Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser

Elena Chistova
FRC CSC RAS
chistova@isa.ru

Abstract

We introduce UniRST, the first unified RST-style discourse parser capable of handling 18 treebanks in 11 languages without modifying their relation inventories. To overcome inventory incompatibilities, we propose and evaluate two training strategies: *Multi-Head*, which assigns separate relation classification layer per inventory, and *Masked-Union*, which enables shared parameter training through selective label masking. We first benchmark mono-treebank parsing with a simple yet effective augmentation technique for low-resource settings. We then train a unified model and show that (1) the parameter efficient Masked-Union approach is also the strongest, and (2) UniRST outperforms 16 of 18 mono-treebank baselines, demonstrating the advantages of a single-model, multilingual end-to-end discourse parsing across diverse resources.[1]

1 Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) represents discourse as a hierarchical tree of elementary discourse units (EDUs) connected by rhetorical relations. Over the years, RST has inspired the creation of multiple discourse treebanks across different languages. However, large-scale annotated corpora are scarce and predominantly available in English. For other languages, the high cost of annotation and inconsistent guidelines have resulted in smaller, heterogeneous resources with incompatible relation inventories.

The English RST Discourse Treebank (RST-DT) (Carlson et al., 2001), the primary benchmark for RST parsing, defines 56 fine-grained rhetorical relations, usually mapped to 18 coarse-grained classes for training and evaluation. Many discourse treebanks in other languages define considerably fewer relations. Aligning them with the RST-DT inventory often requires collapsing relations, such as merging CAUSE with EFFECT, CONTRAST with CONCESSION, or ELABORATION with ENTITY-ELABORATION. This process erases distinctions that can be crucial for downstream applications such as coreference resolution, narrative analysis, and opinion mining. Moreover, when no direct equivalents exist, alignment is frequently based on surface-level label similarity, which compromises annotation reliability across languages.

End-to-end RST parsing involves three interconnected subtasks: EDU segmentation, tree structure prediction, and nuclearity and relation labeling. The definitions of these tasks are shaped by the relation inventory and constraints of each treebank. For instance, segmentation decisions can be influenced by fine-grained intra-sentential relations. Mono- or multinuclearity of certain overlapping relations (LABEL_NS, LABEL_SN, LABEL_NN) varies across treebanks. When datasets with different inventories are merged and collapsed into a coarser label set, inconsistencies in relation definitions and nuclearity distributions can introduce substantial noise into both training and evaluation.

Despite these challenges, training on multiple treebanks offers clear benefits. RST-style parsers are known to generalize poorly across domains (Liu and Zeldes, 2023), and training a unified parsing model on all available treebanks may yield broader applicability. The skewed label distributions within individual corpora complicate model training, particularly in low-resource settings; pooling datasets with overlapping labels can mitigate this issue. Although larger treebanks provide sufficient data for accurate EDU segmentation and local relation labeling, they remain too small to support robust learning of global document structures. Leveraging all annotated structures across corpora can thus strengthen structural prediction. Altogether, these considerations motivate the development of univer-

[1] Our models and code: https://github.com/tchewik/UniRST.

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 197–208

sal discourse parsers that effectively integrate all available resources, regardless of language, genre, or domain.

In this work, we propose methods for building a unified RST parser from heterogeneous treebanks. Our contributions are:

1. The first large-scale RST parsing study covering 18 treebanks in 11 languages.

2. Data augmentation technique allowing for strong end-to-end mono-treebank RST parsing baselines even in low-resource settings.

3. Two strategies for jointly modeling divergent relation inventories: Multi-Head and Masked-Union.

4. Evaluations showing that: (i) dataset-specific segmentation heads are essential for handling varying EDU definitions; (ii) the Masked-Union approach enables sufficient model training by leveraging label overlap while respecting treebank-specific relation inventories, and (iii) our unified model outperforms 16 out of 18 mono-treebank baselines.

2 Related Work

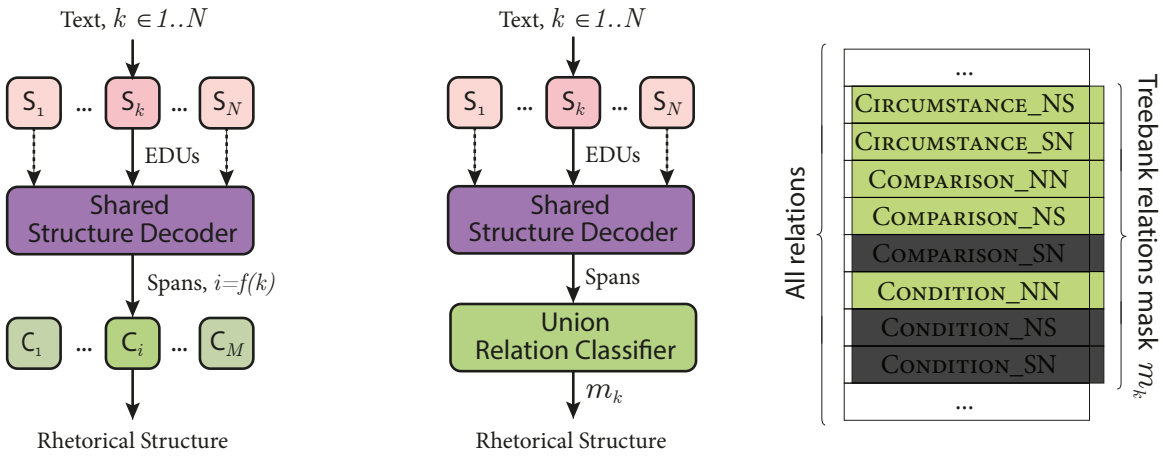
Cross-Lingual RST Parsing Cross-lingual rhetorical structure parsing has gained increasing attention in recent years. Braud et al. (2017) introduced a unified set of coarse-grained (harmonized) rhetorical relations and presented the first data-driven cross-lingual RST parser, transferring across English, Brazilian Portuguese, Spanish, German, Basque, and Dutch. Their study demonstrated that rhetorical structure parsing from pre-segmented texts successfully transfers beyond English and across typologically diverse languages. Building on this foundation, Liu et al. (2020) leveraged multilingual embeddings and proposed EDU-level machine translation to enrich training data. Subsequently, Liu et al. (2021) introduced DMRST, a unified framework performing joint EDU segmentation and discourse tree parsing, enabling end-to-end RST parsing evaluation across multiple languages under harmonized inventories. Extending this line of work, Chistova (2024) applied DMRST to parallel English–Russian data, highlighting the importance of aligned corpora for assessing cross-lingual transfer in the context of RST treebank incompatibilities.

