Bilingual Rhetorical Structure Parsing with Large Parallel Annotations

Anonymous ACL submission

Abstract

Discourse parsing is a crucial task in natural language processing that aims to reveal the higher-level semantic relations in a text. Despite growing interest in cross-lingual discourse parsing, challenges persist due to limited parallel data and inconsistencies in the Rhetorical Structure Theory (RST) application across languages and corpora. To address this, we introduce a parallel Russian annotation for the large and diverse English GUM RST corpus. Leveraging recent advances, our end-to-end RST parser achieves state-of-the-art results on both English and Russian corpora. It demonstrates effectiveness in both monolingual and bilingual settings, successfully transferring even with limited second-language annotation. To the best of our knowledge, this work is the first to evaluate the potential of cross-lingual end-to-end RST parsing on a manually annotated parallel corpus.

1 Introduction

Discourse parsing aims to reveal the higher-level organization of text. While the task has gained significant traction in recent years, cross-lingual rhetorical structure parsing remains a complex challenge. This stems from the inherent diversity of annotation schemes across languages within the Rhetorical Structure Theory (RST) framework and the scarcity of parallel corpora. Existing large RST corpora are inconsistent in annotation guidelines, genre representation, source selection, and relation definitions. Therefore, current studies might underestimate the true potential of RST parsers for language transfer. This study addresses these challenges by introducing a Russian version of the RST part of the Georgetown University Multilayer (GUM) corpus, encompassing all 213 original documents. This large parallel corpus provides a valuable resource for bilingual discourse analysis, enabling the development of robust RST models that can effectively capture the rhetorical structure of text in both languages.

As previous research suggests (Da Cunha and Iruskieta, 2010; Iruskieta et al., 2015; Cao et al., 2018), differences in rhetorical structures across languages primarily arise at the lower structural levels, while the global document organization exhibits some universality. Currently, top-down, unified-model frameworks (Nguyen et al., 2021; Liu et al., 2021) have proven highly effective for end-to-end RST parsing. Hypothetically, these parsers should begin by constructing a language-independent high-level structure, with language-specific nuances incorporated primarily at lower levels. This study investigates the effectiveness of an end-to-end top-down RST parser adaptation across genres in a second language, utilizing both monolingual and bilingual training data. Recognizing the substantial cost of RST annotation, we further investigate the efficient amount of second-language annotation for parser transfer.

The main contributions[1] of this work are:

1. A parallel Russian annotation of a large and diverse English GUM RST corpus dubbed RRG, enabling the development and evaluation of cross-lingual RST models. This resource enables the development and evaluation of cross-lingual RST models following the same annotation framework, addressing a critical gap in the field.

2. A unified end-to-end RST parser achieving state-of-the-art performance on diverse benchmarks in both English and Russian:

    - English: RST-DT (53.0% end-to-end Full F1), $\mathrm{GUM}_{9.1}$ (47.9% F1 – En, 47.6% F1 – bilingual),
    - Russian: RRT (45.3% F1), new RRG (44.6% F1 – Ru, 45.4% F1 – bilingual).

[1] Links to the dataset and trained models will be available upon publication.

2 Related Work

Our work is closely related to two topics, namely end-to-end and cross-lingual RST parsing, in which we review prior work in this section.

Top-down Document-level RST Parsing The paradigm of top-down rhetorical parsing has recently emerged and is receiving significant attention for its exceptional capabilities for efficient end-to-end analysis through a unified model. Zhang et al. (2020) proposed a top-down strategy for parsing rhetorical structure from a sequence of EDUs (Elementary Discourse Units). An encoder-decoder module with an internal stack iteratively ranks the split points, ultimately assigning each EDU to its corresponding rhetorical role. To account for the variation in document structure context at different levels of granularity, Kobayashi et al. (2020) presented a multi-level tree construction approach developing distinct paragraph- and sentence-level discourse unit representations. Multiple monolingual language models were tested in this framework by Kobayashi et al. (2022). Koto et al. (2021) simplified the parsing by reformulating it as a sequence labeling for sequences of EDUs. Zhang et al. (2021a) proposed computing an additional loss based on the dissimilarity between 3D representations of both gold and predicted trees, guiding the latter towards closer alignment with the original structures. Addressing the limitations of previous methods, Nguyen et al. (2021) devised an end-to-end document-level parsing model. This architecture presents two key advantages: (1) it seamlessly integrates tree construction and EDU segmentation through token-level splitting decisions, and (2) it employs beam search for non-greedy RST parsing. Liu et al. (2021) introduced a joint model where a shared LM encoder is employed for both segmentation and tree construction. The tree is built via attention over the sequence of EDUs within the current unit. We adopt this approach, with further details provided in Section 4.

Cross-lingual Rhetorical Parsing The qualitative comparison conducted by Iruskieta et al. (2015) laid the foundation for multilingual rhetorical structure analysis. Applied to a small parallel corpus across English, Spanish, and Basque (318 EDUs per language), their method revealed significant similarities in rhetorical structures between languages. Differences primarily manifested in segmentation (sentence-level discourse structure). This insight inspired subsequent efforts to bridge the gap between languages. (Cao et al., 2018) developed a Spanish-Chinese bilingual RST Treebank consisting of 50 texts per language with varying lengths (111-1774 words). Braud et al. (2017) laid the groundwork for cross-lingual parsing experiments by harmonizing RST treebanks across languages and introducing 18 unified coarse-grained rhetorical labels. Subsequent work by Iruskieta and Braud (2019) leveraged multilingual word embeddings to adapt mono- and multilingual parsers to the Basque with limited RST annotations. Liu et al. (2020, 2021) then developed a novel neural parser utilizing EDU-level machine translation (MT). These advancements, while addressing data sparsity, also reveal challenges like ensuring the rhetorical naturalness of the texts translated segment-by-segment. The recent Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022) offers RST annotations for 50 Chinese texts (9710 EDUs) spanning 5 of 10 genres found in the GUM corpus following the same relation inventory. Notably, 19 documents drawn from multilingual sources like Wikipedia, Wikinews, and wikiHow have English counterparts in GUM, although content and presentation may diverge across languages.

3 RST Corpora

This work employs three previous Rhetorical Structure Theory (RST) datasets for two languages: English (RST-DT[2] (Carlson et al., 2001), $\text{GUM}_{\text{V9.1}}$[3] (Zeldes, 2017)) and Russian ($\text{RuRSTreebank}_{\text{V2.1}}$ (Pisarevskaya et al., 2017)). Furthermore, we suggest an additional parallel annotation for the Georgetown RST annotations ($\text{GUM}_{\text{V9.1}}$) in Russian. This section discusses the datasets and preprocessing steps.

The general corpora analysis outlined in Table 1 reveals differences between the corpora extending beyond variation in genres, tree sizes, and relation labels inventory. For instance, in the RST-DT corpus, 79.4% of non-elementary sentences[4] (those containing at least one relation) are spanned by well-formed rhetorical subtrees. This high prevalence, along with explicit sentence and paragraph boundary annotation, fostered research on sentence-

[2] `https://catalog.ldc.upenn.edu/LDC2002T07`; under an LDC license.

[3] `https://github.com/amir-zeldes/gum/releases/tag/V9.1.0`; CC BY 4.0.

[4] For sentence splitting we used spaCy and `razdel` libraries for English and Russian, respectively.

	Genres	Sources	Docs	Classes	Tokens per tree			Spanned non-EDU sent., %	EDUs	EDUs per tree	Relation pairs
					min	max	median				
RST-DT (En)	1	1	385	41	30	2624	396	79.4	21789	56.6	21404
GUM (En)	12	12+	213	27	167	1879	989	72.5	26319	123.6	26106
RRT (Ru)	2	17+	233	24	2	1148	89	76.7	28372	11.7	25957
RRG (Ru)	12	12+	213	27	137	1629	833	77.0	25239	118.5	25026

Table 1: Statistics of the corpora.

level RST analysis (Soricut and Marcu, 2003; Joty et al., 2012; Nejat et al., 2017; Lin et al., 2019; Zhang et al., 2021b). In contrast, the GUM corpus takes a different approach by ignoring formal sentence and paragraph boundaries and omitting paragraph markers altogether. These differences underscore that variations in rhetorical structure, even within the same genre[5], stem not only from diverse relation sets and text sources, as Liu and Zeldes (2023) suggest, but also from fundamental differences in annotation principles.

