of a Chinese RST subtree automatically translated into English, EDU by EDU. Admittedly, automatic translation is error-prone. For example, Figure 5.1b does not recognize relative clauses as propositional clauses in EDU_10 and EDU_13. However, previous studies (Cheng and Li, 2019; Liu et al., 2021b) and the results in Section 5.4 show that combining automatically translated RST trees from other languages nevertheless enhance parsing performance. Besides presenting the overall parsing performance, Sections 5.5 and Section 5.6 fur-

\(^1\)https://github.com/nidhaloff/deep-translator
ther examine the parsers’ and humans’ performances from two perspectives: different genres and intra-paragraph versus inter-paragraph distinctions. Analyses show that even though the SOTA parsing scores look satisfactory overall, performances on some genres and macro-level discourse units await improvements.

Figure 5.1: The original and automatic zh→en EDU-wise translated sub-trees of `gcdt_academic_dingzhen`.

(a) The original annotation.

(b) The zh→en translation.
5.4 Multilingual Results

This section experiments with four sets of data combinations and four training strategies for each combination. The data combinations include:

- ComboA: training on GCDT+GUM-5 and developing/testing on GCDT;
- ComboB: training on GCDT+GUM-12 and developing/testing on GCDT;
- ComboC: training on GCDT+GUM-5 and developing/testing on GUM-5;
- ComboD: training on GCDT+GUM-12 and developing/testing on GUM-12.

For each data combination, the following training strategies are involved:

- StrategyA: joint training of both original datasets with the multilingual RoBERTa embedding;
- StrategyB: first pre-training on both original datasets and then training on the target dataset with the multilingual RoBERTa embedding;
- StrategyC: translating the other dataset into the target language and joint training the original target dataset and the translated other dataset using the multilingual RoBERTa embedding;
- StrategyD: translating the other dataset into the target language and joint training the original target dataset and the translated other dataset using the better-performing target language RoBERTa embedding.

Table 5.4 shows the multilingual parsing performances of these 16 experiments (4 data combinations \( \times \) 4 training strategies). The results can be interpreted from the following aspects.