Training on Incompatible Treebanks Research on integrating incompatible treebanks has largely focused on syntax parsing. Early work by Johansson (2013) introduced two adaptation techniques for training syntax parsers on treebanks with differing annotation schemes. Their methods involved concatenating the feature spaces of two treebanks and using a parser trained on one treebank to guide the other. These approaches were applied to treebanks pairs within the same language (German, Swedish, Italian, and English). Stymne et al. (2018) explored three strategies: treebank concatenation with and without fine-tuning, and the inclusion of treebank-specific embeddings. Their results showed consistent improvements in dependency parsing for most of the nine languages evaluated when using treebank-specific embeddings. A similar approach was applied by Barry et al. (2019) to train a cross-lingual parser for low-resource Faroese syntax parsing. Johansson and Adesam (2020) trained a Swedish constituency parser on six incompatible treebanks by sharing word representations across corpora while maintaining separate neural parsing modules for each treebank, thus accommodating both constituency and dependency annotations. Kankanampati et al. (2020) leveraged two Arabic dependency treebanks to build a parser with a unified attachment scorer. Sayyed and Dakota (2021) conducted multilingual experiments with treebank-specific biaffine parsing layers for UD and SUD syntactic annotations, ultimately finding that combining distinct annotation schemes could degrade parsing performance.

Notably, in syntactic parsing, terminal nodes correspond to words, so efforts to resolve annotation inconsistencies are confined to structure building and label assignment. In contrast, rhetorical structure parsing additionally requires segmentation, which is affected by treebank-specific constraints on elementary discourse units. In our work, we aim to develop the first end-to-end RST parser benefiting from each annotation scheme in a wide range of diverse discourse treebanks.

3 UniRST

We address joint training over heterogeneous RST corpora while preserving each treebank's native relation inventory, EDU segmentation, and relational definitions. Building on the DMRST architecture (Liu et al., 2021), we explore extensions that enable training across incompatible tree-



(a) MH architecture

(b) MU architecture

Figure 1: Model variants in the UniRST framework. (a) Multi-Head: independent classifiers per relation inventory. (b) Masked-Union: shared classifier with treebank-specific label masking.

banks. Specifically, we propose two strategies: *Multi-Head* (MH), which maintains separate classification heads per inventory, and *Masked-Union* (MU), which uses a single classifier constrained by treebank-specific masks. For reference, we additionally implement *Unmasked-Union* (UU), which lacks label masking and serves as a lower bound. Unless otherwise noted, models use treebank-specific segmentation heads, though shared segmentation is also tested. Figure 1 illustrates the architectures.

3.1 DMRST

DMRST (Liu et al., 2021) is an end-to-end RST parsing model that integrates EDU segmentation, discourse tree construction, and relation/nuclearity labeling. Its pipeline has four stages: (1) a pretrained language model encodes input tokens, (2) an LSTM-CRF module detects EDU boundaries, (3) a recurrent pointer network decoder constructs the discourse tree, and (4) a biaffine classifier assigns nuclearity and relation labels. The model is trained jointly, with dynamically weighted loss balancing segmentation, structure prediction, and labeling. This unified design enables consistent end-to-end parsing.

UniRST extends this backbone to multi-treebank training. The pretrained encoder and recurrent decoder are shared across corpora, while segmentation and relation classification are treebank-tailored. This design aims to achieve robust structural prediction while respecting each corpus's definitions and constraints.

3.2 Multi-Head (MH)

Our first method for multi-inventory RST parsing assigns a separate classification head to each distinct relation inventory. Given the set of inventories $\mathcal{G} = \{G_1, \ldots, G_M\}$, treebanks sharing the same inventory (e.g., `eng.gum`, `rus.rrg`, `zho.gcdt`) share a relation/nuclearity classifier $\mathbf{W}^{(m)} \in \mathbb{R}^{d \times |G_m|}$. In this configuration, cross-treebank information about relation and nuclearity is exchanged only implicitly, through fine-tuning of the language model and shared structural decoder.

3.3 Masked-Union (MU)

Let $\mathcal{U} = \bigcup_k \mathcal{L}_{T_k}$ be the unified set of all relation types across treebanks. MU employs a single shared classifier $\mathbf{W} \in \mathbb{R}^{d \times |\mathcal{U}|}$ that predicts over this unified label space. To enforce inventory constraints, for each treebank T_k, we apply a binary mask $\mathbf{m}^{(k)} \in \{-1 \times 10^9, 1\}^{|\mathcal{U}|}$ to the classifier logits. This parameter-efficient design promotes explicit parameter sharing and enables direct transfer for overlapping relations (e.g., ELABORATION_NS) across all components of the model.