3.1 Annotations for English

RST-DT The RST-DT corpus remains the primary benchmark for RST parsing, offering fine-grained annotations for WSJ news articles of various lengths.

GUM The Georgetown University Multilayer corpus is an expending multi-genre corpus containing multiple layers of linguistic annotation, including RST. Featuring both written and spoken language across 12 genres, it remains the largest monolingual RST annotation corpus to date.

3.2 RRT (RuRSTreebank)

We exclude the scientific portion of the RuRSTreebank corpus in our experiments, as these are reported to be the first attempts at RST annotation for Russian following the earliest incompatible guidelines (Chistova et al., 2021). The resulting dataset comprises news articles and blogs from diverse sources. It includes 5 news sources and 17 blogs covering topics such as travel, life stories, IT, cosmetics, health, politics, environment, and psychology. Despite the diversity, most documents are only partially annotated. Among the 233 document annotations, only one text is fully covered by a single tree; the remaining documents have random under-annotations. The maximum number of trees in a single `*.rs3` document reaches 42, with an average of 11.7 trees per document. This has influenced previous attempts to build a Russian parser (Chistova et al., 2021; Chistova and Smirnov, 2022), in which many efforts are directed towards predicting a look-alike forest for each full document. where efforts focus on predicting a similar forest for each full document. However, we emphasize the clear randomness of tree boundaries within the text, treating each connected tree as a separate document in our study[6]. Our approach's validity is implicitly supported by the absence of rhetorical relations for higher-level textual organization (such as HEADING or TOPIC-CHANGE) in the RRT. Additionally, we've observed that in corpora for other languages, the fully annotated tree often represents only a portion of the original text. Following established practices in end-to-end discourse parsing for RRT, we address inconsistencies in the assignment of specific relations documented by Pisarevskaya et al. (2017). The dictionary in Appendix B assists in remapping these relations during corpus preprocessing.

3.3 RRG

The Russian RST dataset from Georgetown University Multilayer corpus (RRG) was constructed by manually translating the RST annotations in $\text{GUM}_{9.1}$.

Translation We prioritized manual literary translation and genre-specific text adaptation for 213 English texts. This differs from the common practice in cross-lingual RST research relying on EDU-level machine translation.

Rhetorical Structure Alignment The translated texts were manually aligned to the original structures unit-by-unit, following the guidelines for

[5]See Appendix A for genre-wise comparison.

[6]The original train/dev/test corpus splitting is preserved. The documents are only split into `docname_part_*.rs3` files processed independently. Documents containing only a single EDU are excluded. Within the refined corpus used for experiments, 12.8% of trees are constructed of 2 to 4 elementary discourse units.

EDU segmentation in Russian developed for RRT[7]. We added or removed elementary discourse units from the tree based on the discourse segmentation in the Russian sentences. Rhetorical relations and nuclearity were assigned following the GUM RST annotation guidelines[8].

Annotation Polishing Our efforts to detect and correct misassigned labels and misaligned EDUs in the RRG draft began with an examination of the class distribution. It helped us identify obvious annotation errors, including some inherited from the original English corpus (such as rare and unlikely classes like RESTATEMENT_SN). To further refine the annotations, we trained the RST label classifier for Russian proposed by Chistova et al. (2021) on the draft dataset. This classifier served as an outlier detection tool, allowing us to detect potentially mislabeled examples. Specifically, we focused on cases where the classifier confidently predicted an incorrect class and excluded the true (annotated) class from its top 3 most probable predictions. Following the GUM relation annotation guidelines, we fixed any corrupted structures identified through this analysis.

4 End-to-End RST Parser

The rhetorical structure parsers suggested in recent years (Zhang et al., 2020; Kobayashi et al., 2020; Zhang et al., 2021a; Nguyen et al., 2021) often focused on developing innovative features to address either specific aspects of the structure construction or its global optimization. However, these approaches often overlook the integration of previously established effective features. They also frequently neglect the end-to-end performance, a fundamental aspect of any practical framework. We are building a hybrid deep model solving both segmentation and tree construction that benefits from the techniques suggested by recent work.

4.1 Base Model

As a base end-to-end deep model, we use the DMRST (Liu et al., 2021) architecture visualized in Figure 1.

The framework consists of four main modules: (1) EDU segmentation via document-level labeling, (2) hierarchical EDU encoding, (3) span-splitting decoding for tree construction, and (4) nuclearity-relation prediction using a bi-affine classifier. The

[7] https://rstreebank.ru/eng

[8] https://wiki.gucorpling.org/gum/guidelines

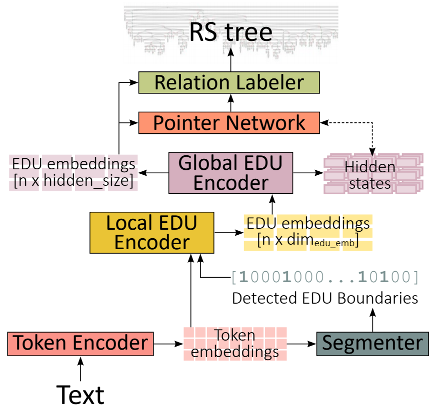


Figure 1: Architectural overview of DMRST.

encoded EDU sequence is iteratively parsed during decoding, and the classifier predicts the nuclearity and relations between adjacent units. Training minimizes the dynamic weighted average (DWA) (Liu et al., 2019) of losses for EDU segmentation, tree structure parsing, and nuclearity+relation labeling.

4.2 Modifications to the Base Model

To improve end-to-end parsing performance, we introduce modifications to the base model, focusing primarily on EDU segmentation and encoding.

Segmentation: `ToNy` The BiLSTM-CRF segmenter known by this name (Muller et al., 2019) is a simple yet robust neural token labeler. Original DMRST parser implements a feedforward token classifier (with an additional similar classifier for the right neighbor only for loss penalization)[9]. We replace the original DMRST segmentation module with a BiLSTM-CRF layer without additional losses.

Local EDU Encoding: `E-BiLSTM` Rather than averaging subword embeddings for local EDU encoding like the original method, we utilize another BiLSTM layer, which enables us to achieve better sequence encodings. The concatenation of hidden states at the final time step of each pass captures the context of the phrase more precisely than an average of its subword embeddings.

[9] Directly comparing segmentation scores from the report with ToNy's paper raises concerns due to differing methodological choices. DMRST employs a different pretrained language model, potentially augmented data, and document-level segmentation, contrasting with ToNy's sentence-level StanfordNLP splitter. Furthermore, original ToNy functions as a standalone segmenter, while DMRST incorporates segmentation into its unified encoder training for joint optimization with tree construction.

No augmentations One of the distinctive features of the original DMRST is data augmentation using corpora unification and EDU-level machine translation. However, we emphasize that annotated corpora for different languages can present different interpretations of RST with nuances in the tree constraints and relation definitions. Furthermore, EDU-level MT can result in unnatural discourse structures in the target language and offer little linguistic knowledge (although it can augment examples of some relations in the training set). Therefore, we do not consider either corpora unification or machine translation. Instead, we build a full parallel RST corpus with consistent relation inventory.

DWA Window Size Dynamic weighting is crucial for ensuring that each component of the parser receives the necessary attention during training:

$$\mathcal{L}_{total} = \sum_{k=1}^{3} \lambda_k \mathcal{L}_k,\ w_k(i-1) = \frac{\mathcal{L}_k(i-1)}{\mathcal{L}_k(i-2)} \quad (1)$$

$$\lambda_k(i) = \text{softmax}(\frac{w_k(i-1)}{Temp}) \times 3, \quad (2)$$

where the loss $\mathcal{L}_{total}$ is the DWA of task-specific losses with weights λ_i; w_k are the relative descending rates for tasks 1 (segmentation), 2 (tree construction), and 3 (relation labeling), i is an iteration index, and $Temp$ controls the softness of the task weighting. However, relying solely on the last two batches (Equation 1) is susceptible to local trend amplification, especially with smaller batches encompassing rhetorical trees of varying sizes and complexities. To address this issue, we introduce a DWA window size parameter b:

$$w_k(i-1) = \frac{\sum_{j=1}^{b} \mathcal{L}_k(i-j)}{\sum_{j=b+1}^{2b} \mathcal{L}_k(i-j)} \quad (3)$$

By analyzing a broader range of loss values, the model can effectively identify long-term trends and adjust task weights accordingly. This modification improved training stability with smaller batches, particularly on the RRT dataset comprising a large number of single-relation discourse trees.