Firstly, multilingual joint training with GCDT+GUM (StrategyA) outperforms monolingual results in all three test scenarios: GCDT, GUM-5, and GUM-12. For
Table 5.4: Multilingual parsing results with pretraining and automatic translation on the test sets of GCDT+GUM combinations with highest-performing Chinese, English, and multilingual RoBERTa embeddings (mean±std over five runs).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Span</th>
<th>Nuc</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monolingual baselines (see Table 5.2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train/Dev/Test on GCDT w/ xlm-roberta-base</td>
<td>74.35±0.54</td>
<td>54.17±1.20</td>
<td>50.45±1.09</td>
</tr>
<tr>
<td>Train/Dev/Test on GCDT w/ hfl/chinese-roberta-wwm-ext</td>
<td>75.51±0.68</td>
<td>57.08±0.81</td>
<td>51.76±0.97</td>
</tr>
<tr>
<td>Train/Dev/Test on GUM-5 w/ xlm-roberta-base</td>
<td>72.45±0.97</td>
<td>56.78±0.80</td>
<td>47.69±0.88</td>
</tr>
<tr>
<td>Train/Dev/Test on GUM-5 w/ roberta-base</td>
<td>73.85±0.70</td>
<td>58.95±0.79</td>
<td>50.35±1.18</td>
</tr>
<tr>
<td>Train/Dev/Test on GUM-12 w/ xlm-roberta-base</td>
<td>66.12±0.59</td>
<td>52.58±0.52</td>
<td>45.06±0.45</td>
</tr>
<tr>
<td>Train/Dev/Test on GUM-12 w/ roberta-base</td>
<td>68.59±0.58</td>
<td>55.32±0.27</td>
<td>46.29±0.46</td>
</tr>
<tr>
<td><strong>ComboA: Train on GCDT+GUM-5 and Dev/Test on GCDT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrategyA: joint training w/ xlm-roberta-base</td>
<td>74.24±0.48</td>
<td>56.68±0.86</td>
<td>52.21±0.83</td>
</tr>
<tr>
<td>StrategyB: +pretraining w/ xlm-roberta-base</td>
<td>76.97±0.32</td>
<td>57.94±0.82</td>
<td>53.38±0.51</td>
</tr>
<tr>
<td>StrategyC: +en→zh trans. w/ xlm-roberta-base</td>
<td>74.80±0.78</td>
<td>56.58±0.98</td>
<td>51.18±1.15</td>
</tr>
<tr>
<td><strong>StrategyD: +en→zh trans. w/ hfl/chinese-roberta-wwm-ext</strong></td>
<td>77.66±0.42</td>
<td>59.29±0.59</td>
<td>54.66±0.76</td>
</tr>
<tr>
<td><strong>ComboB: Train on GCDT+GUM-12 and Dev/Test on GCDT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrategyA: joint training w/ xlm-roberta-base</td>
<td>74.33±0.49</td>
<td>57.24±0.99</td>
<td>52.61±1.13</td>
</tr>
<tr>
<td>StrategyB: +pretraining w/ xlm-roberta-base</td>
<td>76.95±0.65</td>
<td>59.40±0.64</td>
<td>55.28±0.23</td>
</tr>
<tr>
<td>StrategyC: +en→zh trans. w/ xlm-roberta-base</td>
<td>73.99±0.79</td>
<td>56.31±1.43</td>
<td>51.51±1.34</td>
</tr>
<tr>
<td>StrategyD: +en→zh trans. w/ hfl/chinese-roberta-wwm-ext</td>
<td>78.11±0.39</td>
<td>59.42±0.90</td>
<td>54.41±1.23</td>
</tr>
<tr>
<td><strong>ComboC: Train on GUM-5+GCDT and Dev/Test on GUM-5</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrategyA: joint training w/ xlm-roberta-base</td>
<td>72.56±0.71</td>
<td>60.63±0.43</td>
<td>52.57±0.77</td>
</tr>
<tr>
<td>StrategyB: +pretraining w/ xlm-roberta-base</td>
<td>73.44±0.36</td>
<td>59.40±0.56</td>
<td>50.57±0.97</td>
</tr>
<tr>
<td>StrategyC: +zh→en trans. w/ xlm-roberta-base</td>
<td>72.21±1.11</td>
<td>60.67±1.25</td>
<td>52.32±1.05</td>
</tr>
<tr>
<td><strong>StrategyD: +zh→en trans. w/ roberta-base</strong></td>
<td>74.73±0.40</td>
<td>62.65±0.72</td>
<td>54.32±0.82</td>
</tr>
<tr>
<td><strong>ComboD: Train on GUM-12+GCDT and Dev/Test on GUM-12</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrategyA: joint training w/ xlm-roberta-base</td>
<td>70.32±0.37</td>
<td>57.49±0.73</td>
<td>49.14±0.34</td>
</tr>
<tr>
<td>StrategyB: +pretraining w/ xlm-roberta-base</td>
<td>66.00±0.24</td>
<td>53.13±0.22</td>
<td>45.47±0.42</td>
</tr>
<tr>
<td>StrategyC: +zh→en trans. w/ xlm-roberta-base</td>
<td>70.28±0.55</td>
<td>57.63±0.55</td>
<td>49.26±0.39</td>
</tr>
<tr>
<td><strong>StrategyD: +zh→en trans. w/ roberta-base</strong></td>
<td>71.41±0.47</td>
<td>59.17±0.35</td>
<td>50.63±0.48</td>
</tr>
</tbody>
</table>
example, training on GCDT+GUM-12 using xlm-roberta-base achieves Span, Nuclearity, and Relation F1s of 74.33, 57.24, and 52.61 on GCDT. Even though the F-score on Span is lower, training with combined data performs better on Relation than training with only GCDT data and the same embedding.

Secondly, training with more genres from GUM (ComboB: GCDT+GUM-12) achieves slightly better performance than training only using the same genres (ComboA: GCDT+GUM-5) when tested on GCDT.

Thirdly, pretraining on the GCDT+GUM-combined training sets and continuing to train on the target corpus training set (StrategyB) improves performance over Strategy A on Chinese GCDT but deteriorates on English GUM. One possible explanation is that there is less room for improvement with more English training data. On the contrary, pretraining for the smaller Chinese dataset added to the comparatively little information available to the parser.