3.4 Unmasked-Union (UU)

UU mirrors the MU architecture but omits the treebank-specific masking, thereby allowing predictions over the entire concatenated label set without restriction. Consequently, it can produce labels that do not exist in the target corpus, limiting its practical utility. We include UU scores as a lower-bound baseline.

Treebank	Language	# Docs	# Tokens	# EDUs	# Labels	# Classes	# Rels
ces.crdt (2024)	Czech	54	14,623	1,345	23	34	1,288
deu.pcc (2014)	German	176	32,836	2,842	25	37	2,665
eng.gentle (2023) (test only)	English	26	17,799	2,328	15	27	2,552
eng.gum v11.1 (2025)	English	255	250,290	34,428	15	27	32,173
eng.oll (2008)	English	327	46,177	3,026	21	35	2,716
eng.rstdt (2002)	English	385	205,829	21,789	18	42	21,404
eng.sts (2008)	English	150	70,422	3,208	21	35	3,058
eng.umuc (2024)	English	87	61,292	5,421	28	46	5,334
eus.ert (2013)	Basque	88	45,780	2,509	24	31	2,421
fas.prstc (2021)	Persian	150	66,694	5,789	18	26	5,638
fra.annodis (2012)	French	86	32,699	3,307	18	20	3,221
nld.nldt (2012)	Dutch	80	24,898	2,326	27	45	2,246
por.cstn (2011)	Portuguese	140	58,793	5,527	22	38	5,387
rus.rrg (2024)	Russian	213	172,405	25,222	15	27	25,010
rus.rrt (2017)	Russian	233	262,495	28,247	17	25	25,892
spa.rststb (2011)	Spanish	267	58,717	3,351	29	43	3,084
spa.sctb (2018)	Spanish	50	16,515	744	20	26	694
zho.gcdt (2022)	Chinese	50	62,905	9,403	15	28	9,345
zho.sctb (2018)	Chinese	50	15,496	744	20	26	684

Table 1: Treebank statistics.

4 Data

This study leverages training data from 18 RST treebanks covering 11 languages, aiming to create the most universal end-to-end RST parser to date. The treebanks span Czech, German, English, Basque, Persian, French, Dutch, Brazilian Portuguese, Russian, Spanish, and Chinese. Treebank statistics are summarized in Table 1.

For the English RST-DT benchmark, we adopt the coarse-grained relation labels used in prior work. Corpora annotated using the GUM RST schema (`eng.gum`, `zho.gcdt`, `rus.rrg`) retain their predefined coarse-grained labels. For other corpora, if applicable, we merged infrequent classes (less than 10 instances) with related ones based on nuclearity, following the mapping suggested by Braud et al. (2017). This ensures both label diversity and sufficient representation for training. Detailed class distributions are illustrated in Appendix B.

To ensure consistency and reproducibility, we use the standardized training, validation, and test splits[2] provided by the DISRPT 2025 shared task for segmentation, connective identification, and relation classification across discourse annotation frameworks.

4.1 Data Augmentation

While several large RST treebanks dominate end-to-end discourse parsing research, smaller corpora remain underutilized due to limited training data. To address this gap and establish strong mono-treebank baselines, we propose a simple yet effective data augmentation technique to improve performance in low-resource settings. Crucially, our method enriches training data without modifying the original texts or local annotations.

DMRST model employs a recurrent structure prediction module that relies heavily on contextual signals. As each annotated tree yields a single training instance, the number of examples is limited, particularly in smaller treebanks. To address this, we introduce an augmentation approach based on extracting structurally coherent subtrees from annotated documents. While these subtrees omit full-document context, their internal discourse structure remains valid and informative.

Our procedure involves: (1) identifying sentence boundaries to avoid extracting subtrees spanning sentence fragments; (2) extracting all connected subtrees not spanning sentence fragments and including at least three rhetorical relations; and (3) sampling a proportion p_{aug} of these subtrees for augmentation. Sampling is critical to prevent overfitting, particularly for the segmentation subtask.

This augmentation allows the model to train on a wider range of partial structures, potentially improving end-to-end RST parsing training in low-resource settings. We set p_{aug} to 50% to enrich the training data multifold.[3]

[2] We employ the open version of `eus.ert` treebank from `https://ixa2.si.ehu.eus/diskurtsoa/en/`, containing 88 annotations.

[3] For RST-DT, $p_{\text{aug}} = 50\%$ produces 5.4 times more training samples. Over all treebanks, it multiplies number of training samples by 7.7.