5 Experimental Setup

In this study, we adopt the multilingual `xlm-roberta-large`[10] (Conneau et al., 2020). Hyperparameters are fixed as specified in Appendix D. We average results across five runs with varying model seeds (fixed-split corpora: GUM, RRT, RRG) or different train/dev splits (RST-DT). Bilingual experiments (Section 8) additionally involve randomly selecting 25%, 50%, and 75% of the second-language data for each of the five runs.

6 Monolingual Evaluation and Discussion

6.1 Segmentation

Segmentation performance is shown in Table 3 alongside other metrics for end-to-end parsing.

English The previous best segmentation performance belongs to the DisCut[11] method (Metheniti et al., 2023), achieving 97.6 F1 on RST-DT[12] and 95.5 F1 on $\text{GUM}_{9.0}$. Our improved DMRST+ToNy surpasses this on RST-DT with an average of 97.9% F1. The final model also outperforms the original DMRST configuration on $\text{GUM}_{9.1}$ reaching an average F1 score of 95.5% compared to 94.7%.

Russian Building upon the ToNy method (2019), Chistova and Smirnov (2022) achieve an F1 score of 89.1% on the Russian RuRSTreebank corpus (version 2.1). The DISRPT shared tasks (2019; 2021; 2023) featured an early and flawed version of RRT, which had non-hierarchical annotations of academic genres. Thus, the performance in segmentation and relation classification reported for their version of the dataset is not consistent with the version used in the current work on end-to-end discourse parsing for Russian. The details on the current version (RuRSTreebank v2.1) are outlined in Section 3.2. While the architecture modifications did not significantly impact segmentation performance on the RRT corpus, they consistently improved it on the RRG corpus, with an average increase from 96.3% F1 to 96.9% F1.

6.2 Assessing the Joint Model

Our experiment on joint training of segmentation and parsing modules within a unified architecture produced intriguing results, revealing a fundamental tension between the two tasks. Models with

[10]MIT License.

[11]A simple token classifier for sentences on top of the `XLM-RoBERTa-large`.

[12]Inter-annotator agreement for segmentation on a subset of 53 (Carlson et al., 2001) double-annotated texts within the RST-DT corpus yielded a score of 98.3% F1 (Soricut and Marcu, 2003). However, this evaluation remains limited to a small part of the corpus that does not align with its test section. The human agreement scores reported in Table 2 are obtained on the same part of the corpus (Joty et al., 2015).

	Corpus	Method	S	N	R	Full
En	RST-DT	Human	78.7	66.8	57.1	55.0
		Feng and Hirst (2014)	68.6	55.9	45.8	44.6
		DPLP (2014)	64.1	54.2	46.8	46.3
		CODRA (2015)	65.1	55.5	45.1	44.3
		Surdeanu et al. (2015)	65.3	54.2	45.1	44.2
		Li et al. (2016)	64.5	54.0	38.1	36.6
		HILDA (2016)	65.1	54.6	44.7	44.1
		Braud et al. (2016)	59.5	47.2	34.7	34.3
		Braud et al. (2017)	62.7	54.5	45.5	45.1
		Yu et al. (2018)	71.4	60.3	49.2	48.1
		Mabona et al. (2019)	67.1	57.4	45.5	45.0
		Zhang et al. (2020)	67.2	55.5	45.3	44.3
		Nguyen et al. (2021)	74.3	64.3	51.6	50.2
		Koto et al. (2021)	73.1	62.3	51.5	50.3
		Zhang et al. (2021a)	76.3	65.5	55.6	53.8
		DMRST + Cross-translation (2021)	76.7	66.2	56.5	–
		Yu et al. (2022)	76.4	66.1	54.5	53.5
		Kobayashi et al. (2022)	77.8 ± 0.3	**68.0 ± 0.5**	**57.3 ± 0.2**	55.4 ± 0.4
		DMRST (this work)	**78.7 ± 0.4**	**68.0 ± 0.6**	**57.3 ± 0.2**	**55.7 ± 0.3**
		+ `ToNy`	78.4 ± 0.7	67.4 ± 0.8	56.8 ± 0.9	55.2 ± 0.9
		+ `ToNy` + `E-BiLSTM`	78.5 ± 0.5	67.5 ± 0.7	57.0 ± 0.5	55.3 ± 0.5
	GUM v9.1	DMRST (this work)	72.7 ± 0.7	60.8 ± 0.6	52.8 ± 0.5	51.7 ± 0.4
		+ `ToNy`	72.8 ± 0.3	**61.4 ± 0.6**	**53.1 ± 0.5**	**52.0 ± 0.5**
		+ `ToNy` + `E-BiLSTM`	**73.1 ± 0.3**	61.3 ± 0.2	53.0 ± 0.3	**52.0 ± 0.3**
Ru	RRT	DMRST (this work)	81.0 ± 0.5	63.3 ± 0.9	54.2 ± 0.9	54.0 ± 0.9
		+ `ToNy`	80.9 ± 1.0	**63.4 ± 0.9**	**54.7 ± 0.9**	**54.6 ± 0.9**
		+ `ToNy` + `E-BiLSTM`	**81.2 ± 0.4**	62.9 ± 0.9	53.8 ± 1.2	53.6 ± 1.2
	RRG	DMRST (this work)	**71.5 ± 0.4**	**57.6 ± 0.2**	**49.1 ± 0.3**	**47.9 ± 0.2**
		+ `ToNy`	71.1 ± 0.5	56.6 ± 1.4	48.2 ± 1.5	47.2 ± 1.4
		+ `ToNy` + `E-BiLSTM`	70.7 ± 0.4	56.4 ± 0.5	48.3 ± 0.5	47.1 ± 0.5

Table 2: RST parsing performance evaluated on the gold EDU segmentation. Micro F1 scores (original Parseval); average and standard deviation. Missing values are not reported in the cited work.

higher F1 scores on gold-standard segmentation (Table 2) performed worse on both segmentation and end-to-end parsing metrics than models with lower gold-segmentation scores but better utilization of their predicted segments (Table 3). This pattern suggests that the encoder representations are being pulled in two opposing directions during fine-tuning. Sentence segmentation relies heavily on local cues within sentences, leading segmentation-optimized models to develop encodings for fine-grained syntactic patterns. However, building a document-level parse tree requires capturing long-range context and global relationships, demanding encodings that recognize complex discourse units. Therefore, directly comparing jointly trained models on gold-EDU trees may not be reliable in this scenario. The following discussion delves into the end-to-end parsing evaluated in Table 3.

English The enhanced models achieve state-of-the-art results for end-to-end English RST parsing. Leveraging `ToNy` segmentation for the RST-DT dataset and both `ToNy` and BiLSTM EDU encoding for the GUM dataset, we obtain a substantial improvement in unlabeled tree construction, measured by the Span metric (average increase of 0.8% for RST-DT and 1.9% for GUM). This gain is noteworthy considering the widespread use of unlabeled rhetorical trees in RST parsing applications (Guzmán et al., 2014; Khosla et al., 2021). Nuclearity assignment, crucial for tasks like summarization and sentiment analysis (Goyal and Eisenstein, 2016; Fu et al., 2016; Huber and Carenini, 2020), also benefits from our approach. The best models achieve an average F1-score of 64.8% (+0.7) on RST-DT and 56.1% (+1.9) on GUM for the Nuclearity metric. Finally, the full rhetorical structure construction for both datasets achieves 53.0% for RST-DT and 47.9% for GUM.

Russian While the enhanced model noticeably improved performance on other corpora, it surprisingly failed to do so on RRT. This disparity might be attributed to the overfitting of the ToNy segmenter, potentially caused by the larger batch size necessary for stable RRT training (Appendix D). RRT's smaller median tree size (Table 1) results in the highest Span score on gold-standard segmentation across all corpora (Table 2). Building trees on EDUs predicted with 92% F1 (Table 3) significantly drops the Span metric (15% F1 gap). Similar to the original GUM corpus, the model incorporating both modifications achieved the best results on RRG, exhibiting an average F1-score of 44.6%.