Lastly, results show that augmenting with automatic translation and using monolingual embeddings (StrategyD) achieve the best performance on three data combinations, even though StrategyB achieves the best result on ComboB. By automatically translating EDUs to the target language, one enables replacing multilingual embeddings with better-performing monolingual embeddings. Table 5.4 evinces that StrategyD wins over StrategyC in applicable settings, mirroring the findings in Section 5.2.

In sum, Sections 5.1-5.4 report SOTA mono- and multilingual parsing performances on GCDT and GUM V8.0.0 using the DMRST parser (Liu et al., 2021b). Results show that monolingual RoBERTa embeddings, data combinations, pretraining, and automatic EDU-wise translations improve GCDT parsing. The following Sections 5.5 and 5.6 delve into analyzing performance from two aspects:
across GCDT’s multiple genres and between intra- versus inter-paragraph discourse units.

5.5 Genre-wise Performances

Domain adaption is difficult in discourse parsing (Ji et al., 2015; Atwell et al., 2021; Nishida and Matsumoto, 2022) due to the status quo that most of the benchmark RST parsers are trained only on English Wall Street Journal news in the RST-DT corpus. Since GCDT exhibits five different genres for Chinese, this section reports the per-genre performances on GCDT. Table 5.5 selects three models trained in the monolingual GCDT and translation-augmented scenarios (i.e., GCDT+GUM-5 and GCDT+GUM-12) using the best-performing Chinese RoBERTa embedding (Cui et al., 2021).

Table 5.5: GCDT genre-wise performances on sample models trained on GCDT, translation-augmented GCDT+GUM-5, and GCDT+GUM-12 combinations using hfl/chinese-roberta-wwm-ext.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Trained on GCDT</th>
<th>Trained w/ trans. on GCDT+GUM-5</th>
<th>Trained w/ trans. on GCDT+GUM-12</th>
<th>Human Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Span  Nuc Rel</td>
<td>Span  Nuc Rel</td>
<td>Span  Nuc Rel</td>
<td>Span  Nuc Rel</td>
</tr>
<tr>
<td>academic</td>
<td>74.64  54.07  48.33</td>
<td>72.25  47.37  43.54</td>
<td>75.12  51.20  44.98</td>
<td>80.38  59.33  49.76</td>
</tr>
<tr>
<td>bio</td>
<td>72.87  54.26  52.71</td>
<td>74.81  57.75  53.49</td>
<td>77.52  59.69  55.43</td>
<td>81.57  63.92  55.69</td>
</tr>
<tr>
<td>interview</td>
<td>74.68  56.33  52.53</td>
<td>80.38  61.39  55.70</td>
<td>77.85  56.96  48.73</td>
<td>83.55  62.50  54.61</td>
</tr>
<tr>
<td>news</td>
<td>76.63  56.52  50.54</td>
<td>83.15  64.13  57.07</td>
<td>78.80  60.33  54.35</td>
<td>80.98  61.96  54.35</td>
</tr>
<tr>
<td>whow</td>
<td>77.89  57.76  54.79</td>
<td>80.20  66.34  62.71</td>
<td>80.20  65.68  61.06</td>
<td>91.99  77.70  69.34</td>
</tr>
<tr>
<td>Overall</td>
<td>75.45  55.85  52.07</td>
<td>77.97  59.71  55.04</td>
<td>78.06  59.44  53.87</td>
<td>84.27  66.15  57.77</td>
</tr>
</tbody>
</table>

Table 5.5 provides per-genre parsing results of the models on the five test genres. On the one hand, the average performance on how-to guides (whow) is much higher than academic articles for both models and humans. This demonstrates a satisfying human-model alignment regarding which genre is the hardest or easiest (see Zeldes and Simonson 2016 for analyses about GUM). Conversely, model results are the farthest from the human ceiling scores on the highest performing whow genre. It is evident
from the table that characteristics of genres trigger the different performances. Future multi-genre experiments could be conducted across datasets to study out-of-domain effects in multilingual RST parsing scenarios.

Nevertheless, the current annotation agreements and parsing results on GCDT genres have limitations. GCDT only includes five test documents, one from each genre. Adding more test documents for each genre would be required to strengthen observations regarding genre-wise performances. Going beyond the dissertation, GCDT will include more double annotations for each genre. At the same time, GUM also sets the stage for double annotations and annotations for out-of-domain genres.