	Baseline									+ Augmentation								
	Gold seg				End-to-end					Gold seg				End-to-end				
Treebank	**S**	**N**	**R**	**Full**	**Seg**	**S**	**N**	**R**	**Full**	**S**	**N**	**R**	**Full**	**Seg**	**S**	**N**	**R**	**Full**
ces.crdt	60.2	31.1	18.2	16.9	90.1	47.3	24.0	13.7	12.3	58.9	31.1	18.0	17.1	90.6	46.2	23.4	11.5	11.2
deu.pcc	68.7	38.8	24.7	23.6	95.2	58.9	33.9	21.0	20.0	67.2	42.7	26.4	25.6	96.0	59.7	36.9	21.7	21.2
eng.gum	73.3	60.5	52.6	51.5	95.5	66.9	55.4	48.3	47.4	72.8	59.9	52.6	51.4	95.2	66.1	54.3	47.9	46.9
eng.oll	65.9	48.2	29.5	29.3	89.7	51.4	36.7	21.9	21.5	61.8	43.2	29.0	28.4	91.2	54.1	36.4	24.5	24.0
eng.rstdt	77.5	66.6	56.1	54.6	97.6	73.8	63.4	53.3	51.8	78.3	67.5	57.0	55.2	97.7	74.9	64.5	54.6	52.9
eng.sts	46.5	35.1	21.7	21.1	89.7	38.2	28.8	18.2	18.0	44.1	32.9	20.5	19.6	88.4	33.5	24.5	16.1	15.7
eng.umuc	68.8	50.8	33.1	32.4	89.6	51.2	36.9	24.3	23.7	67.1	48.4	31.8	31.2	89.0	49.1	35.1	24.3	24.0
eus.ert	71.0	47.3	29.9	29.2	89.7	54.8	38.0	23.1	22.7	66.5	44.5	25.7	25.7	89.0	52.3	35.3	19.9	19.8
fas.prstc	65.0	51.3	40.2	40.1	93.8	55.3	44.6	34.4	34.4	65.3	50.9	40.7	40.4	93.8	55.3	42.9	34.3	34.0
fra.annodis	62.5	51.6	33.0	33.0	92.1	53.2	44.7	28.6	28.6	62.5	51.4	32.9	32.9	91.5	52.4	43.3	27.6	27.6
nld.nldt	63.8	47.4	30.7	29.1	96.3	58.2	42.9	28.5	26.9	61.7	46.4	30.6	28.8	96.4	57.2	42.8	28.9	27.3
por.cstn	76.0	62.2	50.9	50.8	93.9	68.2	53.3	43.9	43.8	76.1	61.6	49.9	49.9	94.0	66.3	53.3	43.0	43.0
rus.rrg	71.2	57.2	49.4	48.2	97.2	67.6	54.3	47.1	46.0	70.3	55.9	47.9	46.8	96.8	65.6	52.2	44.9	43.9
rus.rrt	79.7	61.4	51.5	51.3	91.1	62.8	49.0	41.4	41.3	80.0	61.9	52.2	51.9	91.0	63.0	49.9	42.5	42.3
spa.rststb	68.1	52.2	35.0	35.0	91.5	54.9	40.9	28.1	28.1	71.1	54.4	38.9	38.9	91.9	57.7	44.0	32.8	32.8
spa.sctb	66.7	41.5	35.2	35.2	74.1	34.3	25.4	23.1	23.1	66.7	43.6	31.7	31.7	84.1	47.5	35.3	27.3	27.3
zho.gcdt	76.3	58.1	52.3	50.7	91.2	61.0	46.1	40.9	39.6	75.3	58.2	51.9	50.4	91.9	64.0	49.5	43.5	42.3
zho.sctb	66.7	37.7	32.1	32.1	91.1	56.3	34.3	28.5	28.5	60.2	40.9	32.3	32.3	92.5	52.3	38.1	29.6	29.6

Table 2: Performance of the treebank-specific models, with and without train data augmentation.

5 Experimental Setup

We employ `xlm-roberta-large` as the multilingual encoder across all experiments. The batch size is set to 2, with a hidden size of 200 for the segmenter and 512 for the parsing module. The DMRST model is trained with a learning rate of 1e-5, while the encoder is fine-tuned using a learning rate of 2e-5. Early stopping is set to a patience of 5 in mono-treebank settings and reduced to 3 in UniRST due to the larger concatenated dataset.

Evaluation follows the original Parseval metrics for rhetorical structure parsing, with micro F1 scores reported for segmentation (Seg), span (S), relation (R), nuclearity (N), and full structure (Full). Each model is trained using three different random seeds, and all reported results are averaged across these runs.

6 Experimental Results

6.1 Mono-Treebank Evaluations

Table 2 reports the performance of treebank-specific models trained with and without data augmentation. Augmentation yielded substantial gains on smaller corpora such as `eng.oll`, `spa.sctb`, `zho.sctb`, and `zho.gcdt`, but improvements were not uniform across all treebanks. Interestingly, on the `eng.rstdt` benchmark with diverse document lengths, augmentation led to an average 1.1% F1 improvement in unlabeled structure prediction (S), highlighting its potential even for larger datasets. On the other hand, the data augmentation resulted in performance degradation on two large-scale GUM-based corpora (`eng.gum`, `rus.rrg`), likely due to segmenter overfitting on long documents. Overall, augmentation yielded the best mono-treebank parsing performance on 10 of the 18 treebanks. For comparison, Appendix A summarizes previous end-to-end RST parsing results on eight treebanks. DMRST+ denotes the architecture used as a baseline in this work.

Model	**In-treebank**			**All avg.**		
	Seg	**S**	**N**	**Seg**	**S**	**N**
ces.crdt	90.5	49.5	24.5	76.4	32.6	16.0
deu.pcc	96.0	59.7	36.9	76.6	32.9	19.3
eng.gum	95.5	66.9	55.4	78.9	41.3	30.2
eng.oll	91.2	54.1	35.4	77.3	31.9	18.7
eng.rstdt	97.7	74.9	64.5	78.4	40.2	28.8
eng.sts	89.7	38.2	32.6	75.8	29.1	16.8
eng.umuc	89.6	51.2	36.9	77.6	35.1	22.4
eus.ert	89.7	54.8	38.0	77.5	33.8	18.4
fas.prstc	93.8	58.3	44.6	78.3	36.2	20.3
fra.annodis	92.1	53.2	44.7	77.5	32.6	16.5
nld.nldt	96.4	57.0	43.9	80.2	35.9	22.0
por.cstn	94.1	68.8	57.2	78.2	37.1	25.2
rus.rrg	97.2	67.6	54.3	80.0	40.9	28.7
rus.rrt	91.0	63.0	49.9	81.2	40.9	27.3
spa.rststb	91.9	57.7	44.0	78.5	35.4	22.6
spa.sctb	84.1	47.5	35.3	74.0	29.9	16.2
zho.gcdt	91.9	64.0	49.5	76.9	36.7	23.3
zho.sctb	92.5	52.3	38.1	57.4	17.1	10.2
UniRST	—	—	—	92.9	60.7	47.3

Table 3: Evaluation across all treebanks. We only assess segmentation (Seg), unlabeled structure construction (S), and nuclearity assignment (N), as relation inventories are incompatible.