	Corpus	Method	Segm.	S	N	R	Full
En	RST-DT	SegBot (2018) & Zhang et al. (2020)	92.2	62.3	50.1	40.7	39.6
		Nguyen et al. (2021)	96.3	68.4	59.1	47.8	46.6
		DMRST (2021)	96.4	69.8	59.4	49.4	48.6
		+ Cross-translation	96.5	70.4	60.6	51.6	50.1
		DMRST (this work)	97.3 ± 0.1	74.3 ± 0.6	64.1 ± 0.7	53.9 ± 0.5	52.4 ± 0.5
		+ ToNy	**97.9 ± 0.1**	**75.1 ± 0.7**	**64.8 ± 0.7**	**54.5 ± 0.9**	**53.0 ± 0.9**
		+ ToNy + E-BiLSTM	97.8 ± 0.1	74.8 ± 0.5	64.5 ± 0.8	**54.5 ± 0.7**	**53.0 ± 0.7**
	GUM v9.1	DMRST (this work)	94.7 ± 0.4	65.0 ± 0.5	54.2 ± 0.5	47.3 ± 0.5	46.4 ± 0.4
		+ ToNy	95.4 ± 0.1	66.4 ± 0.3	55.8 ± 0.5	48.5 ± 0.5	47.6 ± 0.6
		+ ToNy + E-BiLSTM	**95.5 ± 0.1**	**66.9 ± 0.5**	**56.1 ± 0.3**	**48.8 ± 0.4**	**47.9 ± 0.4**
Ru	RRT	DMRST (this work)	**92.4 ± 0.3**	**66.5 ± 1.0**	**52.4 ± 1.2**	**45.3 ± 1.0**	**45.3 ± 1.0**
		+ ToNy	**92.4 ± 0.2**	65.4 ± 1.1	51.3 ± 0.6	44.6 ± 0.5	44.5 ± 0.5
		+ ToNy + E-BiLSTM	92.2 ± 0.2	65.9 ± 0.5	51.0 ± 0.7	43.9 ± 1.0	43.8 ± 1.0
	RRG	DMRST (this work)	96.3 ± 0.1	65.6 ± 0.3	52.8 ± 0.3	45.1 ± 0.2	44.0 ± 0.3
		+ ToNy	96.7 ± 0.2	**66.6 ± 0.9**	53.0 ± 1.7	45.3 ± 1.7	44.3 ± 1.5
		+ ToNy + E-BiLSTM	**96.9 ± 0.2**	66.5 ± 0.4	**53.3 ± 0.6**	**45.8 ± 0.5**	**44.6 ± 0.4**

Table 3: End-to-end parsing performance. Micro F1 scores (original Parseval); average and standard deviation.

En	Ru	En					Ru				
		Segm.	S	N	R	Full	Segm.	S	N	R	Full
100%	0%	95.5 ± 0.1	66.9 ± 0.5	56.1 ± 0.3	48.8 ± 0.4	47.9 ± 0.4	95.5 ± 0.3	63.9 ± 0.7	51.4 ± 1.0	43.4 ± 0.6	42.2 ± 0.6
	25%	95.5 ± 0.1	66.4 ± 0.7	55.1 ± 1.0	48.2 ± 1.0	47.4 ± 1.0	96.4 ± 0.3	66.3 ± 0.6	53.8 ± 0.6	45.9 ± 0.7	44.9 ± 0.6
	50%	95.5 ± 0.1	66.6 ± 0.5	55.4 ± 0.6	48.7 ± 0.6	47.7 ± 0.7	96.6 ± 0.2	67.0 ± 0.5	54.2 ± 0.6	46.6 ± 0.8	45.5 ± 0.8
	75%	95.6 ± 0.2	67.2 ± 0.2	55.7 ± 0.5	48.9 ± 0.6	47.9 ± 0.5	96.8 ± 0.2	67.0 ± 0.4	54.0 ± 0.5	46.2 ± 0.5	45.0 ± 0.5
	100%	95.3 ± 0.1	66.4 ± 0.7	55.2 ± 0.6	48.6 ± 0.6	47.6 ± 0.7	96.8 ± 0.1	66.9 ± 0.4	54.3 ± 0.3	46.5 ± 0.4	45.4 ± 0.4

Table 4: Performance of the models trained with second language data injection.

7 Cross-Dataset Compatibility in Russian RST Parsing

This section explores the cross-dataset compatibility of Russian RST parsing by comparing two relation inventories derived from RRT and RRG parsers using a data-driven approach.

Relation Labeling To categorize the discourse unit pairs connected in the annotated corpora, we trained the relation classifier for Russian developed by Chistova et al. (2021). It is an ensemble of a feature-rich classifier and an ELMo-driven classifier. The feature-rich classifier includes a comprehensive dictionary of discourse cues in Russian, various morpho-syntactic features, a sentiment classifier, and USE vectors (Cer et al., 2018). The neural classifier is based on the BiMPM architecture (Wang et al., 2017), and utilizes the ELMo model for Russian as well as pre-trained fastText embeddings (Bojanowski et al., 2017) and character n-gram embeddings to encode a discourse unit. The RRT dataset, which includes 24 classes, yielded a 48.9% macro F1 score, while the RRG dataset, which includes 27 classes, yielded a 46.3% macro F1 score (see Appendix C for detailed results). Cross-dataset classification results illustrated in Appendix C Figure 5 indicate a notable overlap among the majority of classes from the two datasets while also highlighting the challenge of RST treebanks unification across languages and frameworks.

8 Cross-Lingual Evaluation

In this section, we explore the capabilities of our best `+ToNy+E-BiLSTM` model in two scenarios: (1) its performance on an unseen or under-annotated language, and (2) its bilingual adaptation when trained on a fully-annotated parallel corpus. We assess the performance of a model on a new language, analyzing how expanding the parallel training data influences its ability to parse diverse writing and speech styles. With the English training data held constant, we investigate its ability to adapt to different genres in Russian.

Direct Transfer By employing documents that differ only in language, we isolate the impact of language on RST parsing within zero-shot generalization, offering a more nuanced evaluation compared to typical mixed-source approaches. As demonstrated in Table 4, the RST parser achieves remarkable results on Russian test documents in zero-shot setting (0%), showcasing the strength of multilingual language models. It performs nearly on par with the monolingual parser specifically trained on Russian data (RRG, Table 3). Although the Russian parser exhibits improvements across all met-

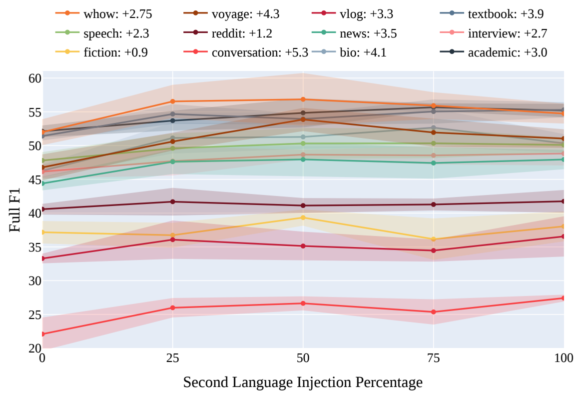


Figure 2: Impact of second language injection on the end-to-end Full performance.

Test Language Train Data	English GUM	 GUM+RRG	Russian GUM	 RRG	 GUM+RRG
academic	56.3	55.5 (–0.8)	52.1	55.7	55.2 (–0.5)
bio	51.5	52.5 (+1.0)	46.3	52.2	50.3 (–1.9)
conversation	29.3	30.2 (+0.9)	22.1	25.9	27.4 (+1.5)
fiction	38.5	40.2 (+1.7)	37.2	36.7	38.0 (+1.3)
interview	55.1	54.7 (–0.4)	46.1	47.3	48.8 (+1.5)
news	55.0	52.9 (–2.1)	44.4	45.9	47.9 (+2.0)
reddit	44.0	42.3 (–1.7)	40.6	41.5	41.8 (+0.3)
speech	57.6	57.2 (–0.4)	47.8	50.2	50.1 (–0.1)
textbook	57.0	56.4 (–0.6)	51.4	53.6	55.3 (+1.7)
vlog	41.7	40.6 (–1.1)	33.3	35.5	36.6 (+1.1)
voyage	44.1	43.4 (–0.7)	46.8	49.3	51.0 (+1.7)
whow	57.0	56.8 (–0.2)	52.0	54.1	54.7 (+0.6)
all	47.9	47.6 (–0.3)	42.2	44.6	45.4 (+0.8)

Table 5: Mono- vs. bilingual model evaluation (avg. end-to-end Full F1).

rics (segmentation: +1.4%, Span splitting: +2.6%, Nuclearity assignment: +1.9%, Full: +2.4%), the gap remains relatively narrow, demonstrating the effectiveness of the original GUM-based parser across languages. Reversing the direction (Russian to English) revealed a substantial performance drop (Table 11, Appendix E). Its F1 score for English segmentation is only 86.9%. This disparity likely stems from heavy reliance on commas to separate elementary discourse units in Russian (examples in Figure 4, Appendix E). With only 18.5% of EDUs ending with commas in GUM compared to a staggering 37.5% in RRG, the segmenter became overly reliant on a feature less common in English.