5.6 Intra- versus Inter-Paragraph Performances

Paragraph breaks signify thematic discontinuities (Ji, 2008), and inter-paragraph discourse relations are shown to be different from intra-paragraph ones (Sporlede and Lascarides, 2004; Wang et al., 2017; Feng and Hirst, 2014). For example, elaboration, topic-comment, and textual-organization are more common across paragraphs than within a paragraph in RST-DT. Previous RST parsers also use sentence and paragraph boundaries as features or in a multi-stage pipeline (Liu and Lapata, 2017; Wang et al., 2017, 2019; Kobayashi et al., 2020). This section examines the DMRST parser’s performance on intra-paragraph versus inter-paragraph discourse units compared to human agreements. Since GUM does not include double annotation, this section presents results on Chinese GCDT and English RST-DT. GCDT double-annotates the five test documents, which enables a fair comparison between model and human performances. RST-DT is the largest double-annotated RST corpus with 53 double-annotated documents, 48 occur in the train partition and 5 in test.
5.6.1 Paragraph Performances on GCDT

To separate intra-paragraph versus inter-paragraph performances, paragraph boundaries are aligned to discourse units in an RST tree. The GCDT test set has 918 intra-paragraph and 169 inter-paragraph instances of discourse units. Table 5.6 presents the overall, intra-paragraph and inter-paragraph performances for three models trained on GCDT and translation-augmented GCDT+GUM combinations, as well as on human agreements.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Span</td>
<td>Nuc</td>
<td>Rel</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trained on GCDT</td>
<td>74.52</td>
<td>56.76</td>
<td>51.61</td>
</tr>
<tr>
<td>Intra-</td>
<td>66.15</td>
<td>50.41</td>
<td>45.63</td>
</tr>
<tr>
<td>Inter-</td>
<td>8.37</td>
<td>6.35</td>
<td>5.98</td>
</tr>
<tr>
<td>Trained on GCDT+GUM-5 w/ en→zh trans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>76.91</td>
<td>61.27</td>
<td>54.65</td>
</tr>
<tr>
<td>Intra-</td>
<td>67.71</td>
<td>53.73</td>
<td>47.56</td>
</tr>
<tr>
<td>Inter-</td>
<td>9.20</td>
<td>7.54</td>
<td>7.08</td>
</tr>
<tr>
<td>Trained on GCDT+GUM-12 w/ en→zh trans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>77.28</td>
<td>60.35</td>
<td>54.19</td>
</tr>
<tr>
<td>Intra-</td>
<td>67.53</td>
<td>52.25</td>
<td>46.46</td>
</tr>
<tr>
<td>Inter-</td>
<td>9.75</td>
<td>8.10</td>
<td>7.73</td>
</tr>
<tr>
<td>Human Performance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>84.27</td>
<td>66.15</td>
<td>57.77</td>
</tr>
<tr>
<td>Intra-</td>
<td>71.00</td>
<td>53.73</td>
<td>45.72</td>
</tr>
<tr>
<td>Inter-</td>
<td>14.17</td>
<td>12.42</td>
<td>12.05</td>
</tr>
</tbody>
</table>

Monolingual and multilingual trained models have similar performances on intra-versus inter-paragraph scenarios. In general, intra-paragraph recalls are the highest, with R_Rel around 55 for models; and inter-paragraph precisions are the lowest, with P_Rel only around 7. On the one side, the models’ performances are almost as good as humans in the intra-paragraph scenarios, reaching an F1_Rel of around
50. However, the performance gap on inter-paragraph units is substantial in precision and recall. Even though human agreement also needs improvements by adjudication and refining the annotation guidelines, the models’ performances are still 7-8 points below the human ceiling on inter-paragraph units.