Method	Segmentation	Gold seg				End-to-end				
		S	N	R	Full	Seg	S	N	R	Full
MH	single	73.5	58.4	47.6	46.6	93.4	63.4	50.7	41.7	40.8
	multiple	73.6	59.0	48.5	47.6	93.7	63.7	51.3	42.4	41.6
UU	single	73.8	58.7	47.0	46.8	93.4	64.3	51.9	42.8	41.8
MU	single	74.1	59.3	48.8	47.8	93.7	64.5	**52.1**	43.2	42.3
	multiple	**74.4**	**59.6**	**49.3**	**48.3**	**93.9**	**64.8**	**52.1**	**43.4**	**42.5**

Table 4: Performance of the UniRST model in different setups.

Treebank	Seg	S	N	R	Full
ces.crdt	94.2 (+4.1)	57.9 (+10.6)	38.6 (+14.6)	27.3 (+3.3)	26.8 (+14.5)
deu.pcc	96.5 (+0.5)	66.3 (+6.6)	45.5 (+8.6)	32.8 (+11.1)	31.1 (+9.9)
eng.gum	95.2 (-0.3)	66.7 (-0.2))	54.7 (-0.7)	48.0 (-0.3)	46.9 (-0.5)
eng.oll	93.8 (+2.6)	56.7 (+2.6)	40.6 (+4.2)	27.6 (+3.1)	27.1 (+3.1)
eng.rstdt	97.8 (+0.1)	75.6 (+0.7)	65.1 (+0.6)	55.2 (+0.6)	53.5 (+0.6)
eng.sts	91.0 (+1.3)	40.4 (+2.2)	30.7 (+1.9)	19.4 (+1.2)	18.8 (+0.8)
eng.umuc	88.8 (-0.8)	52.0 (+0.8)	40.1 (+3.2)	26.1 (+1.8)	25.6 (+1.9)
eus.ert	92.0 (+2.3)	62.8 (+8.0)	47.4 (+9.4)	35.4 (+12.3)	35.3 (+12.6)
fas.prstc	94.6 (+0.8)	61.7 (+6.4)	50.2 (+5.6)	40.7 (+6.3)	40.5 (+6.1)
fra.anodis	90.9 (-1.2)	58.1 (+4.9)	47.3 (+2.6)	31.1 (+2.5)	30.7 (+2.1)
nld.nldt	97.6 (+1.2)	59.3 (+2.1)	45.3 (+2.5)	33.5 (+4.6)	31.7 (+4.4)
por.cstn	94.3 (+0.4)	67.7 (-0.5)	54.9 (+1.6)	45.7 (+1.8)	45.4 (+1.6)
rus.rrg	96.5 (-0.7)	66.8 (-0.8)	53.5 (-0.8)	45.5 (-1.6)	44.1 (-1.9)
rus.rrt	90.6 (-0.4)	63.0 (0.0)	49.8 (-0.1)	42.6 (+0.1)	42.4 (+0.1)
spa.rststb	92.5 (+0.6)	63.5 (+5.8)	50.1 (+6.1)	35.3 (+2.5)	35.2 (+2.4)
spa.sctb	86.0 (+1.9)	55.8 (+8.3)	48.0 (+12.7)	40.8 (+13.5)	40.8 (+13.5)
zho.gcdt	92.1 (+0.2)	62.9 (-1.1)	48.7 (-0.8)	44.0 (+0.5)	42.7 (+0.4)
zho.sctb	94.3 (+1.8)	64.3 (+12.0)	50.5 (+12.4)	40.7 (+11.1)	40.7 (+11.1)

Table 5: UniRST performance per treebank. Improvements over the strongest mono-treebank baseline, as listed in Table 2, are shown in parentheses.

To assess generalization, each best-performing treebank-specific model was evaluated on all 18 corpora. Table 3 reveals a consistent transferability gap: models tend to overfit to treebank-specific language, domains, relation usage, and document styles. Segmentation scores also decline in transfer settings, though less severely than Span or Nuclearity scores. In certain cases, however (e.g., `eng.oll`, `eng.gum`), segmentation drops sharply, reflecting variation in EDU definitions across corpora. Despite strong in-treebank Span F1 (e.g., 74.9% for `eng.rstdt`, 68.8% for `por.cstn`), transfer performance degrades substantially (dropping to 40.2% and 37.1%, respectively). This disparity demonstrates that in-domain success is a poor indicator of cross-corpus robustness and highlights the need for more generalizable RST parsers, such as UniRST.

6.2 UniRST

Performance of the Multi-Head and Masked-Union strategies is reported in Table 4. UniRST performs best when segmentation is handled by treebank-specific heads, which capture differences in EDU annotation schemes, whereas a universal segmentation head primarily learns broader segmentation patterns. The Masked-Union (MU) strategy consistently outperforms Multi-Head (MH), offering both greater efficiency and higher parsing accuracy. Its masking mechanism ensures that each treebank's inventory is respected, while still enabling transfer for overlapping relations, which in turn improves parsing performance over the unmasked baseline. The strongest configuration is MU with treebank-specific segmentation heads. We refer to this variant as "UniRST" throughout the remainder of the paper.