Mixed Train Data The objective of this experiment is to estimate the data requirements for successful cross-lingual parser transfer in RST parsing, a task that relies on laborious expert annotation. We evaluate cross-lingual transfer performance across different amounts of annotation, ranging from 25% to 100% of the target language corpus. Our evaluation considers an ideal scenario involving full parallel data. Table 4 presents the model's performance as the number of labeled examples in the second language increases. We observe a gradual improvement in the model's ability to construct rhetorical trees with attached nuclearities. However, the rhetorical labeling accuracy plateaus at approximately 50% of second language annotations. The genre-specific performance of the model is illustrated in Figure 2. A more detailed evaluation is provided in Appendix E. Genres such as *wikihow*, *textbook*, *academic*, *voyage*, *bio* (Wikipedia), *speech*, *interview*, and *news* exhibit the highest adaptation to the second language. Spoken discourse genres achieved the lowest parsing scores but showed notable adaptation (*vlog*: 33.3% to 36.6% F1; *conversation*: 22.1% to 27.4% F1).

The bilingual model outperforms the monolingual RRG model (44.6% F1), achieving an impressive Full end-to-end score of 45.4% F1. Despite a slight F1 decrease in English, the bilingual parser excelled in 9 out of 12 genres in Russian (as detailed in Table 5). This underscores its efficacy in cross-lingual transfer.

9 Conclusion

This study addresses the challenges of cross-lingual discourse parsing. We introduce a large parallel Russian annotation of the multigenre GUM RST corpus and assess the performance of an end-to-end top-down model in bilingual rhetorical structure parsing. The top-down unified parser employing a multilingual language model established a strong baseline on end-to-end parsing in both languages. Further analysis explored direct parser transfer without second-language data. Surprisingly, transferring the English parser to Russian achieved comparable quality to the monolingual parser. However, the reverse transfer suffered due to nuances in Russian discourse segmentation, underlining the critical role of language-specific features in language transfer. We investigated the effectiveness of porting the analyzer with limited second-language data. Our findings demonstrate that even with minimal data, such transfer remains effective. Finally, training the bilingual parser on the entire parallel dataset yielded the best discourse parsing performance in Russian, and strong performance in English.

Limitations

While the written sections of the corpus are well-adapted into Russian, accurately capturing the nuances of Russian spontaneous speech in documents outlining English spoken discourse (*vlog*, *conversation*) through translation can be challenging. This presents an exciting opportunity for future research to explore the unique RST features of spoken discourse in Russian.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal Sentence Encoder. *arXiv preprint arXiv:1803.11175*.

Elena Chistova, Artem Shelmanov, Dina Pisarevskaya, Maria Kobozeva, Vadim Isakov, Alexander Panchenko, Svetlana Toldova, and Ivan Smirnov. 2021. RST discourse parser for Russian: an experimental study of deep learning models. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*, pages 105–119. Springer.

Elena Chistova and Ivan Smirnov. 2022. Discourse-aware text classification for argument mining. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2022)*, 21, pages 93–105.

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

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

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

Xianghua Fu, Wangwang Liu, Yingying Xu, Chong Yu, and Ting Wang. 2016. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In *Proceedings of The 8th Asian Conference on Machine Learning*, pages 17–32. PMLR.

Naman Goyal and Jacob Eisenstein. 2016. A joint model of rhetorical discourse structure and summarization. In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 25–34, Austin, TX. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2020. MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.

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

Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49:263–309.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915, Jeju Island, Korea. Association for Computational Linguistics.

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

Sopan Khosla, James Fiacco, and Carolyn Rosé. 2021. Evaluating the impact of a hierarchical discourse representation on entity coreference resolution performance. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1645–1651, Online. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down rst parsing utilizing granularity levels in documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2022. A simple and strong baseline for end-to-end neural RST-style discourse parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6725–6737, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. SegBot: a generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4166–4172.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas. Association for Computational Linguistics.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

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

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

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2284–2295, Hong Kong, China. Association for Computational Linguistics.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298, Saarbrücken, Germany. Association for Computational Linguistics.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.

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

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Towards building a discourse-annotated corpus of Russian. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017)*, pages 201–212.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Mihai Surdeanu, Tom Hicks, and Marco Antonio Valenzuela-Escárcega. 2015. Two practical Rhetorical Structure Theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado. Association for Computational Linguistics.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, page 4144–4150. AAAI Press.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

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

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. 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 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021a. Adversarial learning for discourse rhetorical structure parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3946–3957, Online. Association for Computational Linguistics.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2021b. A language model-based generative classifier for sentence-level discourse parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2432–2446, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Sentence Subtrees Coverage

Examining tree-covered non-elementary sentences in the analyzed corpora (see Table 6) reveals evident disparities in formal structure between annotation schemas, even within the recurring *news* genre.

Corpus	Genre	En	Ru
RST-DT	*news*	79.4	–
GUM, RRG	*academic*	72.0	77.0
	bio	61.1	72.8
	conversation	65.8	68.2
	fiction	70.4	78.5
	interview	71.4	78.1
	news	69.0	79.7
	reddit	73.0	77.0
	speech	85.8	86.7
	textbook	78.5	76.4
	vlog	75.3	77.6
	voyage	71.3	71.4
	whow	77.5	78.4
RRT	*blogs*	–	71.6
	news	–	82.9

Table 6: Spanned non-EDU sentences, %

While (Soricut and Marcu, 2003) briefly mention a 95% coverage of sentences spanned by well-formed rhetorical subtrees in RST-DT, our analysis, based on automatic sentence segmentation and counting within binarized trees (the standard format for RST parsing), suggests a more conservative estimate of 86%. Notably, even among non-elementary sentences (those containing at least two elementary units) there remains a prevalence of 79.4% well-formed rhetorical trees in the corpus. This value exceeds what has been observed in other examined corpora.

B RRT Preprocessing Details

Table 7 provides information about the common renaming of mislabeled samples in RRT.

The mislabelings, which persist in version 2.1 and are consequently addressed during corpus preprocessing, can be attributed to the following factors:

- **Relation selection errors.** TThe Antithesis relation is intentionally excluded from the corpus during annotation. However, a few instances of this class within the corpus clearly imply the Attribution relation. Furthermore, Restatement_SN(NS), Preparation_NS, Elaboration_SN are considered impossible according to the annotation manual.

Original Annotation	Preprocessing
antithesis	Attribution
cause, effect, cause-effect	Cause-effect
condition, motivation	Condition
evaluation, interpretation, interpretation-evaluation	Interpetation-evaluation
RESTATEMENT_SN	CONDITION_SN
RESTATEMENT_NS	ELABORATION_NS
SOLUTIONHOOD_NS	SOLUTIONHOOD_SN
PREPARATION_NS	ELABORATION_NS
ELABORATION_SN	PREPARATION_SN
BACKGROUND_NS	ELABORATION_SN

Table 7: Common renaming of mislabeled relations during RRT preprocessing.

- **Artifacts of shifting relation definitions.** In pursuit of objectivity and annotation agreement, Pisarevskaya et al. (2017) combined or eliminated certain initial relations (cause, effect, motivation, evaluation, interpretation). Nevertheless, remnants of these fine-grained labels persist within the corpus.

C Relation Classification Results

Table 8 presents a detailed rhetorical relation classification performance for each corpus employing a standalone classifier. The task is treated in the context of the end-to-end system, with merged relation and nuclearity.