5.6.2 Paragraph Performances on RST-DT

Since GCDT only includes five double-annotated documents, one from each of the five genres, it could be possible that the results above are due to coincidences. To strengthen the observation, this section looks into a similar intra- versus inter-paragraph distinction on the English RST-DT benchmark. RST-DT includes 53 double-annotated documents in its release, with 48 from the train partition and 5 from test. To maximize structural analysis support and avoid evaluating over-fitted training data, this section alters the test set to the 53 double-annotated documents while using the other documents for training and development. The DMRST parser is trained on RST-DT, and parsing performances are analyzed on all 53 documents against human agreements. Moreover, because RST-DT double annotations are conducted from raw texts and could result in different EDU segmentation, this section reports the DMRST parser’s performance on predicted instead of gold EDUs for RST-DT. There are 2,291 intra-paragraph and 698 inter-paragraph discourse units in the 53 documents.

Table 5.7 shows DMRST parser and human performances on the 53 double annotated documents. Overall, the parser’s performance on RST-DT is similar to GCDT. Firstly, intra-paragraph precision metrics are lower than overall precision, whereas intra-paragraph recall metrics are higher. However, since the rise in RST-DT’s intra-paragraph recall is much higher than GCDT, RST-DT has a higher intra-paragraph F1 score than overall F1. Secondly, like parsers’ results on GCDT, inter-paragraph
Table 5.7: Intra- versus inter-paragraph precision, recall and F1 on Span, Nuclearity, and Relation for DMRST model trained on RST-DT with \textit{roberta-base} and human performances when the test set is modified to be the 53 double-annotated documents.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seg</td>
<td>Span</td>
<td>Nuc</td>
</tr>
<tr>
<td>Trained on RST-DT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>99.51</td>
<td>76.06</td>
<td>64.97</td>
</tr>
<tr>
<td>Intra-</td>
<td>99.51</td>
<td>66.20</td>
<td>57.91</td>
</tr>
<tr>
<td>Inter-</td>
<td>99.51</td>
<td>9.87</td>
<td>7.07</td>
</tr>
<tr>
<td>Human overall performance reported in Morey et al. (2017)</td>
<td>/</td>
<td>78.7</td>
<td>66.8</td>
</tr>
</tbody>
</table>

precision metrics are the lowest-performing scenario for RST-DT parsers. More interestingly, inter-paragraph recall is also remarkably low for the RST-DT parser. This result is hypothesized to be relevant to the non-normal distribution of document lengths in RST-DT as discussed in Section 4.3.2, which complicates the models in detecting paragraph boundaries and labeling inter-paragraph relations. Even though Table 5.7 does not reproduce the reported inter-annotator agreement in Morey et al. (2017) due to the difficulties in allowing different EDU segmentation decisions between annotators, the results confirm the bottleneck with parsing inter-paragraph relations in the RST-DT dataset. Nevertheless, GCDT’s annotation procedure to first double-annotate EDUs and then annotate relations based on adjudicated EDU boundaries is more effective in separating annotators’ disagreements in EDU segmentation from relation annotation.

In sum, this section observes consistent results across GCDT and RST-DT that inter-paragraph performances are far below expectations in precision and recall. However, due to RST-DT’s tolerance for different EDU segmentation versions between annotators, the current paragraph-level performance comparison for RST-DT is most
likely biased. The lack of comparable double annotations could be resolved by double annotating more documents for GCDT and GUM with adjudicated EDU segments and extending the intra-versus-inter analyses beyond the paragraph level.

5.7 Interim Summary

This chapter establishes several SOTA parsing benchmarks on the newly released GCDT dataset. In addition to presenting results in monolingual settings, Sections 5.1-5.4 jointly train on GCDT and the similar English GUM corpus and demonstrate that multilingual training and automatic EDU translation boost parser performance. The results show that joint training and translation improve performance on the smaller Chinese and larger English datasets using monolingual and multilingual embeddings. However, pretraining only helps with the smaller Chinese data. Moreover, it is important to notice that monolingual RoBERTa embeddings outperform multilingual embeddings in applicable settings. Sections 5.5-5.6 further conduct genre and paragraph analyses and show that parsing performance varies widely between genres and deteriorates largely for inter-paragraph relations. Still, the best overall performance is achieved using Chinese and English data in a multilingual training regime. Beyond this dissertation, GCDT and GUM plan to conduct more double annotations and extend the current intra-versus-inter analyses to higher discourse structures, such as sections and topic segments.
Chapter 6