As shown in Table 3, UniRST achieves higher average performance across combined test set compared to any mono-treebank parser. This demonstrates the robustness of UniRST model as a cross-lingual parser capable of learning shared representations that generalize effectively across diverse

RST corpora.

Detailed results by treebank are provided in Table 5. The unified model outperforms the strongest mono-treebank baselines on 16 out of 18 treebanks. Notable improvements in end-to-end Full F1 are observed across most datasets, particularly for smaller-scale treebanks such as `ces.crdt`, `deu.pcc`, `eus.ert`, `spa.sctb`, `zho.gcdt`, and `zho.sctb`. Similar to data augmentation in mono-treebank training, joint training does not benefit the large-scale `eng.gum` and `rus.rrg` corpora, whose annotations appear sufficient on their own. Importantly, the performance drop on `eng.gum` under joint training remains marginal. The only corpus where UniRST fails to exceed 50% Span F1 and 25% Full F1 is `eng.sts`. Given the limited documentation of this dataset, the cause is unclear, but the low scores may stem from poor inter-annotator agreement or inconsistently applied segmentation and structural constraints. Joint training nonetheless improved performance, suggesting that it provides some stabilization even under noisy conditions. Across nine corpora in English, Persian, Portuguese, Russian, Spanish, and Chinese, UniRST achieves more than 40% Full end-to-end F1 while preserving original relation inventories.

To further assess out-of-domain generalization, we evaluate GUM-compatible models on the GENTLE benchmark, which follows GUM annotation guidelines.[4] As shown in Table 6, UniRST achieves the highest Full end-to-end parsing score. The `eng.gum` model performs best in segmentation (93.0% F1) and structure prediction (58.0%) due to its alignment with GENTLE's language and annotation conventions. However, UniRST outperforms it on Relation and Full F1, highlighting the benefits of shared relation classification training across multiple treebanks. Notably, UniRST supports 11 languages, while `eng.gum` is English-only. Training on multiple multi-domain treebanks, including five English treebanks, did not lead to a substantial improvement in out-of-domain performance over the GUM-specific model. These findings highlight the importance of treebank-specific annotation schemes and show that the universal model remains most effective within the domains and genres present in its training data.

[4]GENTLE includes annotations for eight unconventional genres: dictionary entries, esports commentaries, legal documents, medical notes, poetry, mathematical proofs, syllabuses, and threat letters. None of these genres are represented in the training corpora used in this work.

Model	Seg	S	N	R	Full
eng.gum	**93.0**	**58.0**	**47.2**	39.1	38.6
rus.rrg	85.2	44.7	34.9	28.8	28.3
zho.gcdt	76.4	34.1	23.5	18.4	18.0
UniRST	92.7	57.4	46.0	**39.9**	**39.4**

Table 6: Performance of the GUM-compatible models on GENTLE out-of-domain benchmark.

7 Conclusion

While previous approaches to multilingual parsing have often advocated for reducing relation inventories to a small standardized set of RST relations, such strategies fail to fully account for the broader divergences among RST treebanks. These include differences in discourse segmentation, the treatment of mono- versus multinuclearity, and the granularity, specificity, and definitions of rhetorical relations. In this work, we introduced UniRST, the first unified RST-style discourse parser capable of effectively processing 18 treebanks across 11 languages without altering their original relation inventories. To address the challenge of inventory incompatibility, we proposed two approaches: Multi-Head and Masked-Union. Our results show that the latter yields superior performance, particularly when paired with treebank-specific segmentation heads. UniRST outperforms 16 out of 18 mono-treebank baselines, demonstrating that end-to-end multilingual discourse parsing is achievable despite considerable annotation diversity. The results indicate that embracing annotation heterogeneity can benefit multilingual discourse parsing.

Limitations

The main limitation of a multilingual RST parser that preserves multiple relation inventories lies in the need to account for inventory differences in downstream applications. This issue is not unique to our approach, as it also arises when deploying separate treebank-specific models per language or domain. Even under label harmonization to a reduced set, variation in the number and distribution of relations across languages can persist. While UniRST demonstrates strong generalization across most treebanks, it shows a marginal performance drop on two large, multi-domain corpora (`eng.gum`, `rus.rrg`), likely because their annotations are sufficient to support strong mono-treebank models. Furthermore, `eng.sts` remains the only dataset where Span F1 remains below 50%, with both mono- and

multi-treebank models performing poorly. These observations suggest that data quality and annotation consistency substantially affect performance, and that future work may benefit from treebank filtering or weighting.

Acknowledgments

The research was carried out using the infrastructure of the shared research facilities «High Performance Computing and Big Data» of FRC CSC RAS (CKP «Informatics»).

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. GENTLE: A genre-diverse multilayer challenge set for English NLP and linguistic evaluation. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 166–178, Toronto, Canada. Association for Computational Linguistics.

James Barry, Joachim Wagner, and Jennifer Foster. 2019. Cross-lingual parsing with polyglot training and multi-treebank learning: A Faroese case study. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 163–174, Hong Kong, China. Association for Computational Linguistics.

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

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.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, M. Eloize, R. Kibar Aji Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. CST-News - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, Brazil.

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*.

Elena Chistova. 2024. Bilingual rhetorical structure parsing with large parallel annotations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9689–9706, Bangkok, Thailand. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: An online search interface to check rhetorical relations. In *4th Workshop on RST and Discourse Studies*, pages 40–49, Fortaleza, Brasil.