Figure 3 shows confusion matrices for the same classification models focusing only on the coarse-grained relation. Overlapping RST relation-nuclearity classes across two corpora are illustrated in Figure 5. Confidently predicted relations (entropy >75th percentile) are shown on the right, with the target corpus's ground truth relations on the left. Only frequent transitions (>2.5% of gold class) are included. These figures reveal recurring patterns of overlapping relations in the two annotation types. The classes ORGANIZATION_NS, MODE, CONTEXT_SN, and ORGANIZATION_NS in the RRG corpus do not correspond with certain classes in RRT when examining the mentioned discourse unit features. The RRT-trained classifier consistently assigns the CONDITION class to both RRG's CONTINGENCY (contingency-condition) and CONTEXT (context-circumstance) classes. For parser efficiency, RRG merges its specific adversative classes (antithesis, concession, contrast) into a single ADVERSATIVE category. This unified category maps to two distinct relations in the RRT: CONTRAST and CONCESSION, leading to inconsistencies in

	P	R	F1	Num.
	RRT			
Attribution_NS	87.21	97.40	92.02	77
Attribution_SN	77.05	94.95	85.07	198
Background_SN	00.00	00.00	00.00	10
Cause-effect_NS	50.88	37.18	42.96	78
Cause-effect_SN	43.18	48.72	45.78	78
Comparison_NN	35.71	26.32	30.30	38
Concession_NS	83.33	90.91	86.96	22
Concession_SN	40.00	20.00	26.67	10
Condition_NS	53.47	75.00	62.43	72
Condition_SN	62.38	67.74	64.95	93
Contrast_NN	70.94	76.60	73.66	188
Elaboration_NS	52.72	71.21	60.59	639
Evidence_NS	26.67	08.89	13.33	45
Evidence_SN	00.00	00.00	00.00	12
Interpretation-evaluation_NS	45.24	39.58	42.22	144
Interpretation-evaluation_SN	33.33	15.38	21.05	13
Joint_NN	72.18	60.12	65.60	682
Preparation_SN	56.44	48.72	52.29	117
Purpose_NS	89.06	78.08	83.21	73
Purpose_SN	55.00	57.89	56.41	19
Restatement_NN	33.33	22.73	27.03	22
Sequence_NN	59.72	30.50	40.38	141
Solutionhood_SN	51.16	48.89	50.00	45
same-unit_NN	59.02	45.00	51.06	80
Macro avg.	51.58	48.41	48.92	2896
	RRG			
adversative_NN	24.32	17.31	20.22	52
adversative_NS	35.85	33.33	34.55	57
adversative_SN	36.23	51.02	42.37	49
attribution_NS	84.00	72.41	77.78	29
attribution_SN	69.47	88.35	77.78	103
causal_NS	29.55	16.46	21.14	79
causal_SN	07.14	05.88	06.45	17
context_NS	60.56	42.16	49.71	102
context_SN	35.24	30.58	32.74	121
contingency_NS	71.43	71.43	71.43	14
contingency_SN	86.49	84.21	85.33	38
elaboration_NS	50.66	69.33	58.54	551
evaluation_NS	33.80	23.30	27.59	103
evaluation_SN	50.00	07.14	12.50	14
explanation_NS	54.41	26.62	35.75	139
explanation_SN	20.00	03.57	06.06	28
joint_NN	60.69	71.48	65.64	568
mode_NS	46.43	31.71	37.68	41
mode_SN	00.00	00.00	00.00	3
organization_NS	73.68	96.55	83.58	29
organization_SN	78.57	65.13	71.22	152
purpose_NS	85.07	82.61	83.82	69
purpose_SN	75.00	85.71	80.00	7
restatement_NN	37.50	32.14	34.62	28
restatement_NS	16.67	04.00	06.45	25
same-unit_NN	82.61	45.97	59.07	124
topic_SN	63.27	73.81	68.13	42
Macro avg.	50.69	45.64	46.30	2584

Table 8: Performance of the relation classification model on Russian corpora.

nuclearity correspondence. The classifiers exhibit similar error patterns across both corpora. For instance, despite having its own dedicated Evidence relation within the broader EXPLANATION category, the RRG classifier consistently misidentifies the RRT’s EVIDENCE samples as ATTRIBUTION, mirroring 14% of the RRT classifier’s predictions. This suggests a bias in both models towards interpreting references to information sources as attributions, regardless of the intended meaning. Meanwhile, RRT’s CAUSE-EFFECT class absorbs EXPLANATION’s Justify and Motivation, encompassing both event causality and justifications (except for EVIDENCE).

D Implementation Details

Table 9 shows the hyperparameters used in our experiments.

	RST-DT	GUM	RRG	RRT
batch size (# of trees)	2	1	1	6
b_{DWA} (# of trees)	12	12	12	24
	`LM`			
hidden size	1024			
sliding window length	400			
learning rate	2e-05			
	`Parser`			
hidden size	1024	1024	1024	768
dropout (segmenter input)	0.4			
dropout (encoder input)	0.5			
learning rate	1e-04			
	`ToNy`			
hidden size	200			
	`E-BiLSTM`			
hidden size	512			

Table 9: Parameters used in the experiments.

The experiments are performed on an NVIDIA Tesla v100 GPU. A single run takes 4 to 8 GPU hours, depending on the dataset and batch size.

E Genre-wise Evaluation

Tables 10, 11, and 12 offer in-depth performance metrics for the end-to-end RST parsing in both languages. Additionally, Figure 4 provides an example for segmentation mistakes made by the Russian-trained monolingual parser on English text.

	en					**ru**				
	Segm	S	N	R	Full	Segm	S	N	R	Full
academic	94.6 ± 0.6	72.7 ± 1.3	64.0 ± 1.9	56.9 ± 1.7	56.3 ± 1.7	94.6 ± 0.5	72.6 ± 1.8	62.9 ± 1.1	55.8 ± 0.8	55.7 ± 0.8
bio	97.7 ± 0.6	68.1 ± 1.8	57.0 ± 2.9	53.2 ± 2.1	51.5 ± 2.1	98.5 ± 0.3	69.0 ± 1.8	58.4 ± 1.1	52.8 ± 1.2	52.2 ± 1.2
conversation	95.5 ± 0.3	49.5 ± 1.3	39.0 ± 1.5	29.8 ± 1.4	29.3 ± 1.6	95.5 ± 0.5	48.5 ± 1.2	33.8 ± 1.1	27.4 ± 1.2	25.9 ± 1.4
fiction	93.9 ± 0.7	59.3 ± 2.4	47.8 ± 2.9	39.7 ± 2.2	38.5 ± 2.3	96.2 ± 0.6	61.0 ± 1.1	47.3 ± 1.2	38.2 ± 0.6	36.7 ± 1.0
interview	95.1 ± 0.4	73.8 ± 0.6	65.7 ± 1.3	55.3 ± 1.2	55.1 ± 1.1	96.6 ± 0.3	71.6 ± 1.9	60.3 ± 0.7	47.3 ± 0.8	47.3 ± 0.8
news	94.6 ± 0.8	69.0 ± 1.9	60.4 ± 2.4	56.7 ± 2.0	55.0 ± 2.1	96.3 ± 0.5	65.7 ± 2.1	54.2 ± 3.2	47.6 ± 1.2	45.9 ± 1.7
reddit	93.3 ± 0.6	60.5 ± 1.1	51.5 ± 1.4	44.5 ± 1.6	44.0 ± 1.4	97.7 ± 0.3	61.1 ± 1.3	48.6 ± 1.6	42.7 ± 1.4	41.5 ± 1.7
speech	97.5 ± 0.4	79.1 ± 1.7	67.4 ± 2.4	57.8 ± 1.8	57.6 ± 2.0	96.0 ± 0.6	70.5 ± 2.5	58.7 ± 1.9	50.9 ± 0.5	50.2 ± 0.5
textbook	97.5 ± 0.3	78.7 ± 1.3	66.1 ± 1.8	57.4 ± 2.0	57.0 ± 1.9	97.4 ± 0.3	76.0 ± 2.0	62.7 ± 2.3	54.6 ± 2.0	53.6 ± 2.0
vlog	95.6 ± 0.5	61.9 ± 1.0	48.8 ± 2.0	43.5 ± 1.5	41.7 ± 1.7	97.9 ± 0.3	65.8 ± 2.1	43.2 ± 2.1	38.8 ± 1.5	35.5 ± 1.3
voyage	94.6 ± 0.5	67.2 ± 1.9	51.6 ± 2.2	44.6 ± 2.0	44.1 ± 1.9	99.0 ± 0.1	73.7 ± 0.9	58.1 ± 0.8	50.4 ± 1.2	49.3 ± 1.0
whow	97.3 ± 0.3	75.7 ± 0.9	64.3 ± 1.9	58.6 ± 1.7	57.0 ± 1.7	97.8 ± 0.5	75.5 ± 1.7	64.4 ± 2.3	55.5 ± 2.1	54.1 ± 2.2
all	95.5 ± 0.1	66.9 ± 0.5	56.1 ± 0.3	48.8 ± 0.4	47.9 ± 0.4	96.9 ± 0.2	66.5 ± 0.4	53.3 ± 0.6	45.8 ± 0.5	44.6 ± 0.4