Discussion and Future Work

Understanding the discourse structure of documents is one of the most challenging tasks in computational linguistics (Kong et al., 2019; Kang et al., 2019). In the past few years, there have been joint efforts from both linguistics and natural language processing to improve discourse understanding tasks such as Rhetorical Structure Theory (RST) parsing. Many RST annotations in diverse languages and genres have become available to the public (Carlson et al., 2001; Zeldes, 2017; Stede and Neumann, 2014; Cao et al., 2018). SOTA RST parsers incorporate multilingual and contextualized neural language models and start to outperform strong baselines on RST-DT, an English news-only RST benchmark (Liu et al., 2021b). However, RST parsing is still far from satisfactory. Not only is RST parsing inferior in non-news genres and at higher discourse levels, but a sizeable Chinese RST corpus and subsequent SOTA RST parsers that incorporate Chinese are also missing.

This dissertation is titled “Cross-Paragraph Discourse Structure in Rhetorical Structure Theory Parsing and Treebanking for Chinese and English” and comprises six chapters. Chapter 1 introduces the status quo of RST datasets and parsers, highlighting this dissertation’s contributions. Chapter 2 presents previous research from a few areas: English RST and Chinese discourse datasets, genres and paragraphs in RST, and structure-informed and multilingual RST parsing. Chapters 3-5 organize contributions of this dissertation into three aspects: creating the largest Chinese RST
dataset, analyzing paragraph structures in RST, and experimenting with multilingual RST parsing. Chapter 3 introduces the largest Chinese RST corpus, i.e., GCDT, which closely follows established RST guidelines and is highly comparable to existing English RST corpora. Chapter 4 discusses the role of paragraph structures in English and Chinese RST corpora from two perspectives: paragraph containment and inter-versus intra-paragraph relation distributions. Chapter 5 conducts multilingual parsing experiments on Chinese and English corpora with pretraining and automatic EDU-wise translation and achieves SOTA performances on monolingual and multilingual setups. The dissertation also presents parsers’ performances and human agreements on genre diversity and paragraph structure. It shows that the SOTA parsing performances are unsatisfactory in some genres and the inter-paragraph scenario. As the author of this dissertation, I hope the GCDT dataset can alleviate the lack of training resources for hierarchical discourse parsing in Chinese. Results in this dissertation also encourage future work to improve RST parsing on diverse genres and macro-level structures.

There are a few promising future research directions in improving macro-level RST parsing. Since RST parsing aims to simulate gold human annotations, one can further analyze the role of macro-level boundaries in human annotations. A sample research question would be: how much would human annotations differ when given the macro-level boundary information versus when such information is hidden from them? This would be an annotation experiment where annotators are asked first to annotate the texts without boundary information and then revise for a second version with given macro-level boundaries. I hypothesize that analyses on annotation revisions can provide more linguistic insights into structure-informed RST parsing, particularly on macro-level performances.
In addition to the paragraph structures discussed in this dissertation, another even higher level of structure, i.e., topic segments, also deserves exploration. Topic segmentation, also known as text tiling, dates back to Hearst (1997), where a document is divided into multi-paragraph segments where each segment discusses a specific topic, and segment boundaries mark the transition between topics. Most topic segmentation datasets are sourced from Wikipedia (Arnold et al., 2019; Koshorek et al., 2018; Choi, 2000), and recent topic segmentation models (Li et al., 2018, 2020; Glava and Somasundaran, 2020; Aumiller et al., 2020; Arnold et al., 2019; Lukasik et al., 2020) either use unsupervised approaches or treat section headers as gold topic boundaries. Despite a couple of new topic segmentation annotations on dialogues (Xia et al., 2022; Kirihara et al., 2022), there lacks a general guideline on how topic structures should be annotated across genres and languages.