Richard Johansson. 2013. Training parsers on incompatible treebanks. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–137, Atlanta, Georgia. Association for Computational Linguistics.

Richard Johansson and Yvonne Adesam. 2020. Training a Swedish constituency parser on six incompatible treebanks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5219–5224, Marseille, France. European Language Resources Association.

Yash Kankanampati, Joseph Le Roux, Nadi Tomeh, Dima Taji, and Nizar Habash. 2020. Multitask easy-first dependency parsing: Exploiting complementarities of different dependency representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2497–2508, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Mary Ellen Okurowski Lynn Carlson, Daniel Marcu. 2002. RST Discourse Treebank LDC2002T07.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Draoulec, and Laure Vieu. 2012. Manuel d'annotation en relations de discours du projet annodis.

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, A. Nasedkin, S. Nikiforova, I. Pavlova, and A. Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, pages 194–204.

Lucie Polakova, Jiří Mírovský, Šárka Zikánová, and Eva Hajicova. 2024. Developing a Rhetorical Structure Theory treebank for Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4802–4810, Torino, Italia. ELRA and ICCL.

Andrew Potter. 2008. Interactional coherence in asynchronous learning networks: A rhetorical approach. *The Internet and Higher Education*, 11(2):87–97.

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

Zeeshan Ali Sayyed and Daniel Dakota. 2021. Annotations matter: Leveraging multi-task learning to parse UD and SUD. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3467–3481, Online. Association for Computational Linguistics.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.

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

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

Karolina Zaczynska and Manfred Stede. 2024. Rhetorical strategies in the UN security council: Rhetorical Structure Theory and conflicts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28, Kyoto, Japan. Association for Computational Linguistics.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

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.

A Reference Results from Prior Work

Table 7 summarizes previously reported results for end-to-end RST parsing. It is important to note that prior results may differ in experimental setup,[5] limiting direct comparability. All results reported in Section 6.1 are obtained through single-treebank

[5]Most notably, in the use of multicorpus training with harmonized label sets, or non-standard train/dev/test splits.

training using the original relation sets and the standardized DISRPT 2025 splits. To the best of our knowledge, the remaining treebanks are evaluated here for the first time in a full end-to-end RST parsing setting.

System	Seg	S	N	R	Full
eng.rstdt					
SegBot (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)	96.3	68.4	59.1	47.8	46.6
DMRST+ (2024)	97.8	74.8	64.5	54.5	53.0
deu.pcc					
DMRST (2021)	96.5	70.4	60.6	n/c	n/c
eus.ert					
DMRST (2021)	88.7	53.3	39.1	n/c	n/c
nld.nldt					
DMRST (2021)	95.5	62.3	46.6	n/c	n/c
por.cstn					
DMRST (2021)	92.8	62.5	51.6	n/c	n/c
rus.rrg					
DMRST+ (2024)	96.9	66.5	53.3	45.8	44.6
rus.rrt					
DMRST+ (2024)	92.2	65.9	51.0	43.9	43.8
spa.rststb					
DMRST (2021)	92.8	62.5	51.6	n/c	n/c

Table 7: Reference end-to-end parsing evaluations across RST treebanks. **n/c** indicates incompatible (harmonized) label sets.

B Relation Classes across Treebanks

Figures 2 and 3 illustrate the distribution of all relation labels across 19 treebanks (including the test-only `eng.gentle`). UniRST handles all 96 unique LABEL_NUCLEARITY relations as they appear in each corpus. Note that while some treebanks (e.g., GUM-style and RST-DT) internally group ANTITHESIS, CONTRAST, and CONCESSION as ADVERSATIVE, and CAUSE with RESULT as CAUSAL, others treat some of these relations separately or organize them under alternative groupings.

During preprocessing, only relations with equivalent definitions and comparable granularity were unified under a single label (e.g., CONDITION and CONTINGENCY; ADVERSATIVE and coarse-grained CONTRAST). CONDITION is a coarse-grained label encompassing, in most treebanks, the underrepresented fine-grained relations OTHERWISE, UNLESS, and UNCONDITIONAL, each of which appears too infrequently to be modeled reliably on its own. Labels without clear counterparts, such as GRADATION_SN in `ces.crdt` (Polakova et al., 2024) or FRAME_NS in `fra.annodis` (Muller et al., 2012), remain unique to their respective treebanks.

Variations in the representation of overlapping labels across treebanks reflect underlying genre and linguistic differences. For instance, `zho.gcdt` features more instances of ELABORATION_SN than ELABORATION_NS, in stark contrast to other languages, where the satellite in ELABORATION typically follows the nucleus.