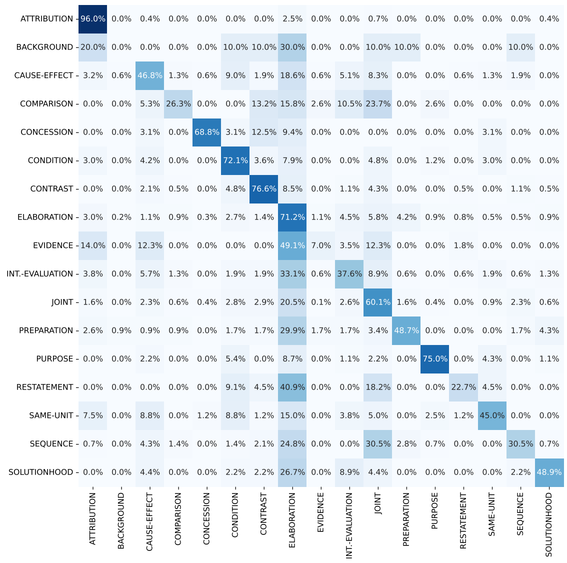
Table 10: Detailed evaluation of the monolingual parsers.

	ru → en					**en** → ru				
	Segm	S	N	R	Full	Segm	S	N	R	Full
academic	83.1 ± 1.3	52.0 ± 4.3	43.2 ± 3.2	39.0 ± 3.0	38.7 ± 2.9	93.1 ± 0.9	69.2 ± 0.8	61.5 ± 0.2	52.1 ± 0.9	52.1 ± 0.9
bio	94.4 ± 0.5	63.0 ± 1.8	50.1 ± 2.8	45.9 ± 3.0	44.8 ± 3.2	97.3 ± 0.4	66.3 ± 1.0	54.6 ± 0.5	47.5 ± 0.8	46.3 ± 0.9
conversation	91.6 ± 0.6	42.4 ± 1.7	30.8 ± 2.0	23.5 ± 1.2	22.8 ± 1.4	94.4 ± 0.7	45.5 ± 2.5	32.9 ± 3.3	23.2 ± 2.5	22.1 ± 2.4
fiction	85.3 ± 0.8	47.8 ± 2.6	35.9 ± 2.6	28.8 ± 1.9	27.7 ± 1.7	94.9 ± 0.7	60.0 ± 2.4	48.1 ± 2.8	38.1 ± 1.8	37.2 ± 1.7
interview	83.2 ± 1.4	43.9 ± 3.6	37.1 ± 2.3	29.6 ± 2.9	29.5 ± 2.7	95.6 ± 0.8	69.7 ± 1.3	58.2 ± 1.0	46.9 ± 0.8	46.1 ± 1.0
news	84.5 ± 1.8	45.9 ± 3.3	38.7 ± 3.4	36.9 ± 2.9	34.8 ± 2.6	93.5 ± 1.2	61.9 ± 1.2	51.8 ± 1.7	45.8 ± 0.9	44.4 ± 1.0
reddit	83.1 ± 1.4	37.1 ± 2.7	30.7 ± 1.8	24.9 ± 2.5	24.6 ± 2.3	97.1 ± 0.4	59.8 ± 2.1	48.5 ± 1.7	41.1 ± 0.8	40.6 ± 0.8
speech	83.7 ± 1.6	44.6 ± 2.1	34.8 ± 1.3	29.8 ± 2.4	29.5 ± 2.5	94.5 ± 0.5	69.8 ± 1.1	56.6 ± 0.9	48.6 ± 1.6	47.8 ± 1.3
textbook	87.8 ± 1.4	56.2 ± 2.3	45.7 ± 2.3	39.9 ± 2.3	39.2 ± 2.1	95.1 ± 0.3	71.2 ± 0.9	58.0 ± 1.8	51.9 ± 0.6	51.4 ± 0.6
vlog	88.1 ± 1.9	52.7 ± 3.4	35.7 ± 3.1	32.8 ± 3.5	30.2 ± 3.9	97.2 ± 0.1	61.6 ± 1.8	41.5 ± 0.6	36.1 ± 1.0	33.3 ± 0.7
voyage	85.1 ± 1.2	46.6 ± 2.6	34.9 ± 2.3	28.8 ± 1.7	28.7 ± 1.5	96.7 ± 0.3	71.6 ± 1.4	55.2 ± 1.6	48.8 ± 2.4	46.8 ± 1.9
whow	90.6 ± 1.8	58.7 ± 3.8	49.8 ± 3.9	42.9 ± 3.2	42.1 ± 3.0	96.5 ± 0.5	74.0 ± 1.6	61.9 ± 1.8	54.1 ± 1.7	52.0 ± 1.9
all	86.9 ± 1.0	49.0 ± 2.2	38.6 ± 2.1	33.1 ± 1.9	32.2 ± 1.9	95.5 ± 0.3	63.9 ± 0.7	51.4 ± 1.0	43.4 ± 0.6	42.2 ± 0.6

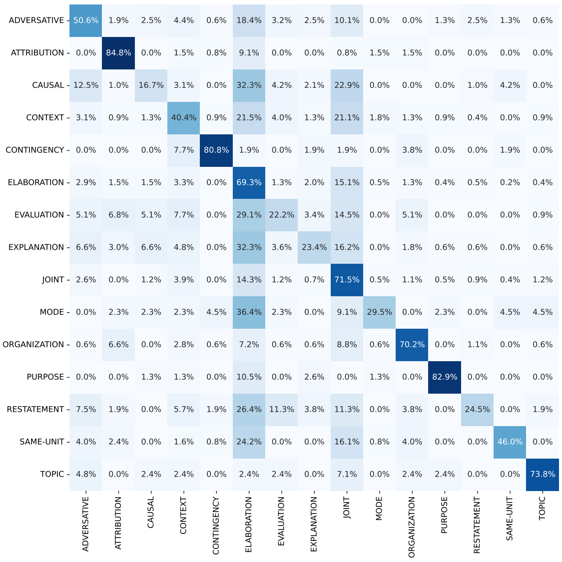
Table 11: Evaluating monolingual parsing transfer to a second language.