Recent research has started to connect topic segmentation with RST. Li et al. (2018, 2020); Lukasik et al. (2020) create neural models for both topic and EDU segmentation tasks. Xing et al. (2022) use paragraph boundaries in RST-DT as gold topic boundaries for topic segmentation experiments. Nevertheless, there is no overlapping corpus with both topic and RST annotations. When preparing for this dissertation, I collaborated with an undergraduate student and conducted pilot topic annotations on the GUM corpus. However, due to the variations in genres and the lack of previous annotations or guidelines for reference, the pilot annotations were based on untenable assumptions such as binary topic splits and deep hierarchical structure. A few revisions are proposed for a second round of pilot annotation. Though beyond the scope of this dissertation, I plan to include topic segmentation into GUM and GCDT and to re-evaluate RST trees following the methodologies in Chapter 4 and Section 5.6 but at the higher topic level.
On top of providing linguistic insights towards improving RST parsers, another research direction would be examining the current evaluation process. Morey et al. (2017) argue for the original Parseval metrics rather than RSTParseval because the latter results in artificially high agreements. However, a natural RST tree always implies an imbalance between macro-level and micro-level relations, for example, a 5:1 ratio on average between intra-paragraph and inter-paragraph instances as shown in Section 5.6. A hypothetical question is whether macro-level discourse structures should be weighed unequally from micro-level relations. This imbalance in evaluation metrics is also relevant to annotation ambiguity in RST. In other words, if human annotators tend to disagree with each other in specific scenarios, should there be fewer penalties on the model? Given that multiple manually annotated RST trees can be simultaneously compatible (Mann and Thompson, 1988; Taboada and Mann, 2006b), some weighting variations will also represent human evaluation better. Furthermore, structural weights can be correlated with performances on downstream applications, for example, whether some portions of the RST structure are more crucial for downstream applications than others. This might lead to another way of decomposing RST trees besides the multi-stage approach.

Lastly, the limited training data is one of the most crucial hindering factors in RST parsing. Though combining multilingual training sources can somewhat alleviate the problem (see Section 2.6 and Chapter 5), the complex annotation schema necessitates a large amount of human labor to achieve an adequate corpus size. For example, annotating 50 documents (62K tokens) for GCDT took almost six months for one expert annotator, including document selection, tokenization, EDU segmentation, and relation annotation. The expense of human labor motivates the growth of silver-quality datasets, such as AMALGUM (Gessler et al., 2020) and MEGA-DT (Huber and Carenini, 2020a) that are created by distant supervision and can reach a more
voluminous size. Even though these datasets do not contain fully gold annotations, they are helpful pretraining resources and have shown performance benefits in parsing target datasets (Guz et al., 2020; Huber and Carenini, 2020a; Gessler et al., 2020).

In closing, this dissertation evaluates multilingual RST corpora and parsing models regarding intra-paragraph and inter-paragraph performances in Mandarin Chinese and English. The underlying motivation is that paragraph-level structures can remedy the smaller size of higher-level relations in RST training data. Zooming out of RST and looking at computational linguistics, the core of the RST theory is to conceptually model the coherent structure of texts, ideally across languages and genres (Taboada and Mann, 2006a). Downstream applications incorporating RST treasure its hierarchical discourse structure at all text levels, i.e., within and across sentences, paragraphs, topic segments, sections, etc. However, the existing corpora and SOTA parsers still perform unsatisfactory RST parsing at the macro level and for non-news genres. Contributions of this dissertation and the proposed future directions serve the same long-term goals: (1) to further understand the functionality of macro-level structures in medium-to-long documents; (2) to strengthen RST parsers’ capability in capturing document-level structures; and (3) to provide more effective representations for downstream document-level natural language understanding tasks.
References


Helmut Gruber and Birgit Huemer. Two Views on Text Structure: Using Rhetorical Structure Theory and Register & Genre Theory in Improving Students’ Aca-


Patrick Huber and Giuseppe Carenini. From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 185–197, Barcelona, Spain (Online), December 2020b. International Committee on


Feng Jiang, Xiaomin Chu, Sheng Xu, Peifeng Li, and Qiaoming Zhu. A macro discourse primary and secondary relation recognition method based on topic...


Jing Li, Billy Chiu, Shuo Shang, and Ling Shao. Neural Text Segmentation and Its Application to Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020. ISSN 1558-2191. doi: 10.1109/TKDE.2020.2983360. Conference Name: IEEE Transactions on Knowledge and Data Engineering.


Philippe Muller, Chloé Braud, and Mathieu Morey. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking*


Uladzimir Sidarenka. *Sentiment analysis of German Twitter*. PhD thesis, University of Potsdam, 2019. URL [https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/43742](https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/43742).


Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT...


Zining Zhu, Chuer Pan, Mohamed Abdalla, and Frank Rudzicz. Examining the rhetorical capacities of neural language models. In Proceedings of the Third