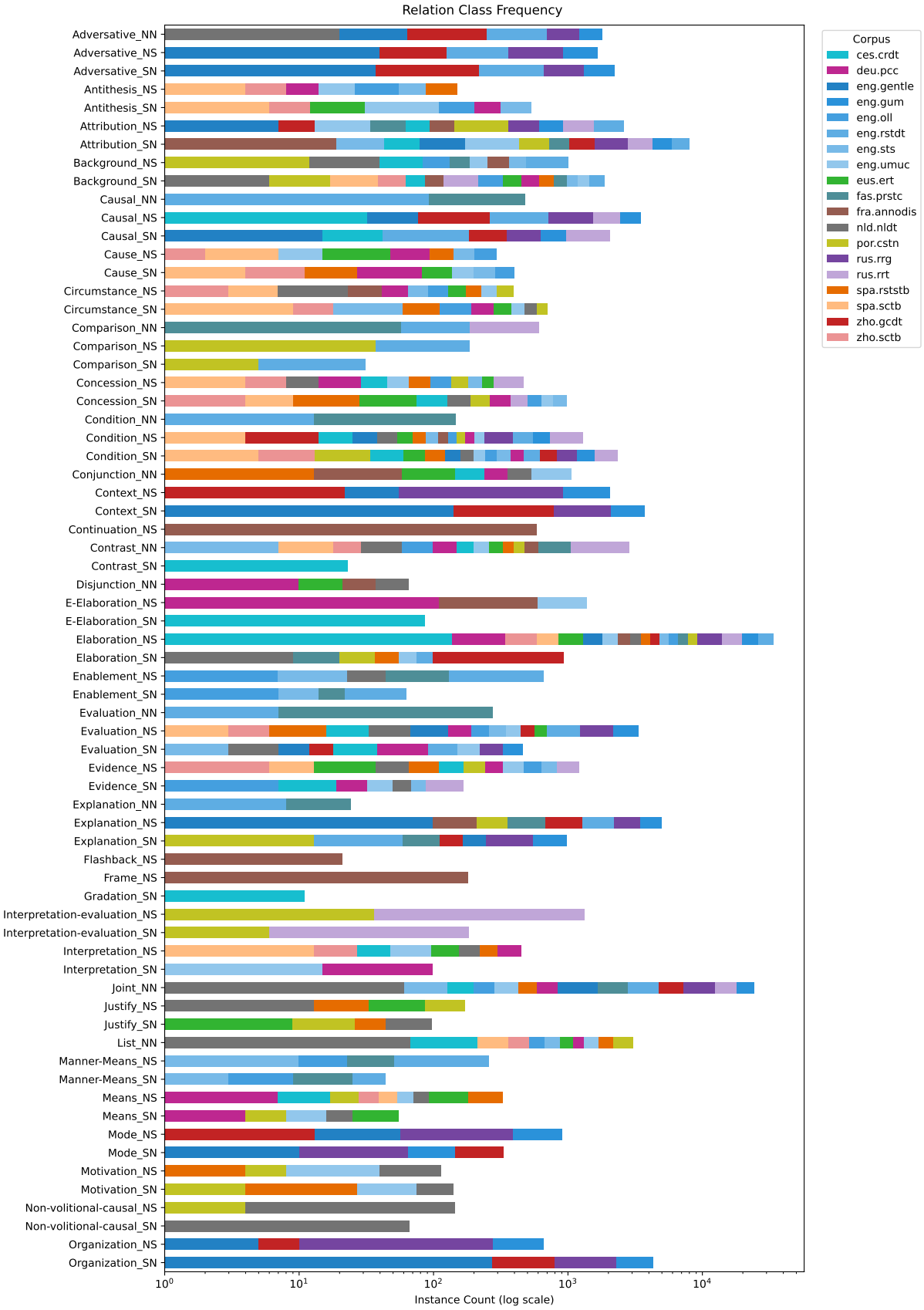


Figure 2: Relation class frequency across treebanks.

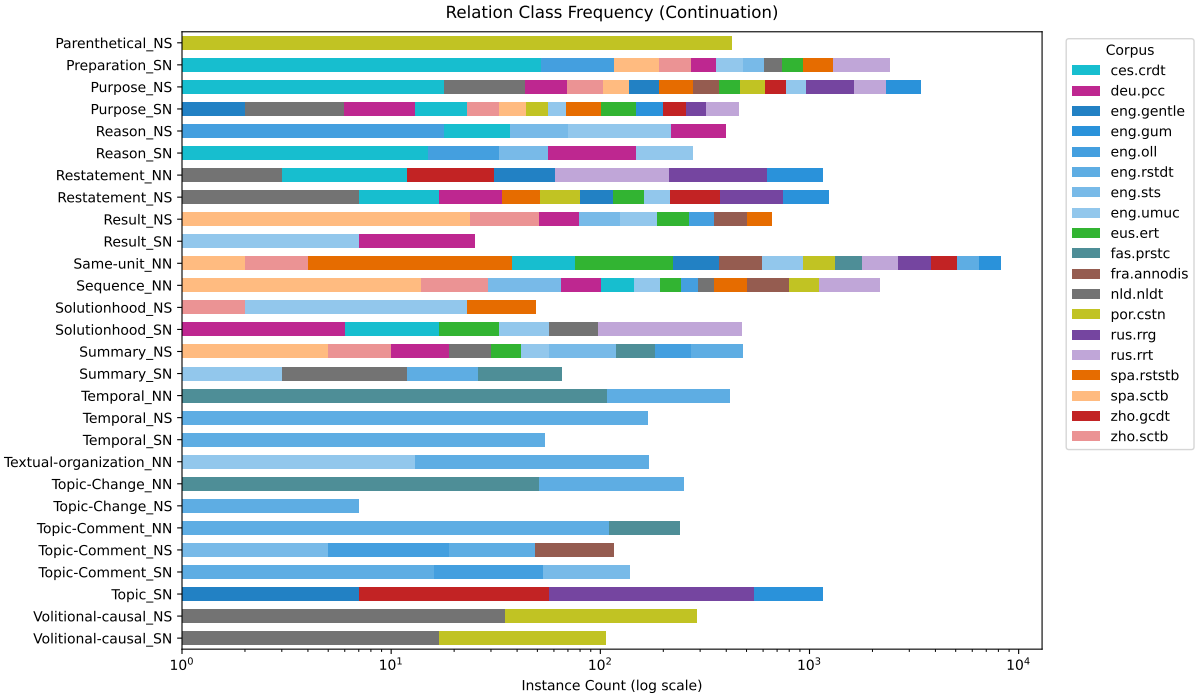


Figure 3: Relation class frequency (continuation).