	en+ru → en					**en+ru** → ru				
	Segm	S	N	R	Full	Segm	S	N	R	Full
academic	94.2 ± 0.4	71.6 ± 1.1	63.1 ± 2.0	55.9 ± 2.1	55.5 ± 2.3	94.9 ± 0.6	72.9 ± 1.7	63.2 ± 1.6	55.3 ± 1.0	55.2 ± 1.0
bio	97.6 ± 0.3	70.0 ± 0.9	58.4 ± 1.0	54.0 ± 1.4	52.5 ± 1.5	98.4 ± 0.4	68.1 ± 1.9	57.5 ± 1.7	51.4 ± 1.4	50.3 ± 1.4
conversation	95.1 ± 0.1	51.5 ± 1.5	39.2 ± 0.7	31.1 ± 1.4	30.2 ± 1.3	95.3 ± 0.4	47.8 ± 1.0	34.8 ± 1.3	28.9 ± 0.5	27.4 ± 0.5
fiction	93.3 ± 0.6	59.2 ± 2.8	48.8 ± 2.3	41.2 ± 1.8	40.2 ± 1.8	96.6 ± 0.3	62.8 ± 1.9	49.6 ± 0.7	39.2 ± 2.0	38.0 ± 2.2
interview	94.6 ± 0.5	71.7 ± 1.2	63.5 ± 1.8	55.2 ± 1.3	54.7 ± 1.2	96.9 ± 0.1	70.0 ± 1.7	60.2 ± 1.9	49.2 ± 1.8	48.8 ± 1.8
news	94.8 ± 0.7	67.5 ± 2.4	59.2 ± 1.8	54.5 ± 1.6	52.9 ± 1.7	96.8 ± 0.7	68.5 ± 0.6	56.8 ± 1.7	49.6 ± 1.0	47.9 ± 1.4
reddit	92.6 ± 0.8	58.5 ± 1.5	48.9 ± 2.3	43.0 ± 2.2	42.3 ± 2.2	97.2 ± 0.3	60.9 ± 1.6	49.4 ± 2.0	42.5 ± 1.6	41.7 ± 1.7
speech	97.3 ± 0.3	75.7 ± 1.6	64.8 ± 1.9	57.2 ± 1.1	57.2 ± 1.1	96.3 ± 0.5	69.9 ± 2.4	57.5 ± 1.0	50.7 ± 1.1	50.1 ± 1.1
textbook	97.5 ± 0.4	77.3 ± 1.7	65.3 ± 2.0	57.3 ± 0.8	56.4 ± 0.9	97.1 ± 0.3	77.1 ± 0.6	64.6 ± 1.0	56.1 ± 1.3	55.3 ± 1.1
vlog	95.9 ± 0.4	62.8 ± 2.0	46.1 ± 2.6	42.8 ± 2.8	40.6 ± 2.7	97.8 ± 0.5	66.0 ± 1.7	46.0 ± 3.1	39.8 ± 3.4	36.5 ± 3.0
voyage	94.2 ± 0.5	65.7 ± 2.5	49.5 ± 3.0	43.7 ± 2.6	43.4 ± 2.6	98.5 ± 0.3	76.4 ± 1.5	60.0 ± 1.9	51.7 ± 1.5	51.0 ± 1.4
whow	97.2 ± 0.3	75.5 ± 1.3	65.0 ± 1.8	58.3 ± 1.9	56.8 ± 1.6	97.8 ± 0.3	75.9 ± 1.5	64.5 ± 2.5	56.3 ± 1.1	54.7 ± 1.5
all	95.3 ± 0.1	66.4 ± 0.7	55.2 ± 0.6	48.6 ± 0.6	47.6 ± 0.7	96.8 ± 0.1	66.9 ± 0.4	54.3 ± 0.3	46.5 ± 0.4	45.4 ± 0.4

Table 12: Bilingual parser performance.

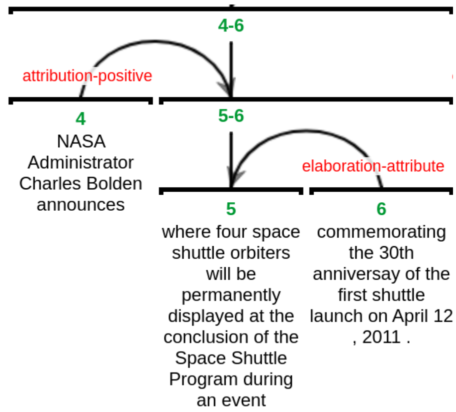


(a) RRT

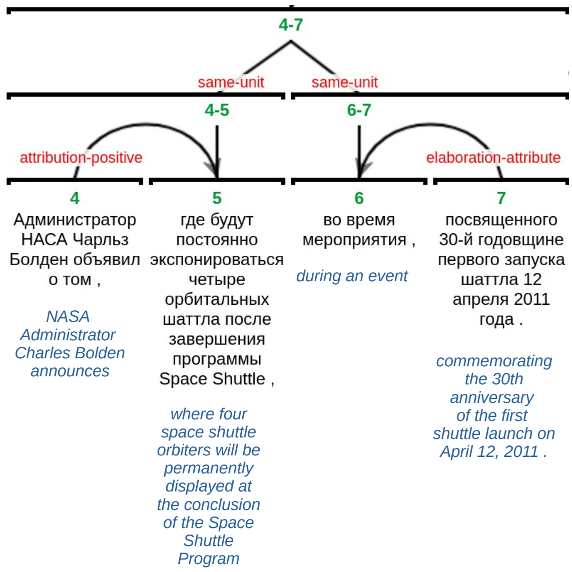


(b) RRG

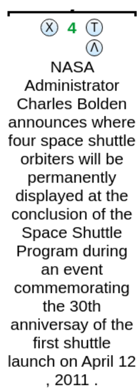
Figure 3: Confusion matrices for the relation classification on Russian corpora; nuclearity omitted.



(a) Original annotation from $GUM_{9.1}$.

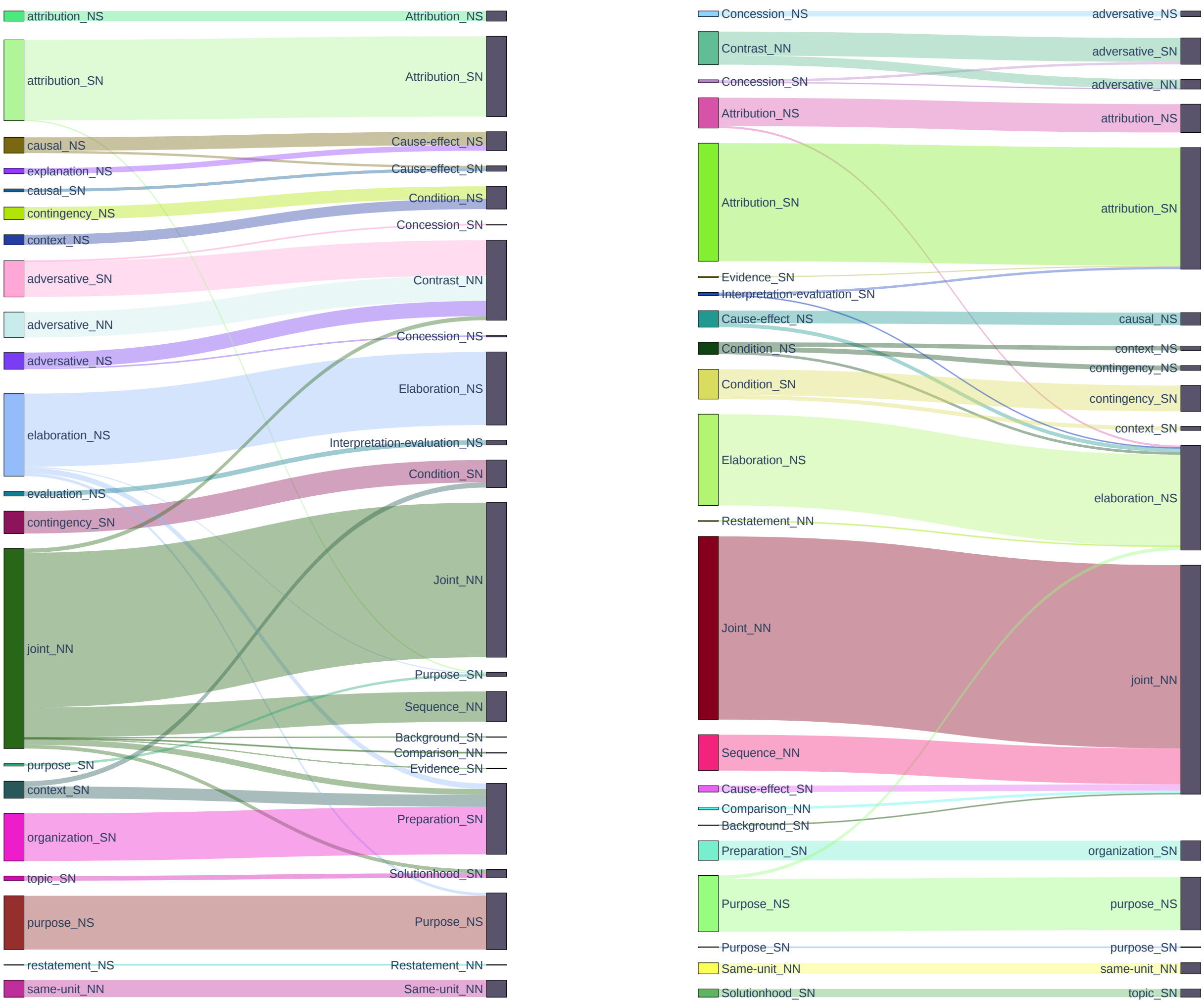


(b) RRG corpus annotation. Commas mark EDU boundaries.



(c) RRG parser prediction for English text.

Figure 4: An example of the direct cross-language segmentation prediction. From `GUM_news_nasa`.



(a) RRT Classifier → RRG

(b) RRG Classifier → RRT

Figure 5: A visual representation of the cross-dataset alignment between ground truth and predicted RST relations.